Meaning-Informed Low-Resource Segmentation of Agglutinative Morphology

Caleb Belth
University of Utah
caleb.belth@utah.edu

Abstract

Morphological segmentation is both an interesting acquisition problem and an important task for natural language processing. Most current computational approaches either use supervised machine learning—which tends to lead to the best-performing models—or operate over bare surface forms of words. However, the empirical conditions of language acquisition seem to fall somewhere in between: children do not have access to pre-segmented input, yet their knowledge of morphological structure develops alongside semantic knowledge. Inspired by this, we suggest a simple model for low-resource segmentation of agglutinative morphology. The model is based on the idea that agglutination tends to mark one meaning per form. It is unsupervised, but is able to exploit features to identify how differences between closely-related surface forms are marked. Trained on hundreds to a few thousand words from languages with agglutinative morphology, the resulting model outperforms an unsupervised model that does not exploit such features, and in some settings even outperforms a supervised model trained on both features and ground-truth segmentations.

1 Introduction

One of the challenges of language learning is to identify the meaning-bearing units—that is *morphological segmentation*. Segmentation has also been important to natural language processing for decades (Kurimo et al., 2010; Batsuren et al., 2022), and continues to be due to the usefulness of subword units for prominent tasks like neural language modeling and machine translation (Sennrich et al., 2016; Kudo, 2018; Brown et al., 2020; Pan et al., 2020).

The problem presents a particular challenge in agglutinative languages, where several grammatical features may be expressed by stringing together affixes. For example, the Hungarian noun *ház* 'house' is combined with a possessive suffix *-aink* and essive case suffix 'ban' to form the word *házainkban* 'in our houses' (example from Ladányi et al. 2020, p. 191). Moreover, agglutination occurs in many low-resources languages (Moeng et al., 2021; Downey et al., 2022), and occurs alongside phonological processes like vowel harmony, which lead to alternation in the form that a given affix is realized as (Ladányi et al., 2020). For example, the Hungarian essive suffix is realized as *-ben*/*-ban* depending on the backness of the vowel to its left, as in *szekrényben* 'in the cupboard' and *barlangban* 'in the cave' (examples from Ladányi et al. 2020, p. 192).

Some approaches to segmentation are supervised, meaning that they learn from segmented training data. For example, the winner of the 2022 SIGMORPHON (Batsuren et al., 2022) segmentation challenge was a sequence-to-sequence transformer model (Peters and Martins, 2022). Other approaches—often preferred due to not requiring annotated training data—are unsupervised approaches, which are usually trained on bare surface forms (e.g., Uchiumi et al. 2015; Xu et al. 2020).

Our approach takes inspiration from language acquisition, where children show evidence of an ability to analyze words in morphologically-complex ways, segmenting them into distinct subunits or morphemes (Marquis and Shi, 2012; Mintz, 2013; Ladányi et al., 2020; Kim and Sundara, 2021). For example, Ladányi et al. (2020) demonstrated that when the common Hungarian suffix *-ban* was attached to a nonce stem (e.g., *púrban*), 15mo Hungarian-learning children indeed analyzed such nonces as suffixed words, as evidenced by their ability to later recognize the stem in bare form (see § 2.1 for more discussion).

We suggest that one mechanism useful to morphological segmentation in agglutinative languages could be the ability to recognize pairs of closely-related word forms, and then infer sim-

Proceedings of the Society for Computation in Linguistics (SCiL) 2024, pages 96-106.
Irvine, California, June 27-29, 2024

ple differences between each pair. For example *tanároknak*, the Hungarian plural (PL) dative (DAT) of 'teacher', differs from *tanárok* 'teachers' in only one feature (case), and the former can be derived from the latter by suffixing *-nak*. This provides the learner evidence that DAT can be marked by the suffix *-nak*. Moreover, if the learner knows from other pairs like *tanár*/*tanárok* 'teacher'/'teachers' that plurals are also marked by suffixation, then they can infer evidence that the DAT suffix is ordered after the plural suffix.

In this paper, we implement this proposal as a simple segmentation model, which uses morphological features to identify closely-related word pairs, from which it infers the concatinative operations that the language uses to mark those features. This approach offers the possibility of improvement over unsupervised approaches that operate over only surface forms, while simplifying the data-annotation demands of supervised approaches needing ground-truth segmentations. For example, Unimorph 3.0 (McCarthy et al., 2020) contains morphological features for 169 languages, but segmentations—via MorphyNet (Batsuren et al., 2021)—for only 15.

When trained on 500-10,000 words, the model achieves 72-100% accuracy segmenting test words in Finnish, Hungarian, Mongolian, and Turkish, out-performing the unsupervised model Morfessor 2.0 (Virpioja et al., 2013) and, in a majority of cases, a supervised neural comparison model. These results suggest that the model could be useful for segmenting low-resource agglutinative languages, since producing a small number of morphological-feature-annotated word forms is often easier than producing ground-truth segmentations, and such feature annotations yield large improvements over segmentations based on bare surface forms.

2 Model

2.1 Cognitive Motivation

Our model is motivated by experimental findings from child language acquisition. We are primarily concerned with the empirical promise of the model to segment agglutinative morphology, but in § 5 we discuss the extent to which we think the model is itself revealing about the mechanisms of the acquisition of morphological structure.

Marquis and Shi (2012) found that 11mo-old French-learning infants could perceive nonce words suffixed with the frequent French verbal suffix *-e* as related to their bare stems. This ability was not attested when an unfamiliar suffix *-u* was attached to nonce stems, suggesting that the infants were decomposing the nonces into stem and affix units rather than recognizing phonological overlap. At 15mo, Mintz (2013) found similar results for the English suffix *-ing*, and likewise Ladányi et al. (2020) for the Hungarian essive suffix *-ban*/*-ben*. The ability to relate forms was unperturbed by the vowel-harmony-induced alternation between suffix forms. Thus, given Hungarian's agglutinative morphology, the ability to relate inflected forms to their stems seems to develop even in the presence of agglutination and alternation.

Many of these results also suggest that the ability to relate closely-related forms may begin developing prior to children acquiring the function of morphemes. For example, Marquis and Shi (2012) found that presenting infants with many nonce words inflected with an unfamiliar suffix, they would begin to relate the inflected nonces to their stems. Moreover, Kim and Sundara (2021) found that the ability emerges for at least some English suffixes (*-s*) as early as 6mo, even when the nonces are presented without referential context, which they take as evidence that the ability begins developing without dependence on meaning.

Together, these studies suggest that infants can relate (concatenatively) inflected forms to their stems, and that this ability at least begins to emerge prior to children learning the function of morphemes. Payne (2022, 2023) has proposed that this early segmentation ability could allow children to identify *collisions*, which are instances of stems appearing in multiple inflected forms. Payne argues that these collisions provide evidence to the learner about what morphological features are marked in the language being acquired, via Clark (2014)'s observation that differences in form are indicative of differences in meaning. Payne's proposal, implemented as an explicit learning model, accurately matches developmental findings.

Given the well-attested ability of infants to relate word forms that differ in a single affix and the plausibility of Payne (2023)'s hypothesis about how learners can use this to discover what morphological features are marked, learners could then use the differences between related word forms and their developing knowledge of marked morphological features to identify what subparts of words correspond to these marked features—that is to pro-

duce meaning-informed segmentations. This idea forms the basis of our proposed model, which we present next.

2.2 Input

The model's input training data is a collection of $<w, r, f>$ words (triples), where w is the word's surface form, r is the word's root meaning, and f is a set of morphological features marked in the word. Notably, r is only the root *meaning* and not the root form. An example input is (1), which we will use as a running example.

(1) a. (*tanár*, TEACHER, {})
b. (*tanárok*, TEACHER, {PL })
c. (*tanároknak*, TEACHER, {PL, DAT })
d. (*személy*, PERSON, {})
e. (*személynek*, PERSON, {DAT })

2.3 Learning Algorithm

The model, which we call MIASEG ['mi.ə.sɛg] for *Meaning-Informed Agglutinative Segmentation*, learns from the input described above by identifying closely-related words and inferring the concatenative difference between their forms as a candidate marking of the feature difference between the words. MIASEG considers two words to be *closely-related* if they have the same root meaning and one has all the features of the other plus one.

Thus, a paradigm P_m corresponding to a root meaning m is represented as the set of input triples whose root meaning equals m, (2).

(2) $P_m \triangleq \{<w, r, f> : r = m\}$

For example, the paradigm P_{TEACHER} contains (1a)-(1c). This is shown in (5; step 1). MIASEG then computes, for each paradigm, the closely-related words in the paradigm—namely those where one word has all the features of the other plus one (3).

(3) $c(P_m) \triangleq (<w_i, r_i, f_i>, <w_j, r_j, f_j>) \in P_m : |f_i \cup f_j \setminus f_i \cap f_j| = 1$

Thus $c(P_{\text{TEACHER}})$ returns the pairs (1a)-(1b) and (1b)-(1c). For each of these pairs (5; step 2), MIASEG computes the string difference between the word forms w_i and w_j (5; step 3) and posits the difference as one way of marking the feature that differs between the two words (5; step 4). MIASEG represents this inference as a triple of the form (4), where $\phi = |f_i \cup f_j \setminus f_i \cap f_j|$ is the marked feature, Δ is the concatenative difference between w_i and w_j, and t specifies whether the difference is a suffix (i.e., comes at the right edge) or a prefix (i.e., comes at the left edge).

(4) $<\phi, \Delta, t>$

For example, the difference between (1b)-(1c) is the presence of an ending *-nak* in (1c), which has the additional feature DAT. Thus, MIASEG infers that the suffix *-nak* is one way of marking DAT: <DAT, *nak*, SUFF >. MIASEG also stores the number of times the difference has been inferred as a marking of the feature (i.e., the frequency of each triple), for prioritizing among multiple analyses during segmentation (§ 2.4). Moreover, because both (1b) and (1c) have the feature PL, MIASEG tabulates that the PL marker probably comes before the DAT marker.

At a different iteration of the loops, MIASEG will find the difference between (1d) and (1e) to be *-nek* and MIASEG will learn that this is another way to mark DAT. Thus, the markings inferred by MIASEG are effectively allomorphs of the morphemes corresponding to each marked feature. The resulting segmentations could be used as the input to a method like Belth (2023a)'s, which constructs underlying forms for morphemes based on surface alternation.

Once the for loops are complete, MIASEG infers a global ordering of features (5; step 6) by creating a directed graph, where each feature forms a node and an edge is formed from f_i to f_j whenever it was inferred that f_i must come before f_j (e.g., PL → DAT). The graph is then topologically sorted, which yields a total linear ordering of the features such that any orderings encoded in the graph edges are preserved in the linear ordering (Cormen et al., 2009, p. 612).[1]

(5) **Input:** Set of $<w, r, f>$ triples
1. **For** each paradigm P_m in data **do**
2. – **For** pair in $c(P_m)$ **do**
3. — Find Δ between w_i and w_j
4. — Posit Δ as marking of $f_i \cup f_j \setminus f_i \cap f_j$
5. —Tabulate implied feature orderings
6. Infer global ordering of features

We discuss the strengths and limitations of this algorithm in § 5. The code is available at https://github.com/cbelth/miaseg.

[1]Extensions may be necessary for languages with variable morpheme order, as this would introduce cycles into the graph. In the current implementation, if the ordering $f_i \rightarrow f_j$ and $f_j \rightarrow f_i$ are both inferred, only the ordering that was inferred the most times at line 5 is inserted into the graph.

2.4 Segmentation

Once the ways in which morphological features can be marked, and the ordering among them, are inferred and recorded, the model can segment words—either the words from which it made these inferences or new (test) words.

Segmentation takes as input a surface form, w (e.g., *csapatoknak*), and set of features f (e.g., {PL, DAT }). MIASEG iterates (6; step 1) over each feature in f in an order consistent with the ordering inferred during training (5; step 6)—left-to-right for prefixes and right-to-left for suffixes (e.g., DAT then PL since PL → DAT was inferred during training).

For each feature, the model looks up the ways in which it was marked in the training data (6; step 2), and tries each marking until one matches the end (for suffixes) or beginning (for prefixes) of w. The markings are considered in descending order of length, using the number of times the marking was attested in the training data as a tie breaker for equal-length matches. When a match is found, the matching ending is separated from the word as a morpheme. For example, DAT was marked as *-nak* and *-nek* in the training data, and *csapatoknak* ends in *-nak*, so *nak* is separated from the word to form *csapatok-nak*. The segmentation algorithm then proceeds to the next feature. For example, the model then looks at the ways in which PL can be marked for a match at the ending of *csapatok*, which will find *-ok*, resulting in *csapat-ok-nak*.

If at any point no attested marking of a feature matches (6; step 5), to prevent this from blocking further segmentation, MIASEG separates k segments from w, where k is the most common length of attested markings (for example $k = 1$ for a feature with attested markings {*a, e, ja*}).

(6) **Input:** <w, f> pair

1. **For** feat in f (ordered) **do**
2. – **For** attested marking of f **do**
3. — **If** marking matches edge of w **then**
4. —— Separate marking from w
5. – **If** no attested marking matched **do**
6. — Separate k segments from w
7. **Return** segmented w

3 Evaluation

Our evaluation attempts to test the effectiveness of the model at segmenting agglutinative languages in relatively low-resources settings, where only hundreds to a few thousands words are available for training.

Table 1: Dataset Sizes

Fin	541,198
Hun	613,549
Mon	11,215
Tur	18,333

3.1 Data

We collected data for Finnish (Fin), Hungarian (Hun), Mongolian (Mon), and Turkish (Tur), all languages with a substantial amount of agglutinative morphology. The languages come from three language families: Finnish and Hungarian are Uralic languages, Mongolian is a Mongolic language, and Turkish is a Turkic language. For all datasets except Turkish, we followed Batsuren et al. (2022) in using data from MorphyNet (Batsuren et al., 2021), which has canonical segmentations extracted from Wiktionary. For Turkish, we followed Belth (2023a,b, 2024) in using the corpus created for MorphoChallenge (Kurimo et al., 2010). We used Çöltekin (2010, 2014)'s publicly-available finite state morphological analyzer to generate morphological analyses.[2] The analyzer is designed for Turkish, and is similar to the approach used by MorphoChallenge to generate ground-truth analyses. For simplicity, we decided to look only at nouns for this paper. For each dataset, we extracted all nouns where we could unambiguously convert the canonical segmentation to a surface segmentation (Cotterell et al., 2016). The resulting dataset sizes are shown in Tab. 1.

We also collected corpus frequency information for each word in each dataset. For Finnish and Mongolian, we used the very large monolingual datasets aggregated by Conneau et al. (2020); Wenzek et al. (2020) from the 2018 CommonCrawl, counting the frequency of each word in the corpus. For Hungarian, we used the Hungarian Web Corpus (Halácsy et al., 2004) frequency file. Any word in our datasets that did not occur in these web corpuses we assumed to be low frequency (given the extremely large size of the web corpuses); we assigned them frequency of 1. The Turkish dataset already contained frequency information.

[2] https://github.com/coltekin/TRmorph

3.2 Setup

We discuss comparison models in § 3.2.1 and the training and evaluation procedures in § 3.2.2.

3.2.1 Comparison Models

We compare MIASEG, which is unsupervised but requires data be annotated with morphological features, to MORFESSOR, which is an unsupervised model that segments bare surface forms, and to TRANSFORMER, a supervised transformer-based encoder-decoder sequence to sequence (seq2seq) model that learns from segmented training data that is annotated with the same morphological features that MIASEG uses.

For MORFESSOR, we used the Morfessor 2.0 model (Virpioja et al., 2013), which is available as a Python package.

TRANSFORMER is the name of a supervised neural seq2seq model that we apply to the task. Neural seq2seq models have had success at many morphological problems, including the 2022 SIGMORPHON (Batsuren et al., 2022) challenge on morphological segmentation, where Peters and Martins (2022)'s DeepSPIN-3 model achieved the best-overall performance on the word-level task. However, to our knowledge, the code for DeepSPIN-3 is not publicly available, and the model does not incorporate morphological features. On the other hand, neural seq2seq models consistently perform well at the recurring SIGMORPHON morphological inflection task (see Kodner et al. 2022 for a recent iteration of the task), and these models commonly incorporate morphological features directly into the model, due to their importance to the inflection task (e.g., Makarov and Clematide 2018; Wu et al. 2021).

Thus, we follow Wu et al. (2021) in using a transformer-based encoder-decoder architecture, which includes both morphological features and word characters in the model's vocabulary. We describe the model's architecture in more detail below (§ 3.2.2).

3.2.2 Training and Evaluation

While unsupervised models like MORFESSOR and MIASEG can be evaluated on how well they segment the training data since they receive no information about the ground-truth segmentations during training, we wish to compare performance to the supervised setting (represented by TRANSFORMER), which necessitates evaluating on a held-out test set. Consequently, we chose to evaluate all three models on held-out test sets.

In relatively low-resource settings, as well as in child language acquisition, higher-frequency words are more likely to be represented than lower-frequency words. To approximate such a situation, we chose to sample training words weighted by frequency. We evaluated at three different training sizes: 500, 1000, and 10000. For each training size, we ran each model on 10 samples with different random seeds. Every word not included in the training sample was included in the held-out test set.

On each of the 10 random seeds, we tuned TRANSFORMER's hyperparameters using a grid search sweep. To do so, we made a random 80%/20% split of the training data, and trained the model with each hyperparameter combination on the 80% part of the split; we evaluated accuracy on the remaining 20%. We chose the hyperparameter combination that yielded the best accuracy on the 20%, remerged the 80%/20% split into the full training set, and then trained a new model with that hyperparameter combination on the entire training split. The hyperparameters we considered were those in (7), yielding 48 combinations.

(7) Embedding Dimension $\in \{256, 512\}$
Dropout = $\in \{0.1, 0.3\}$
Batch Size = $\in \{32, 128, 256\}$
Number of Enc. & Dec. Layers = $\in \{1, 2\}$
Number of Attention Heads = $\in \{4, 8\}$

We evaluate all models in terms of precision, recall, F1, and accuracy. Precision measures, out of all predicted morphemes (across the entire test set), what fraction are actually morphemes. Recall measures, out of all morphemes, what fraction are predicted. F1 measures the harmonic mean of precision and recall. Accuracy measures the fraction of test items that are correctly segmented.

3.3 Results

The results (F1 and accuracy)[3] are shown in Tab. 2. MIASEG outperforms the unsupervised MORFESSOR by a large margin on all datasets, and even outperforms the supervised TRANSFORMER model on 3/4 datasets. Importantly, the accuracy—not just the F1—is fairly high in absolute terms, even at a training size of only 1000 words. This means that a large majority of the test words are correctly segmented.

[3]We report the precision and recall values going into the F1 scores in Tab. 4 in the appendix.

Miaseg's performance is noticeably worse for Finnish than the other datasets, though it still performs competitively with the supervised Transformer and still outperforms the unsupervised Morfessor baseline. The primary reason for this is that the Nom plural is usually marked with [-t], but in the other noun cases (except Acc), it is marked with [-i]. For example *auto-t* is the plural Nom of 'car', while *auto-i-ssa* is the plural In+Ess. Because case markers come after plural markers, [-i] never occurs at a word boundary, so Miaseg never recognizes it as a possible plural marker. This accounts for over 80% of Miaseg's errors on Finnish.

The reason Miaseg is able to achieve such high accuracy on Mongolian is that the Unimorph data from which it was derived only contains nouns with a single affix, which marks 1 of 7 cases (Gen, Acc, Dat, Abl, Ins, Com, Voc). Thus, once the model has been exposed to sufficient nouns to have seen all allomorphs of those case suffixes, it is able to achieve perfect segmentation of the limited set of nouns. In contrast, all other evaluation languages have chains of multiple affixes in their respective datasets. We note though, that the simplicity of the task for Mongolian is also true for Morfessor and Transformer, which never achieve the same performance on Mongolian.

A few randomly-selected example segmentations are shown in Tab. 3 (we excluded Mongolian since the data only contained single affixes). The first example from Turkish, where Miaseg segmented *gazetelerinizi* 'newspapers-PL-PSS2P-ACC' as *gazete-ler-iniz-i* demonstrates that Miaseg is able to segment multiple affixes, having inferred that plurality is marked first and case last.

3.3.1 Error Analysis

We performed error analysis of Miaseg for each language at the training size of 10K. In Finnish, > 99% of the errors are due to failing to find a match for a suffix, probably due to some suffixes not occurring at a word boundary. As discussed above, this aspect of the non-Nom Pl allomorphs led to 80% of Miaseg's errors on Finnish.

For Hungarian, 58% of errors are of the same type as Finnish. 26% of the errors involve shifting a morpheme boundary to the left (e.g., *tolvaj-a* vs. **tolva-ja*) and 16% are due to shifting a morpheme boundary to the right (e.g., *ezán-jaik* vs. **ezánj-aik*). For Turkish, >0.99% of errors involve shifting a morpheme boundary to the left.

The relative prevalence of errors involving shifting a morpheme boundary to the left is likely because Miaseg considers the forms that a feature has been marked with (6; step 2) in descending order of length. Thus, if two forms match (e.g., both *aj* and *a* are allomorphs of the PSS3S;SG suffix and match the end of *tolvaja* 'thief-PSS3S;SG'), the longer will be chosen. If the shorter was the correct form, the morpheme boundary is effectively shifted left.

These error patterns suggest that promising areas for improvement would be handling affixes not appearing at word boundaries and improving the heuristic preference for the longest matching marking during segmentation (6; step 2). Note that this analysis considered errors at the word level, meaning that we identified one of potentially multiple reasons for each incorrectly-segmented word. Thus, of the errors in Finnish (>0.99%) and Hungarian (58%) attributed to failing to find a match for a suffix, it is possible that some also had morpheme boundaries shifted left or right.

4 Prior Work

Unsupervised segmentation methods include Minimum Description Length (MDL) models (e.g., Goldsmith 2001). A prevelant model, at least as a baseline, is Morfessor (Creutz and Lagus, 2002) and derviations of it (Creutz and Lagus, 2005, 2007; Virpioja et al., 2013). Bayesian models are often successful, though many were developed in the context of word segmentation (e.g., Goldwater et al. 2009). Neural models have also been employed, usually using a self-supervised task like segmental language modeling for training (Sun and Deng, 2018; Downey et al., 2022; Wang and Zheng, 2022).

Like Miaseg, some prior unsupervised approaches explicitly model morphological paradigms (Goldsmith, 2001; Xu et al., 2018, 2020). Moreover, we are not the first approach to consider meaning, along with form, for segmentation. Prior approaches learn word embeddings to represent semantic information through distributional information (Schone and Jurafsky, 2001; Soricut and Och, 2015; Narasimhan et al., 2015). In contrast, we use morphological features from Unimorph (McCarthy et al., 2020), not word embeddings, which can be data-intensive to train.

Some models attempt to achieve broad typological coverage. For instance, Morfessor (Creutz and

Table 2: F1 (harmonic mean of precision and recall) and accuracy of models. MIASEG, which is our model, outperforms MORFESSOR, which is unsupervised and cannot make use of morphological features, on all datasets and data sizes. Moreover, on 3/4 datasets, MIASEG outperforms TRANSFORMER, which trains in a supervised fashion on both ground-truth segmentations and morphological features.

		500		1000		10000	
		F1	Acc	F1	Acc	F1	Acc
Fin	MIASEG	0.57 ± 0.03	**0.48 ± 0.03**	0.69 ± 0.04	**0.61 ± 0.04**	0.79 ± 0.00	0.72 ± 0.00
	MORFESSOR	0.27 ± 0.03	0.18 ± 0.02	0.27 ± 0.02	0.17 ± 0.01	0.19 ± 0.01	0.05 ± 0.00
	TRANSFORMER	**0.63 ± 0.04**	**0.48 ± 0.05**	**0.73 ± 0.03**	**0.61 ± 0.04**	**0.90 ± 0.03**	**0.83 ± 0.05**
Hun	MIASEG	0.41 ± 0.05	**0.32 ± 0.05**	**0.63 ± 0.07**	**0.56 ± 0.07**	**0.94 ± 0.01**	**0.93 ± 0.02**
	MORFESSOR	0.19 ± 0.05	0.12 ± 0.03	0.32 ± 0.04	0.18 ± 0.03	0.32 ± 0.01	0.13 ± 0.01
	TRANSFORMER	**0.48 ± 0.03**	0.27 ± 0.04	0.61 ± 0.02	0.43 ± 0.03	0.82 ± 0.06	0.72 ± 0.09
Mon	MIASEG	**0.99 ± 0.00**	**0.99 ± 0.00**	**0.99 ± 0.00**	**0.99 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**
	MORFESSOR	0.55 ± 0.03	0.48 ± 0.04	0.49 ± 0.03	0.39 ± 0.03	0.91 ± 0.01	0.89 ± 0.02
	TRANSFORMER	0.79 ± 0.04	0.73 ± 0.05	0.93 ± 0.02	0.90 ± 0.03	0.98 ± 0.01	0.97 ± 0.01
Tur	MIASEG	**0.83 ± 0.00**	**0.81 ± 0.00**	**0.94 ± 0.01**	**0.92 ± 0.01**	**0.96 ± 0.00**	**0.94 ± 0.00**
	MORFESSOR	0.47 ± 0.04	0.32 ± 0.04	0.46 ± 0.03	0.30 ± 0.03	0.54 ± 0.01	0.36 ± 0.01
	TRANSFORMER	0.75 ± 0.03	0.60 ± 0.04	0.86 ± 0.03	0.77 ± 0.05	0.94 ± 0.01	0.90 ± 0.03

Table 3: A few randomly-selected segmentations from MIASEG.

	Word & Features	Predicted	Expected	
Tur	*gazetelerinizi* (PL;PSS2P;ACC)	*gazete-ler-iniz-i*	*gazete-ler-iniz-i*	✓
Fin	*ilmaperspektiivein* (INS;PL)	*ilmaperspektiive-in*	*ilmaperspektiive-in*	✓
Hun	*hátraküldésünk* (PSS1P;SG)	*hátraküldés-ünk*	*hátraküldés-ünk*	✓
Fin	*eristysselleillä* (PL;AT+ESS)	*eristyssellei-llä*	*eristysselle-i-llä*	✗
Tur	*mikroorganizmalardan* (PL;ABL)	*mikroorganizma-lar-dan*	*mikroorganizma-lar-dan*	✓

Lagus, 2002, 2005, 2007; Virpioja et al., 2013) can easily be applied to data from any language. Xu et al. (2020) directly leverage typology by incorporating a diverse range of morphological processes beyond affixation. The resulting model leads to strong results across typologically and phylogenetically diverse languages.

Other approaches have focused on particular typologically or phylogenetically related groups of languages. Pan et al. (2020) proposed an approach to segmenting agglutinative languages for the task of machine translation. Moeng et al. (2021) developed supervised and unsupervised approaches for morphological segmentation of Nguni Languages. Downey et al. (2022) demonstrated that training a neural model in a self-supervised task on ten Indigenous languages of the Americas that are typologically related but phylogenetically unrelated can transfer to a target language, K'iche'.

Our work is in line with the latter group, as we focus on agglutinative morphology. We believe there are merits to both approaches. While typological coverage is an important goal, we believe focusing on mechanisms that may be useful for particular kinds of morphological structure is also of value, since languages can differ dramatically in their morphological structure. For instance, we should not necessarily expect the acquisition of agglutinative and templatic morphological processes to involve precisely the same mechanisms.

5 Conclusion and Discussion

In this work, we have proposed a model for unsupervised but morphological-feature-informed segmentation of agglutinative morphology. Our proposed model, MIASEG, takes advantage of the fact that in agglutinative morphology, a single morpheme tends to correspond to a single feature. Thus, by identifying closely-related pairs of words—i.e. words where one has exactly one fea-

ture more than the other—and inferring the concatenative difference between them, the model is able to discover the ways in which morphological features are marked. These markings are effectively the allomorphs of a given morpheme.

When trained in low resource settings of 500, 1000, or 10000 words, MIASEG achieved reasonably high accuracy and F1 scores across Finnish, Hungarian, Mongolian, and Turkish. Moreover, MIASEG outperformed the unsupervised model MORFESSOR, which operates over bare surface forms—demonstrating the value of morphological features. In a majority of settings, MIASEG also outperformed a supervised neural model that was able to exploit the same features. This suggests that MIASEG, while a simple approach, can outperform a supervised model in low-resource settings.

We find the results to be encouraging for our proposed approach to agglutinative morphology, though we acknowledge that much of the approach would require work to extend to many types of non-agglutinative morphology.

In particular, MIASEG exploits the fact that in agglutinative morphology, each morpheme tends to mark a single feature. In contrast, fusional morphological processes mark multiple features with a single morpheme, leading to its own set of learning challenges. Moreover, morphological processes also include non-concatenative stem changes, reduplication, and templatic processes.

Even among concatinative operations, Xu et al. (2020, p. 6673) point out that some languages have affixes that never appear at a word edge because the affix is always followed or preceded by another affix. Because our method depends on identifying concatenative differences between word forms that differ in a single marked feature, our model would need to be extended in order to discover such affixes. We saw this issue in Finnish, where MIASEG achieved its lowest performance due to some plural allomorphs never occurring at the edge of a word.

Our approach to segmentation takes inspiration from findings in child language acquisition (§ 2.1). We have proposed that if a learner knows which morphological features are marked in a language, the learner can use this information to identify morpheme boundaries in an approach like the one we have proposed. We intend the model for practical use in low-resource, agglutinative morphological segmentation settings and not as an acquisition model. That said, the fact that the approach is inspired by considerations of acquisition and is reasonably effective makes it somewhat tantalizing to conjecture that a similar mechanism might be at play when children acquire agglutinative morphological processes. In future work, we plan to investigate this proposal more directly.

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.

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.

Caleb Belth. 2023a. Towards a learning-based account of underlying forms: A case study in Turkish. In *Proceedings of the Society for Computation in Linguistics 2023*, pages 332–342, Amherst, MA. Association for Computational Linguistics.

Caleb Belth. 2023b. *Towards an Algorithmic Account of Phonological Rules and Representations*. Ph.D. thesis, University of Michigan.

Caleb Belth. 2024. A Learning-Based Account of Phonological Tiers. *Linguistic Inquiry*, pages 1–63.

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.

Eve V Clark. 2014. The principle of contrast: A constraint on language acquisition. In *Mechanisms of language acquisition*, pages 1–33. Psychology Press.

Çagri Çöltekin. 2010. A freely available morphological analyzer for turkish. In *LREC*, volume 2, pages 19–28.

Çagri Çöltekin. 2014. A set of open source tools for turkish natural language processing. In *LREC*, pages 1079–1086.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised

cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. 2009. *Introduction to algorithms*. MIT press.

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 - Volume 6*, MPL '02, page 21–30, USA. Association for Computational Linguistics.

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.

C. Downey, Shannon Drizin, Levon Haroutunian, and Shivin Thukral. 2022. Multilingual unsupervised sequence segmentation transfers to extremely low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5331–5346, Dublin, Ireland. Association for Computational Linguistics.

John Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational linguistics*, 27(2):153–198.

Sharon Goldwater, Thomas L Griffiths, and Mark Johnson. 2009. A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

Péter Halácsy, András Kornai, Németh László, Rung András, István Szakadát, and Trón Viktor. 2004. Creating open language resources for hungarian.

Yun Jung Kim and Megha Sundara. 2021. 6–month–olds are sensitive to english morphology. *Developmental science*, 24(4):e13089.

Jordan Kodner, Salam Khalifa, Khuyagbaatar Batsuren, Hossep Dolatian, Ryan Cotterell, Faruk Akkus, Antonios Anastasopoulos, Taras Andrushko, Aryaman Arora, Nona Atanalov, Gábor Bella, Elena Budianskaya, Yustinus Ghanggo Ate, Omer Goldman, David Guriel, Simon Guriel, Silvia Guriel-Agiashvili, Witold Kieraś, Andrew Krizhanovsky, Natalia Krizhanovsky, Igor Marchenko, Magdalena Markowska, Polina Mashkovtseva, Maria Nepomniashchaya, Daria Rodionova, Karina Scheifer, Alexandra Sorova, Anastasia Yemelina, Jeremiah Young, and Ekaterina Vylomova. 2022. SIGMORPHON–UniMorph 2022 shared task 0: Generalization and typologically diverse morphological inflection. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 176–203, Seattle, Washington. 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, Sami Virpioja, Ville Turunen, and Krista Lagus. 2010. 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. ACL.

Enikő Ladányi, Ágnes M Kovács, and Judit Gervain. 2020. How 15-month-old infants process morphologically complex forms in an agglutinative language? *Infancy*, 25(2):190–204.

Peter Makarov and Simon Clematide. 2018. Imitation learning for neural morphological string transduction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2877–2882, Brussels, Belgium. Association for Computational Linguistics.

Alexandra Marquis and Rushen Shi. 2012. Initial morphological learning in preverbal infants. *Cognition*, 122(1):61–66.

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*, pages 3922–3931. European Language Resources Association.

Toben H Mintz. 2013. The segmentation of sublexical morphemes in english-learning 15-month-olds. *Frontiers in psychology*, 4:24.

Tumi Moeng, Sheldon Reay, Aaron Daniels, and Jan Buys. 2021. Canonical and surface morphological segmentation for nguni languages. In *Southern African Conference for Artificial Intelligence Research*, pages 125–139. Springer.

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.

Yirong Pan, Xiao Li, Yating Yang, and Rui Dong. 2020. Morphological word segmentation on agglutinative languages for neural machine translation. *arXiv preprint arXiv:2001.01589*.

Sarah Payne. 2022. *When collisions are a good thing: the acquisition of morphological marking*. Bachelor's thesis, University of Pennsylvania.

Sarah Payne. 2023. Contrast, sufficiency, and the acquisition of morphological marking. In *Proceedings of BUCLD*, volume 47, pages 604–617.

Ben Peters and Andre F. T. Martins. 2022. Beyond characters: Subword-level morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 131–138, Seattle, Washington. Association for Computational Linguistics.

Patrick Schone and Dan Jurafsky. 2001. Knowledge-free induction of inflectional morphologies. In *Second meeting of the north american chapter of the 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, Berlin, Germany. Association for Computational Linguistics.

Radu Soricut and Franz Josef Och. 2015. Unsupervised morphology induction using word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1627–1637.

Zhiqing Sun and Zhi-Hong Deng. 2018. Unsupervised neural word segmentation for Chinese via segmental language modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4915–4920, Brussels, Belgium. Association for Computational Linguistics.

Kei Uchiumi, Hiroshi Tsukahara, and Daichi Mochihashi. 2015. Inducing word and part-of-speech with pitman-yor hidden semi-markov models. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1774–1782.

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

Lihao Wang and Xiaoqing Zheng. 2022. Unsupervised word segmentation with bi-directional neural language model. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(1).

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2021. Applying the transformer to character-level transduction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1901–1907, Online. Association for Computational Linguistics.

Hongzhi Xu, Jordan Kodner, Mitch Marcus, and Charles Yang. 2020. Modeling morphological typology for unsupervised learning of language morphology. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6672–6681.

Hongzhi Xu, Mitch Marcus, Charles Yang, and Lyle Ungar. 2018. Unsupervised morphology learning with statistical paradigms. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 44–54.

A Example Appendix

In Tab. 4, we provide the precision and recall values for the models; these values went in to the computation of F1 scores in Tab. 2.

Table 4: Precision and Recalls for models. These correspond to the F1 scores in Tab. 2.

		500		1000		10000	
		P	R	P	R	P	R
Fin	MIASEG	**0.67 ± 0.03**	0.50 ± 0.03	**0.77 ± 0.03**	0.62 ± 0.04	0.84 ± 0.00	0.74 ± 0.00
	MORFESSOR	0.31 ± 0.03	0.24 ± 0.02	0.27 ± 0.02	0.28 ± 0.02	0.14 ± 0.01	0.28 ± 0.00
	TRANSFORMER	0.63 ± 0.04	**0.63 ± 0.05**	0.73 ± 0.03	**0.73 ± 0.03**	**0.89 ± 0.04**	**0.90 ± 0.03**
Hun	MIASEG	0.48 ± 0.05	0.35 ± 0.05	**0.69 ± 0.06**	0.59 ± 0.07	**0.95 ± 0.01**	**0.94 ± 0.02**
	MORFESSOR	0.24 ± 0.05	0.16 ± 0.04	0.30 ± 0.04	0.34 ± 0.04	0.26 ± 0.01	0.43 ± 0.01
	TRANSFORMER	**0.49 ± 0.04**	**0.46 ± 0.02**	0.62 ± 0.02	**0.60 ± 0.03**	0.83 ± 0.07	0.82 ± 0.06
Mon	MIASEG	**0.99 ± 0.00**	**0.99 ± 0.00**	**0.99 ± 0.00**	**0.99 ± 0.00**	**1.00 ± 0.00**	**1.00 ± 0.00**
	MORFESSOR	0.50 ± 0.04	0.61 ± 0.02	0.41 ± 0.03	0.60 ± 0.01	0.89 ± 0.02	0.94 ± 0.01
	TRANSFORMER	0.80 ± 0.03	0.78 ± 0.05	0.93 ± 0.02	0.93 ± 0.02	0.98 ± 0.01	0.98 ± 0.01
Tur	MIASEG	**0.85 ± 0.00**	**0.81 ± 0.00**	**0.94 ± 0.01**	**0.93 ± 0.01**	**0.96 ± 0.00**	**0.96 ± 0.00**
	MORFESSOR	0.48 ± 0.04	0.47 ± 0.04	0.44 ± 0.04	0.48 ± 0.02	0.58 ± 0.01	0.50 ± 0.01
	TRANSFORMER	0.75 ± 0.03	0.75 ± 0.02	0.86 ± 0.03	0.86 ± 0.03	0.94 ± 0.01	0.94 ± 0.01