VORM: Translations and a constrained hypothesis space support unsupervised morphological segmentation across languages

Anonymous ACL submission

Abstract

This paper introduces VORM, an unsupervised morphological segmentation system, leveraging translation data to infer highly accurate morphological transformations, including less-frequently modeled processes such as infixation and reduplication. The system is evaluated on standard benchmark data, as well as on a novel dataset of 37 typologically diverse languages. In both cases, its results compare favourably to other unsupervised systems.

1 Introduction

While supervised neural models achieve near-ceiling performance on morphological segmentation (Batsuren et al., 2022), unsupervised systems leave ample room for improvement, despite substantial progress over the years (Virpioja et al., 2013; Narasimhan et al., 2015; Eskander et al., 2020; Xu et al., 2020). While supervised techniques can be used for several dozen languages, corpus data and word lists are available for many more — progress on unsupervised learning is thus desirable to improve the cross-linguistic applicability of such systems. The downstream benefit of morphological segmentation for training language models has been debated (Sälevä and Lignos, 2023), but good morphological segmentation can also support linguistic insight: training and applying a good unsupervised morphological segmentation procedure to study patterns in massively parallel corpora (Liu et al., 2023) can help identify functional morphemes, such as tense and case, across languages, and may be a component of semi-automated interlinear-glossing methods (McMillan-Major, 2020).

Especially with the latter, practical, goals in mind, **Contribution #1** of this paper is an unsupervised morphological segmentation system that leverages parallel translation data and best-first heuristics inspired by Lignos (2010) to severely constrain the hypothesis space. This allows it to infer the applicability of a broader array of morphological processes (infixation, stem change, reduplication) while maintaining high precision. The system outperforms, for some metrics that more closely reflect canonical than surface segmentation, state-of-the-art unsupervised morphological models on canonical segmentation across two benchmark tests, Morphochallenge 2010 (Kurimo et al., 2010) and the SIGMORPHON 2022 task on morphological segmentation (Batsuren et al., 2022).

With those linguistic goals in mind, evaluation on a more diverse set of languages is further desirable. The two benchmark testsets reflect only a small part of the diversity in morphological typology, with extremely common processes, like reduplication (Todd et al., 2022), not represented among them. Furthermore, all languages come from the Eurasian continent, thus reflecting an areally narrow set of languages. **Contribution #2** of this paper is to present a method of using a corpus of interlinearly-glossed fieldwork data in 37 typologically and areally more diverse languages (Seifart et al., 2024) to generate (both supervised and unsupervised) training data as well as evaluation data with a reproducible training/development/test split.

Materials for the project are at https://osf.io/bew3q/?view_only=259303d2c7814c4b9566f997dcbb1d7e. After further introducing the backgrounds to this work (§2), I will introduce the novel system (§3) and the cross-linguistic data (§4). The experimentation will be set out in §5, with its empirical results in §6.

2 Background

2.1 Unsupervised morphological segmentation

The Morfessor model (Virpioja et al., 2013) is the de facto baseline for unsupervised morphological

segmentation. It leverages word-internal statistical patterns of character sequences, similarly to Byte Pair Encoding (Gage, 1994), commonly used to preprocess text for training language models. Both techniques lead to surface segmentations of the input string. A recent, linguistically inspired, model that leads to surface segmentations is (Eskander et al., 2020), which trains Adaptor Grammars (Johnson et al., 2006) on surface strings, representing morphological segmentation as a context-free grammar parsing problem.

Other unsupervised models leverage the insight that morphological processes do not merely carve up a surface string, but transform base forms into derived forms, that are often not just superstrings of the base form – transforming *believe* into *believing* requires dropping the *e*. Modeling such processes accurately would allow us to represent the canonical segmentation (Kann et al., 2016) of a surface string, i.e., recognizing that *believe* in the (surface segmented form) *believ+ing* is the same canonical morpheme as in *believe+s*.

An early exponent of this class of models is Morsel (Lignos, 2010), which uses a best-first heuristic that maximizes the data coverage of the inferred transformations, leading to derivations consisting of chains of transformations. A similar idea, but leveraging more global optimization over the search space of transformations can be found in Morphochains (Narasimhan et al., 2015) and Morphoforests (Luo et al., 2017). Like Morphoforests, ParaMA2 (Xu et al., 2020) explicitly considers paradigms, groups of transformations that co-occur as a further building block to their model, on top of using the idea that transformations form chains.

Here, I adopt many of the premises of the cited works: leveraging heuristics, considering word pairs and paradigms as ways to constrain the search space, and representing morphological processes as transformations.

2.2 Leveraging translations

Parallel translation data has, in several domains, been proven to help guide (otherwise) unsupervised models towards the right sectors of the hypothesis space. Most pertinently, Rice et al. (2024) use translations of a target language to a reference language to provide an additional semantic signal in a supervised system, in similar ways to (Narasimhan et al., 2015) and Schone and Jurafsky (2001), to determine morphological segmentation: formally overlapping words in the target language translating to the same or semantically similar words in the reference language are thus more likely to be segmented similarly.

Beyond morphology, translation data has been used to project structure of a better-resourced reference language to a target language – examples are PoS tagging and grammatical structure (Johannsen et al., 2016). Word-sense disambiguation has been shown to benefit from using translation data, given that distinct senses often translate differently (Apidianaki, 2008; Hauer and Kondrak, 2023). Shared between all cases, is the idea that a reference language provides insight in the latent structure (semantic distinctions, grammatical relations, shared morphological material) of the target language, either through the projection of that structure or through the variation in the patterns of translation themselves. My approach leverages this latter type of signal.

2.3 Morphological typology

When we approach unsupervised morphological segmentation as a task of being able to induce *for any language* the morphological segments, canonical or superficial, without having access to the correct segments to train on, considering the variation in morphological processes is of relevance. A typologically-oriented overview is (Haspelmath and Sims, 2010), who draws on the distinction between free morphemes (which can occur as a word by themselves) and bound morphemes (which cannot) to list the following basic processes.

First,**affixation** involves concatenating bound morphemes to a free morpheme, such as *believe + -ing*. This includes infixation, whereby a bound morpheme is located inside the free morpheme – such as the Tagalog 'agent trigger' morpheme *-um-* forming *s-um-alat* 'wrote' out of *salat* 'write'. Next, **compounding** involves concatenating two or more free morphemes, like *boathouse* from *boat* and *house*. Third, **reduplication** means reproducing a part of a free morpheme on either end of that morpheme – marginal in English *house house* 'a real house', but widely productive in other languages, e.g. *duhp* 'dive' → *du-duhp* 'be diving' (Ponapean). Fourth, **base modification** involves changing the string 'inside of' the free morpheme, like Germanic ablaut – *gave* as the past tense of *give*. Finally, **conversion**: leaving the form unaltered but changing e.g., the grammatical category, e.g., *hammer* as a noun, converted into a verb;

Given this diversity, the focus on (non-

reduplicative) affixation is narrow. Reduplication is, for instance, extremely common: >80% of languages are described to have some form of reduplication (Rubino, 2013). A smaller proportion of languages has stem-internal modifications such as ablaut (vowel change) or tone change (Bickel and Nichols, 2013). (Yu, 2007) finds infixation in 111 languages of 26 language families. Various forms of base modification similarly happen across the world's languages: Standard Arabic has stem-internal gemination as the morphological causative (*waqafa* 'stop (intransitive)' → *waqqafa* 'stop (transitive)').

Surface segmentation models such as Morfessor and MorphAGram inherently rule out infixation and base modification, and typically don't provide ways of identifying reduplication as distinct from regular affixation (but see Todd et al., 2022 for an extention of Morfessor doing exactly that). Most models of canonical segmentation do not consider processes of reduplication and base modification, with notable exceptions being ParaMA2 (Xu et al., 2020). The present work intends to develop this line of research.

3 The VORM model

The proposed model, VORM ('Vertaling Ondersteunt Redelijke Morphologie' – Dutch for 'Translation supports reasonable morphology') is a heuristic system that leverages translation equivalency in a reference corpus to find an initial set of morphological transformations, which it then applies more broadly. The model consists of three steps: **Determining potential morphological families**, which guide the **Learning of productive morphological transformations**. Third, the learned transformations are applied beyond the potential morphological families in a **Propagation to the full word list** step.

3.1 S1: Determining morphological families

One recurrent challenge in unsupervised systems that use word pairs (Narasimhan et al., 2015; Xu et al., 2020) is to avoid oversegmentation. Recurrent phonotactic/orthographical patterns may give the suggestion of a morphological transformation where there isn't one. (Narasimhan et al., 2015) use distributional semantic information to nudge the model away from unrelated pairs and towards related pairs, building on the insight of (Schone and Jurafsky, 2001) that distributional semantic representations link morphological variants. Here, I propose to use another way to constrain the comparison: translations, available for many languages.

The general procedure is as follows: we consider a bitext B of translations between t and a reference language r, defined as $B = [\langle u_r^1, u_t^1\rangle, \langle u_r^2, u_t^2\rangle, \ldots \langle u_r^n, u_t^n\rangle]$, meaning that B consists of an ordered list of paired utterances $\langle u_r, u_t\rangle$ that are translation equivalent utterances. Let further the utterances $u_l^1 \ldots u_l^n$ for a language l be made up of words from some vocabulary V_l.

The objective is to retrieve sets of word types in t that are likely morphologically related to each other, to feed into the next step. We call such a set a 'morphological family' (cf. Nagy et al., 1989), denoted $m \in M$ where M is the set of morphological families found. Several functions could be defined mapping the bitext B onto the set of morphological families M – standard alignment procedures might be used, were it not for the fact that morphologically rich target languages have a long tail of morphologically complex hapax legomena which risk not getting accurately aligned.

Instead, I designed this procedure by integrating the forward step of the LIU Conceptualizer model, which, given a seed word w_r in r iteratively finds character substrings $[c_t^1, c_t^2, \ldots c_t^n]$ of words in t whose distribution across the utterances in B is statistically most strongly associated with the distribution of w_r. Each such substring c_t defines a morphological family m as all word types $w_t^1, w_t^2, \ldots w_t^n$ that (1) contain c_t as a substring, and (2) occur in an utterance u_t^i whose aligned counterpart in r, u_r^i contains the seed word w_r.

Examples of families for two languages, using the seed language (Vietnamese) and corpora introduced below, are given in Table 1. Vietnamese *cảm* 'feel' has two c_t: *$danke$* (where $ denotes a word boundary) and *fuehl*. The former has an m containing only *danke* itself, whereas the latter matches several dozens words in the bitext lines it co-occurs in with *cảm*, all containing the *fuehl* stem. Vietnamese *cần* 'need', similarly has two associated substrings in Turkish, *$ihtiya* and *$gerek*, each with large, and mostly morphologically related, morphological families.

3.2 S2: Learning productive transformations

The morphological families are next used to learn productive generalizations. This procedure closely follows Morsel (Lignos, 2010). Step 2 starts with initializing a set F of candidate transformations

language	w_r	c_t	m
German	cảm	$danke$	danke
	cảm	fuehl	bauchgefuehl ehrgefuehl f fuehl fuehle fuehlen fuehlich fuehlst fuehlt fuehlte fuehlten gefuehl gefuehle gefuehlen gefuehllos (40 more)
Turkish	cần	$ihtiya	ihtiyaC ihtiyaClar ihtiyaClarI ihtiyaClarInI ihtiyaClarInIn ihtiyaClarInIz ihtiyaClarIna ihtiyaCtan ihtiyac ihtiyacI ihtiyacIm ihtiyacImIz (30 more)
	cần	$gerek	gerek gerekCe gerekCelerle gerekCemi gerekebilecek gerekebilir gerekecek gerekecektir gereken gerekenden gerekenler gerekenlerden (100 more)

Table 1: Examples of extracted morphological families. Orthography follows the Morphochallenge 2010 format.

$f_1, f_2, \ldots, f_n$. The procedure iterates over all $m \in M$. For each m, each 2-permutations of words (w_t^i, w_t^j) in m is considered. All transformations build from a set of allowed transformation F_{all} that transform w_t^i into w_t^j are added to F.

F_{all} is defined to represent the typological diversity of morphological processes. The following are the allowed transformations on the right edge of the string; symmetrical counterparts are defined for the left edge (**prefixation (with assimilation)**, **full/partial-V/partial-C left reduplication**, resp. **left infixation**):

Sufixation: add characters to the right edge of w_t^i so that the result is w_t^j. For instance: *belief-beliefs* is modeled by *-s* suffixation;

Suffixation with assimilation: remove 1 or 2 characters from the right edge of w_t^i and then add any string of characters to the (new) right edge, so that the result is w_t^j: *believe-believing* is modeled by *-e/ing* suffixation;

Full right reduplication; a string of length n on the right edge of w_t^i is suffixed to w_t^i to form w_t^j: Fanbyak *ini-inini* 'to shoot' are modeled by full right reduplication of *ñi*;

Partial-V right reduplication; all strings of one or more vowels[1] in w_t^i and w_t^j are replaced by a wildcard symbol '@', forming the new strings $w_t^{i'}$ and $w_t^{j'}$. Next, a string s of the length n on the right edge of $w_t^{i'}$ is suffixed to $w_t^{i'}$ to form $w_t^{j'}$: Gorwaa *guus-guusas* are modeled this way, reduplicating the final consonant, preceded by an 'a')l

Partial-C right reduplication; all strings of one or more consonants in w_t^i and w_t^j are replaced by the rightmost consonant in the string, forming the new strings $w_t^{i'}$ and $w_t^{j'}$. Next, a string s of the length n on the right edge of $w_t^{i'}$ is suffixed to $w_t^{i'}$ to form $w_t^{j'}$. Partial-C left reduplication is more common: Pangasinan (Rubino, 2001) transforms *plato* 'plate' into *paplato* 'plates' by taking the leftmost single consonant and vowel of a string and adding them to the left edge of that string.

Right infixation; for a pair of words w_t^i and w_t^j, removing a string s^i of length n from an anchor a in w_t^i results in a new string $w_t^{i'}$, and removing a string s^j of length m from the same anchor a in w_t^j results in a string $w_t^{j'}$. If $w_t^{i'}$ is identical to $w_t^{j'}$, the pair of words is modeled by a-anchored right infixation. Anchors are structural positions in the orthographic string constraining where the infix is combined (Yu, 2007), and I use 4 here: before vs. after the last consonant cluster, and before vs. after the last vowel cluster. English *give-gave* are modeled by replacing s^i = 'i' for s^j = 'j', given that $w_t^{i'}$=$w_t^{j'}$='gve', anchored on a = before-last-consonant-cluster.

Next, a best-first heuristic extracts a set of productive transformations F_p from F. The intuition here is that a productive morphological transformation is one that models many word pairs. Let P be the set of all word pairs (w_t^i, w_t^j) such that there is at least one morphological family m for which $w_t^i \in m \wedge w_t^j \in m$, and P_f all such word pairs modeled by a transformation f. We then define the best transformation $f_{\text{best}} = \arg\max_f |P_f|$.[2] Once f_{best} is found, the word pairs in $P_{f_{\text{best}}}$ are removed from P, as are all other word pairs whose second word is modeld by f_{best}. The procedure is repeated until $|P_{f_{\text{best}}}|$ falls below a threshold θ_f.

The derivations found through the best-first heuristic afford two sources of constraints on the application of F_p in the full vocabulary. First, derivations form **chains**: *bookings* may have been

[1] Vowels are defined as all characters that through the Python library unidecode become one of 'a', 'e', 'i', 'o', 'u', 'y'. Consonants are defined as any other character.

[2] Ties are broken first by morphological type, where the ordering given above is followed, then by affix length (longer affixes are preferred).

derived from *booking* with *-s* suffixation, after which *booking* was derived from *book* through *-ing* suffixation. We denote the chain or derivation d as $\langle$*-ing*, *-s*$\rangle$, and we collect all attested chains of transformations. Secondly, chains co-occur with other chains – this can similarly help prevent oversegmentation in ways set out below. For now, we define a pair of chains of transformations d_i, d_j to **co-occur** if there is at least one base form that models some w_i through d_i and some other w_j through d_j.

Finally, an orthogonal procedure allows us to find **compounds**, using the morphological families. We do so by inferring a set of compound templates, strings of n elements. The template consists of $n-1$ fixed elements, and a blank spot where another word $w_t \in V_t$ can go. We find the set of **reliable compound templates** by iterating over all $m \in M$. For each word $w \in m$, we find all of its exhaustive splits w^i, w^j for which $w^i \in V_t \wedge w^j \in V_t$ and $w^i \in m \vee w^j \in m$. The latter constraint provides evidence that this is indeed a compound. For example, *bauchgefuehl* in Table 1 yields two potential compound patterns $\langle$*bauch* + _$\rangle$ and $\langle$_ + *gefuehl*$\rangle$, as both *bauch* 'belly' $\in V_f$ and *gefuehl* $\in V_f$, with the latter moreover being part of m as well (as can be seen in the table). If a pair w^i, w^j is found that forms a reliable compound template, we recursively apply the procedure to each element of the pair to see if further splits can be found. The count of the reliable compound templates is tracked across M, and all reliable compound templates with a frequency of θ_c or greater are kept to constrain compounding in Step 3.

3.3 S3: Propagation to the full word list

The derivations obtained in Step 2 are typically accurate, but only capture a small part of a language's vocabulary. First, not all morphologically related words in the bitext are found in the same morphological family m, but perhaps more importantly, we would like the unsupervised model to be able to generalize beyond the bitext itself. As such, Step 3 models the propagation of the productive transformations F_p, constrained by the set of chains and chain co-occurrences, to a wordlist L, where L may consist of all words in B, or some external source.

First, for each word $w \in L$, all transformations chains that can apply to it are extracted and added to a set of potential analyses $A(w)$ of w. A chain $d = \langle f_1, f_2, \ldots, f_n \rangle$ is applicable to a word w if, for every transformation f, a new string w' can be derived by removing the string added by f from the previously derived string w, where new strings do not have to be in V_t. The resulting new string after successfully applying d to w is denoted s for stem, and is added to a list of potential stems S.

Every stem $s \in S$ now defines a set of words $D(s) = \{w_i, \ldots, w_n\}$, each of which derives s through the application of a chain d. However, some s with very large $M(s)$ did not reflect coherent morphologically related groups of words. For that reason, we impose a further constraint, such that every derivational chain d modeling the relation between a word $w \in D(s)$ and s has to be found to co-occur, as defined in Step 2, with the derivational chains of $\geq |D(s)-1| \times \frac{1}{2}$ other words $w' \in D(s)$. If this is not the case, the word whose derivation co-occurs with the fewest derivations of the other words of $D(s)$ is removed from $D(s)$. This procedure is repeated until the set consists of one member, or the derivations of all words in $D(s)$ co-occur with $\geq |D(s)-1| \times \frac{1}{2}$ other words $w' \in D(s)$.

The central mechanism of this step is a **best first pass**, similar to Step 2, except the model now iteratively finds the stem s_{best} that models the largest $D(s)$ (with ties broken by stem length, preferring shorter stems). Once found, all words in $D(s_{\text{best}})$ are removed from $D(s')$ for all stems $s' \in S$, and a new s_{best} is determined.

After this pass is done, compounds are extracted over all extracted s_{best} by applying the **reliable compound templates** from Step 2. If the substring s filling the blank is a word in V_t, compounding applies, and the new derivation has more than one stem (potentially each with their own derivations).

4 DoReCo-morph: crosslinguistic data

The representational potential of Vorm, including reduplication and infixation, exceeds the set of morphological phenomena present in the datasets typically used. Reduplication and infixation are absent from widely used benchmark sets such as Morphochallenge 2010 (Kurimo et al., 2010). One dataset that can be fill this gap is DoReCo (https://doreco.huma-num.fr/; Seifart et al., 2024), consisting of 52 collections of transcribed fieldwork materials in the same number of languages. Much of these materials have interlinear glosses, exemplified in Table 2, where for each word, the morphological analysis is given. Such data allow us to

w	melo	bo	lo	ghavilighue.
m	melo	bo	lo	ghavi -li -ghu =e
g	tuna	go	3SG.M	paddle -3SG.M.O -NMLZ =EMPH
f	"he went and fished bonito with it."			

Table 2: Interlinear Gloss; Savosavo (Wegener, 2024)

automatically derive a list of words with their morphological analyses, which in turn can be used to train (un)supervised morphological segmentation systems and evaluate them.

The Supplemental Materials for this paper contain a script for deterministically transforming the corpus data into a dataset in the same format as the Morphochallenge data, with word types linked to their canonical and surface segmentation(s). In particular, the unique words (the **w** layer in Table 2) are linked to all their morphological analyses, represented as combinations of the morphemes (**m**) and the glosses (**g**). An analysis of Savosavo *ghavilighue* would thus be: 'ghavi:paddle -li:3SG.M.O -ghu:NMLZ =e:EMPH'. Some preprocessing to normalize orthography and glossing was applied.

These data can be readily used for computational morphology (and perhaps tasks such as inter-linear gloss induction (**?**) and other multilingually oriented tasks). The script also generates a train/development/test split over the data to facilitate testing. While the derived data cannot be reproduced, their generation is exactly reproducible as long as the corpus remains public. The datasets used, along with relevant statistics on the derived data, are presented in Table 7 in the Appendices. This table also gives the citation for each individual language, required as part of the user agreement of the corpus.

Furthermore, the table presents information on the morphological profile of the 38 languages. Average word lengths across languages range from 4.98 to 14.20 in characters and 1.17 to 3.26 in morphemes, thus representing a broad variety of morphological complexity. While little evidence of (the annotation of) infixation or base modification was found among the languages, reduplication is extensively represented in the corpus: a majority of languages displays reduplication, with some languages having it in over 10% of their word types, underscoring the point of Todd et al. (2022) that reduplication is a phenomenon worth modeling.

5 Evaluation

5.1 Evaluation data and metrics

First, VORM is compared with other models on two extant benchmark sets: Morphochallenge 2010 (MC10; Kurimo et al., 2010), with gold standard data for English, Finnish, Turkish, and German canonical and surface (for all but German) segmentation, and the SIGMORPHON 2022 task on morphological segmentation (SGM22; Batsuren et al., 2022) involving Czech, English, French, Hungarian, Italian, Latin, Mongolian, Russian, and Spanish surface segmentation. Third, we consider the novel DOREC0-MORPH dataset of 37 languages.

The standard metrics were applied: EMMA-2 (Virpioja et al., 2011) and Boundary Precision and Recall (BPR) for the MC10 and DOREC0-MORPH, and the measure capturing morpheme identity instead of boundaries released by Batsuren et al. (2022) for SGM22.

5.2 Experimental set-up

The bitexts used in the experiments varied; for the first two datasets, up to a million words of bitext from Opus2018 (Lison and Tiedemann, 2016) subtitles from television and movies gathered from http://www.opensubtitles.org/ were used. Vietnamese was chosen as the reference language for this experiment as it has little morphology, making the word forms seed items with a broad scope. Bitexts for German and Turkish were orthographically normalized to bring them in line with the test data. For the DOREC0-MORPH experiment, bitexts based on the corpora themselves were generated, using the words as well as the free translations (the **f** layer in Table 2). Most free translations were in English, but other languages were found as well – the free translations were tokenized but otherwise used as-is.

The model was tuned on the development split (12% of the data for each language) in the DOREC0-MORPH data, the training split for MC10 and the development split for SGM22, to find optimal values for the free parameters θ_f (minimum number of word pairs modeled by a transformation in Step 2) and θ_c (minimum number of compound template occurrences for it to be used in Step 3) using a grid search over a set of reasonable values, arriving at $\theta_f = 60, \theta_c = 10$ for the first two datasets and $\theta_f = 10, \theta_c = 10$ for the DOREC0-MORPH data.

	EMMA-2			BPR		
	morf	AG	VORM	morf	AG	VORM
eng	85.9	88.7	**89.1**	75.2	**80.0**	52.9
fin	73.4	77.7	**95.8**	62.8	**71.1**	12.5
ger	80.9	85.9	**95.7**		n/a	
tur	61.3	69.3	**95.6**	64.6	**78.9**	16.1

Table 3: Model comparison on the development sets for Morphochallenge 2010 [MC10] (eng = English, fin = Finnish, ger = German, and tur = Turkish), comparing Morfessor (Morf) and the best MorphAGram (AG) model against VORM on EMMA-2 and BPR F1 scores. The best result per language and per metric is boldfaced.

	DeepSPIN-3	Morfessor2	VORM
ces	93.8	29.4	**37.3**
eng	93.6	**37.7**	35.2
fra	95.7	**22.4**	19.1
hun	98.7	**41.0**	40.7
ita	97.4	9.0	**12.9**
lat	99.4	14.5	**22.2**
rus	99.4	**17.7**	15.8
spa	99.0	**20.6**	15.6
avg	97.3	**25.6**	24.9

Table 4: Model comparison on the tests sets for the SIGMORPHON 2022 challenge, for (ces = Czech, eng = English, fra = French, hun = Hungarian,), comparing Morfessor2 and the best overall supervised model (DeepSPIN-3) model against VORM on the Batsuren et al. (2022) evaluation measure. The best unsupervised result per language and per metric is boldfaced.

5.3 Comparison models

For the MC10 and SGM22, I compare VORM against published results. For DORECO-MORPH, I use published software to run Morfessor2 (Virpioja et al., 2013), ParaMA2 (Xu et al., 2020), MorphAGram (Eskander et al., 2020) (in the language-independent setting) as unsupervised models, and Chipmunk (Cotterell et al., 2015), a supervised statistical model as a popular instance of the class of supervised models. The unsupervised models were trained on the full wordlists, and Chipmunk on the training split (48% of the data), and were tested on the test split (40% of the data). For all models, off-the-shelf parameter settings were used.

6 Results

MorphoChallenge 2010 results. Table 3 presents the results for MC10. On EMMA-2, the VORM model presents an improvement over the best MorphAGram variant for all four languages, with the improvement being substantial for Finnish, German, and Turkish. On BPR, conversely, VORM is substantially outperformed by both Morfessor and MorphAGram. This effect may be due to the differences in what the two measures are picking up on. EMMA-2 favours canonical morpheme identity, but, crucially, does not penalize allomorphy, which is indistinguishable from undersegmentation to the model. The same undersegmentation leads to extremely low (often single digit) recall scores on the BPR measure, thus suppressing the reported F1 scores. To be sure, VORM *does* segment – Finnish *elinalueeltaan* 'from their habitat' is analyzed as 'elinaluei + -i/en + -n/lta + -an', with the reference analysis giving 'elin_N alue_N +ABL +3SGPL'. The model correctly identifies the Ablative case and possessive marker, while missegmenting the location of the morpheme boundary, and undersegmenting the compound 'elin + alue'.

SIGMORPHON 2022 results. For the SGM22, the results are presented in Table 4. While no unsupervised model comes anywhere near the results of the supervised models (here, the best-performing supervised model, DeepSPIN-3 (Peters and Martins, 2022) is given as a reference point), VORM occasionally outranks Morfessor2 in its performance. Like with the BPR measure for the Morphochallenge data, surface segmentation is not the model's strongest suit.

DORECO-MORPH. Finally, let us consider the results on the novel dataset, the DORECO-MORPH data. Table 5 present the aggregated results for VORM and its comparison models over the 37 languages, with Table 8 presenting the F1-scores per language. For the EMMA-2 F1-scores, we see that VORM outperforms the other unsupervised models for 21/37 languages. MorphAGram (AG in the table) is the optimal model for 10 languages. While the supervised Chipmunk model is the best overall model in all but 7 languages, it is notable that it is the VORM model has a higher EMMA-2 score in 6 of those. This result lines up with the findings for MC10, where we found VORM performing well on this metric as well. For surface segmentation, measured with BPR, the performance is more mixed: here, Chipmunk is consistently the best overall model, with Morfessor2, ParaMA2, and VORM each being the best model for a similarly sized set of languages.

	EMMA-2					BPR				
	chip	morf	para	AG	vorm	chip	morf	para	AG	vorm
max?		5	1	10	**21**		7	**18**		12
avg.	91.4	84.6	80.3	84.7	**87.2**	86.9	56.8	**57.0**	34.7	55.6
Q1	89.8	83.1	78.2	80.4	**85.8**	83.5	48.9	**51.8**	30.5	45.9
worst	69.9	**74.8**	67.5	71.6	67.8	65.5	31.3	**35.7**	14.4	33.7

Table 5: Aggregated EMMA-2 & BPR F1 scores for the DoReCo dataset for [chip]munk (supervised), [Morf]essor2, [Para]MA2 , Morph[AG]ram, and VORM. Best unsupervised results in bold; best overall results underlined.

6.1 Reduplication and base modification

word	babarak	vivirigĕm
gold	ba :RED bara:long -k:TAM1	vi :RED virigĕ:rush m:TAM1
chip	babarak	vivirig + ĕm
morf2	babara + k	vivi + rig + ĕm
para	babara + -k	vi_rig + -vi- + -em
AG	babara + k	vivi + rig + ĕm
vorm	ba + bara + -k	vi + virigĕ + -m

Table 6: Examples of reduplication in Vera'a (Schnell, 2024) and their analysis across models. Underscores mark the infix slot; tildes mark reduplicative affixes.

To demonstrate the model's capacity to analyze reduplication, consider the examples in Table 6 with their analyses in the five models. We see that only VORM gets the analysis correct, both in its surface segmentations as well as in its canonical analysis, i.e., recognizing *ba* and *vi* as reduplicative morphemes. Other models either undersegment the left edge of the words, or missegment the word (paraMA, morfessor2).

None of our languages has productive base modification processwa, but German has some, in nominal plurals and past tense. Given the low type frequency of these processes, the tuned model for the Morphochallenge dataset did not learn these patterns, but a model considered during the tuning phase, with $\theta_f = 30$, did analyze *huehnerbesitzer* 'chicken owner' correctly as 'hu_hn + -e- + -er + besitz + -er' and *geldbetraege* 'sums of money' as 'geldbetra_g -e- + -e'

7 Discussion

This paper introduces a novel unsupervised morphological segmentation system, VORM, which uses translation-equivalency to narrow down the set of word pairs on which the inferred morphological transformations are based. Along with affixation, the model has the representational capactity for base-modifying transformations as well as reduplication. The grammar induction takes place through a pair of heuristic, best-first processes. In doing so, the model stands in a tradition of unsupervised morphological segmentation that does not consider very large parts of the hypothesis space (Lignos, 2010; Xu et al., 2020). Rather than an imperfect approximation of some more global optimization, I believe the fact that these models consistently do so well reflects the nature of the induction problem, whereby aspects of the usage of morphological transformations (how many words they model, how large the groups of morphologically related words are modeled by the same stem) that are known to affect learning and processing in humans guide the model to fairly correct answers.

This paper forms the first attempt at using the translation signal; while the morphological families that are inferred through it seem accurate, the heuristic procedure of Steps 2 and 3 introduces error, mostly by undersegmenting, but also, less frequently, by over or missegmenting. We have seem in the experiments that this may lead to an extremely low recall on metrics where the exact morpheme identity (and their boundaries) is at stake, like BPR and the measure of (Batsuren et al., 2022). This seems like the main obstacle for the model to be overcome. Conversely, it's superior performance on the EMMA-2 metric suggests that the model has good potential in identifying, with a high accuracy, the morphemes that linguists would recognize.

Further exploration on the DOBECO-MORPH dataset can prove fruitful in identifying more specific morphological challenges to be modeled. Through such exploration, and more detailed analysis of model performance on different challenges, the landscape of what unsupervised learners have to contend with might become more clear. With this paper, I hope to have made a first move in that direction.

References

Marianna Apidianaki. 2008. Translation-oriented word sense induction based on parallel corpora. In *Language Resources and Evaluation (LREC)*.

Jocelyn Aznar. 2024. Nisvai DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Khuyagbaatar Batsuren, Gábor Bella, Aryaman Arora, Viktor Martinovic, Kyle Gorman, Zdeněk Žabokrtskỳ, Amarsanaa Ganbold, Šárka Dohnalová, Magda Ševčíková, Kateřina Pelegrinová, et al. 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.

Balthasar Bickel and Johanna Nichols. 2013. Fusion of selected inflectional formatives (v2020.4). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Natalia Bogomolova, Dmitry Ganenkov, and Nils Norman Schiborr. 2024. Tabasaran DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Niclas Burenhult. 2024. Jahai DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Alexander Yao Cobbinah. 2024. Baïnounk Gubëeher DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Ryan Cotterell, Thomas Mueller, Alexander Fraser, and Hinrich Schütze. 2015. Labeled morphological segmentation with semi-markov models. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 164–174.

Andrew Cowell. 2024. Arapaho DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Chris Lasse Däbritz, Nina Kudryakova, Eugénie Stapert, and Alexandre Arkhipov. 2024. Dolgan DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Christian Döhler. 2024. Komnzo DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

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

Diana Forker and Nils Norman Schiborr. 2024. Sanzhi Dargwa DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Michael Franjieh. 2024. Fanbyak DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

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

Alexandro Garcia-Laguia. 2024. Northern Alta DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Valentin Gusev, Tiina Klooster, Beáta Wagner-Nagy, and Alexandre Arkhipov. 2024. Kamas DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Tom Güldemann, Martina Ernszt, Sven Siegmund, and Alena Witzlack-Makarevich. 2024. Nng DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Geoff Haig, Maria Vollmer, and Hanna Thiele. 2024. Northern kurdish (kurmanji) doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Iren Hartmann. 2024. Hoocak DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Andrew Harvey. 2024. Gorwaa DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Martin Haspelmath and Andrea Sims. 2010. *Understanding morphology*. Routledge.

Katharina Haude. 2024. Movima DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Bradley Hauer and Grzegorz Kondrak. 2023. One sense per translation. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 442–454.

Birgit Hellwig. 2024. Goemai DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Anders Johannsen, Željko Agić, and Anders Søgaard. 2016. Joint part-of-speech and dependency projection from multiple sources. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–566.

Mark Johnson, Thomas Griffiths, and Sharon Goldwater. 2006. Adaptor grammars: A framework for specifying compositional nonparametric bayesian models. *Advances in neural information processing systems*, 19.

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.

Olga Kazakevich and Elena Klyachko. 2024. Evenki DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Soung-U Kim. 2024. Jejuan DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Manfred Krifka. 2024. Daakie DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

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, Uppsala, Sweden. Association for Computational Linguistics.

Constantine Lignos. 2010. Learning from unseen data. In *Proceedings of the Morpho Challenge 2010 Workshop*, pages 35–38, Helsinki, Finland. Aalto University School of Science and Technology.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Yihong Liu, Haotian Ye, Leonie Weissweiler, Philipp Wicke, Renhao Pei, Robert Zangenfeind, and Hinrich Schütze. 2023. A crosslingual investigation of conceptualization in 1335 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12969–13000.

Jiaming Luo, Karthik Narasimhan, and Regina Barzilay. 2017. Unsupervised learning of morphological forests. *Transactions of the Association for Computational Linguistics*, 5:353–364.

Angelina McMillan-Major. 2020. Automating gloss generation in interlinear glossed text. *Society for Computation in Linguistics*, 3(1).

Ulrike Mosel. 2024. Teop DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

William Nagy, Richard C Anderson, Marlene Schommer, Judith Ann Scott, and Anne C Stallman. 1989. Morphological families in the internal lexicon. *Reading Research Quarterly*, pages 262–282.

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.

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.

Maïa Ponsonnet. 2024. Dalabon DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Juan Diego Quesada, Stavros Skopeteas, Carolina Pasamonik, Carolin Brokmann, and Florian Fischer. 2024. Cabécar DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Sabine Reiter. 2024. Cashinahua DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Enora Rice, Ali Marashian, Luke Gessler, Alexis Palmer, and Katharina Wense. 2024. Tams: Translation-assisted morphological segmentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6752–6765.

Sonja Riesberg. 2024. Yali (apahapsili) doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Hiram Ring. 2024. Pnar DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Françoise Rose. 2024. Mojeño Trinitario DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Carl Rubino. 2001. Pangasinan. In Jane Garry and Carl Rubino, editors, *Encyclopedia of the World's Languages: Past and Present*, pages 539–542. H.W. Wilson Press, New York / Dublin.

Carl Rubino. 2013. Reduplication (v2020.4). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Jonne Sälevä and Constantine Lignos. 2023. What changes when you randomly choose bpe merge operations? not much. In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 59–66.

Stefan Schnell. 2024. Vera'a doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

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*.

Frank Seifart. 2024. Bora DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Frank Seifart, Ludger Paschen, and Matthew Stave. 2024. *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Stavros Skopeteas, Violeta Moisidi, Nutsa Tsetereli, Johanna Lorenz, and Stefanie Schröter. 2024. Urum DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Amos Teo. 2024. Sümi DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Nick Thieberger. 2024. Nafsan (south efate) doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Simon Todd, Annie Huang, Jeremy Needle, Jennifer Hay, and Jeanette King. 2022. Unsupervised morphological segmentation in a language with reduplication. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 12–22.

Martine Vanhove. 2024. Beja DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

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

Sami Virpioja, Ville T. Turunen, Sebastian Spiegler, Oskar Kohonen, and Mikko Kurimo. 2011. Empirical comparison of evaluation methods for unsupervised learning of morphology. *Traitement Automatique des Langues*, 52(2):45–90.

Alexandra Vydrina. 2024. Kakabe DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Claudia Wegener. 2024. Savosavo DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Søren Wichmann. 2024. Texistepec Popoluca DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

Alena Witzlack-Makarevich, Saudah Namyalo, Anatol Kiriggwajjo, and Zarina Molochieva. 2024. Ruuli DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

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.

Alan C. L. Yu. 2007. 67pivot theory and the typology. In *A Natural History of Infixation*. Oxford University Press.

A Detailed data table for DoReCo-morph

This Appendix contains the information on the components of the DoReCo corpus used, presented in Table 7

B Results by language for DoReCo-morph

This Appendix supplements section 6 with the results broken down per language.

glottocode	language	family	area	reference	n types	nC	nM	% rdp
apah1238	Yali	Nuclear Trans New Guinea	PNS	Riesberg (2024)	2474	6.26	1.82	
arap1274	Arapaho	Algic	NAM	Cowell (2024)	4483	14.20	2.37	9.32
bain1259	Baïnounk Gubëeher	Atlantic-Congo	AFR	Cobbinah (2024)	3598	6.59	2.74	0.33
beja1238	Beja	Afro-Asiatic	AFR	Vanhove (2024)	7280	7.40	2.51	1.22
bora1263	Bora	Boran	SAM	Seifart (2024)	10723	9.39	2.52	
cabe1245	Cabécar	Chibchan	NAM	Quesada et al. (2024)	2852	6.24	1.66	
cash1254	Cashinahua	Pano-Tacanan	SAM	Reiter (2024)	2221	7.63	2.22	
dolg1241	Dolgan	Turkic	ERS	Däbritz et al. (2024)	5579	8.04	2.43	0.02
even1259	Evenki	Tungusic	ERS	Kazakevich and Klyachko (2024)	5124	7.37	1.17	
goem1240	Goemai	Afro-Asiatic	AFR	Hellwig (2024)	1327	5.14	1.40	0.23
goro1270	Gorwaa	Afro-Asiatic	AFR	Harvey (2024)	3652	6.43	1.96	2.11
hoch1243	Hoocak	Siouan	NAM	Hartmann (2024)	2630	8.66	2.35	0.34
jeha1242	Jahai	Austroasiatic	ERS	Burenhult (2024)	913	5.71	1.56	6.57
jeju1234	Jejuan	Koreanic	ERS	Kim (2024)	3624	6.85	2.07	
kaka1265	Kakabe	Mande	AFR	Vydrina (2024)	4338	5.64	1.52	
kama1351	Kamas	Uralic	ERS	Gusev et al. (2024)	4952	7.49	2.27	
komn1238	Komnzo	Yam	PNS	Döhler (2024)	6182	8.06	2.35	1.36
movi1243	Movima	Isolate	SAM	Haude (2024)	2088	8.16	2.37	4.69
ngal1292	Dalabon	Gunwinyguan	AUS	Ponsonnet (2024)	865	10.74	3.20	7.28
nisv1234	Nisvai	Austronesian	PNS	Aznar (2024)	2436	6.00	1.78	5.99
nngg1234	N‖ng	Tuu	AFR	Güldemann et al. (2024)	1819	5.40	1.32	0.16
nort2641	Northern Kurdish	Indo-European	ERS	Haig et al. (2024)	2186	5.42	1.72	0.14
nort2875	Northern Alta	Austronesian	PNS	Garcia-Laguia (2024)	2046	6.95	1.91	4.64
orko1234	Fanbyak	Austronesian	PNS	Franjieh (2024)	1298	5.09	1.35	0.15
pnar1238	Pnar	Austroasiatic	ERS	Ring (2024)	2560	5.90	1.68	
port1286	Daakie	Austronesian	PNS	Krifka (2024)	962	5.10	1.18	1.35
ruul1235	Ruuli	Atlantic-Congo	AFR	Witzlack-Makarevich et al. (2024)	4094	8.27	2.80	0.68
sanz1248	Sanzhi Dargwa	Nakh-Daghestanian	ERS	Forker and Schiborr (2024)	1612	7.20	2.57	
savo1255	Savosavo	Isolate	PNS	Wegener (2024)	1872	6.61	1.87	3.21
sout2856	Nafsan	Austronesian	PNS	Thieberger (2024)	3046	5.69	1.59	0.36
sumi1235	Sümi	Sino-Tibetan	ERS	Teo (2024)	3252	6.38	2.42	
taba1259	Tabasaran	Nakh-Daghestanian	ERS	Bogomolova et al. (2024)	1861	6.29	2.35	
teop1238	Teop	Austronesian	PNS	Mosel (2024)	1093	5.45	1.41	9.06
texi1237	Texistepec Popoluca	Mixe-Zoque	NAM	Wichmann (2024)	1833	7.59	2.57	3.27
trin1278	Mojeño Trinitario	Arawakan	SAM	Rose (2024)	5113	9.23	3.26	1.10
urum1249	Urum	Turkic	ERS	Skopeteas et al. (2024)	5675	7.47	2.17	
vera1241	Vera'a	Austronesian	PNS	Schnell (2024)	1771	4.98	1.45	11.80

Table 7: Languages in the DoReCo-morph dataset. 'nC' = average number of characters; 'nM' = average number of morpheme per word type. % rdp gives the percentage of tokens containing reduplication. The macroareas are: PNS = Papunesia, NAM = North America, SAM = South America, AFR = Africa, ERS = Eurasia, AUS = Australia.

	EMMA-2					BPR				
	chip	morf	para	AG	vorm	chip	morf	para	AG	vorm
apah	90.2	83.7	**83.8**	80.4	82.8	88.2	62.4	53.0	32.2	**68.5**
arap	92.9	89.8	70.2	**90.5**	87.7	65.5	31.3	35.7	15.8	**36.7**
bain	95.8	74.8	80.2	**93.3**	89.3	87.8	40.5	**46.8**	37.4	36.4
beja	89.8	81.9	80.9	89.5	**94.8**	78.9	43.6	**47.5**	30.5	33.8
bora	87.3	76.7	67.5	86.6	**92.7**	82.8	47.2	**51.4**	32.7	33.7
cabe	94.2	83.2	82.7	77.6	**89.2**	92.7	59.7	63.5	35.8	**74.4**
cash	95.0	86.1	81.9	87.5	**90.1**	90.7	48.9	**51.8**	34.2	36.6
dolg	95.9	84.9	81.6	88.4	**89.9**	88.9	50.6	**53.0**	28.7	45.2
even	69.9	**81.8**	81.4	78.6	79.3	94.2	48.2	**75.2**	14.4	68.8
goem	95.5	90.2	84.1	80.2	**93.9**	95.1	77.7	65.5	42.6	**80.1**
goro	82.6	81.0	79.3	81.1	**86.3**	82.7	57.8	57.2	24.4	**60.8**
hoch	90.4	87.5	78.8	**87.5**	80.7	81.9	52.9	58.0	37.4	**58.4**
jeha	93.4	**91.5**	90.0	84.1	89.9	89.5	67.7	69.1	32.9	**69.7**
jeju	93.8	86.1	82.0	**86.6**	86.3	87.3	51.7	**56.2**	39.8	53.1
kaka	82.7	82.4	82.5	79.8	**87.6**	89.9	68.1	**68.3**	26.7	63.1
kama	95.3	85.0	87.6	91.5	**94.7**	87.3	45.6	**47.9**	29.8	45.9
komn	92.5	82.9	78.8	91.6	**92.7**	82.7	47.8	**49.2**	34.4	36.2
movi	89.8	85.4	76.2	**86.1**	81.0	85.7	51.4	**57.7**	31.6	47.8
ngal	94.8	87.3	67.9	**88.4**	67.8	87.2	48.9	**50.2**	32.4	42.5
nisv	94.7	87.4	85.7	87.4	**90.5**	91.2	**64.5**	60.3	47.2	60.5
nngg	91.6	**88.5**	80.4	71.6	87.5	91.7	**75.8**	61.5	24.7	69.8
nort	93.6	83.1	85.5	85.1	**90.4**	89.7	60.1	**64.0**	39.0	56.7
nort	86.8	83.3	79.6	**85.5**	78.7	76.0	58.8	62.1	32.0	**64.8**
orko	88.9	83.8	78.2	77.0	**87.7**	85.3	67.9	59.4	35.2	**76.7**
pnar	95.1	88.3	84.0	85.6	**89.8**	94.5	68.6	59.4	45.4	**75.8**
port	90.3	86.7	83.6	75.1	**89.6**	94.6	77.0	74.1	40.5	**89.9**
ruul	91.9	84.0	73.9	87.6	**87.7**	78.4	43.0	50.6	28.9	**51.1**
sanz	94.4	**85.8**	76.4	85.7	74.6	87.7	51.3	**59.4**	28.2	56.8
savo	90.9	87.9	83.8	87.3	**90.4**	85.4	**60.0**	52.7	48.1	46.5
sout	92.7	86.3	84.7	83.9	**91.2**	89.4	**68.2**	60.0	40.2	61.3
sumi	94.2	85.1	85.1	86.7	**92.4**	92.7	**57.1**	54.1	47.1	50.0
taba	91.8	79.5	81.2	**86.1**	83.1	83.2	50.5	**55.6**	36.5	50.8
teop	89.5	**85.0**	77.2	76.5	84.2	88.2	**75.4**	58.3	44.3	54.4
texi	92.2	77.0	76.1	**86.0**	85.8	83.5	50.6	**53.9**	34.5	53.6
trin	96.7	83.9	73.1	**91.0**	86.0	87.7	44.7	**47.9**	35.0	40.7
urum	95.8	87.1	86.0	87.0	**91.3**	92.4	59.4	**60.9**	39.0	51.4
vera	87.3	83.8	78.8	78.1	**87.4**	84.6	**68.1**	59.0	43.2	53.7
avg.	91.4	84.6	80.3	84.7	**87.2**	86.9	56.8	**57.0**	34.7	55.6
Q1	89.8	83.1	78.2	80.4	**85.8**	83.5	48.9	**51.8**	30.5	45.9
worst	69.9	**74.8**	67.5	71.6	67.8	65.5	31.3	**35.7**	14.4	33.7

Table 8: EMMA-2 and BPR F1 scores for the DoReCo dataset for [chip]munk (supervised), Morfessor2 [morf2], ParaMA2 [para], MorphAGram [AG], and VORM. Best unsupervised results per language are in bold. Best overall results per language are underlined