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

Barend Beekhuizen
University of Toronto, Mississauga, Department of Language Studies
University of Toronto, Department of Linguistics
barend.beekhuizen@utoronto.ca

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 and a novel, typologically diverse, dataset of 37 languages. Model performance is competitive and sometimes superior on canonical segmentation, but more limited on surface segmentation.

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 recent progress (Virpioja et al., 2013; Narasimhan et al., 2015; Eskander et al., 2020; Xu et al., 2020). Supervised techniques can furthermore only be used for several dozen languages, whereas corpus data and word lists are available for many more. Progress on unsupervised learning is thus desirable to improve the cross-linguistic scope of morphological segmentation. The downstream benefit of morphological segmentation for training language models has been debated (Sälevä and Lignos, 2023), but morphological segmentation has further applications in comparative linguistics: for instance, to study patterns in massively parallel corpora (Liu et al., 2023), or to support semi-automated interlinear-glossing methods (McMillan-Major, 2020).

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 constrain the hypothesis space. This allows it to accurately infer a broader array of morphological processes (infixation, reduplication). The system outperforms, for metrics that reflect canonical than surface segmentation, state-of-the-art unsupervised morphological models for many languages.

With those linguistic goals in mind, evaluation on a more diverse set of languages is further desirable. Existing benchmark datasets reflect only a small part of the diversity in morphological typology, with notable absences of extremely common processes like reduplication (Todd et al., 2022). 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://github.com/dnrb/vorm`. 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) forms a baseline for unsupervised morphological segmentation. It leverages word-internal statistical patterns of character sequences, leading to surface segmentations of the input string. A recent, linguistically inspired, model that similarly leads to surface segmentations is Eskander et al. (2020)'s MorphAGram, which trains Adaptor Grammars (Johnson et al., 2006) on surface strings, representing 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

Proceedings of the 29th Conference on Computational Natural Language Learning, pages 602–626

derived forms, that are often not just superstrings of the base form – transforming believe into believing requires dropping the e.[1] 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 contains the same canonical morpheme as 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 model pair, leveraging more global optimization over the search space of transformations, is 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: heuristic search, constrained by using word pairs and paradigms, 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 regions 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 (canonical or superficial) morphological segments without having access to the correct segments to train on, it is paramount to consider the variation in morphological processes across languages. A typologically-oriented overview of morphology is Haspelmath and Sims (2010), who draw 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 (e.g., 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, e.g. English ablaut (gave as the past tense of give) or stem-internal gemination as the morphological causative in Standard Arabic (waqafa 'stop (intransitive)' → waqqafa 'stop (transitive)'). Finally, in **conversion** the form is unaltered but the grammatical category changes, e.g., English hammer can be used as a noun or verb.

Given this diversity, the focus on non-reduplicative affixation alone is narrow. Reduplication is, for instance, extremely common: over 80% of languages have some form of it (Rubino, 2013). A smaller set of languages has stem-internal modifications such as ablaut or tone change (Bickel and Nichols, 2013) – Yu (2007) finds infixation in 111 languages from 26 language families.

Surface segmentation models such as Morfessor and MorphAGram inherently rule out infixation

[1] Character strings are represented throughout with the typewriter font.

w_r	c_t	m
cåm	$danke$	danke
cåm	fuehl	bauchgefuehl ehrgefuehl fuehl fuehle fuehlen fuehlst fuehlt fuehlte fuehlten gefuehl gefuehle gefuehlen gefuehllos (40 more)

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

and base modification, and typically do not identify reduplication as distinct from regular affixation (but see Todd et al., 2022). 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 develops this line of research.

3 The VORM model

The proposed model, VORM ('Vertaling Ondersteunt Redelijke Morfologie' – 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 **[S1]**, which guide the learning of productive morphological transformations **[S2]**. Next, the learned transformations are applied beyond the potential morphological families by propagating the inferred transformations to the full vocabulary **[S3]**. Figure 1 presents a simplified illustration of the model to follow along with the technical introduction.

3.1 S1: Determining morphological families

One common challenge in unsupervised systems that use word pairs (Narasimhan et al., 2015; Xu et al., 2020) is to avoid oversegmentation. Recurrent phonotactic or orthographic patterns may lead to the induction of spurious transformations. 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 are often similar for morphological variants. Here, I propose to use another way to constrain the comparison, namely bitexts and word alignments.

The general procedure is as follows: we consider a bitext B of translations between a target language 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 translation-equivalent utterances $\langle u_r, u_t\rangle$. 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 goal 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 a set of morphological families. Several functions could be defined mapping the bitext B onto a set of morphological families M. Word alignment models are a sensible candidate, except for the fact that morphologically rich target languages have a long tail of morphologically complex words which risk not getting accurately aligned, as indeed found by Beekhuizen (2025).

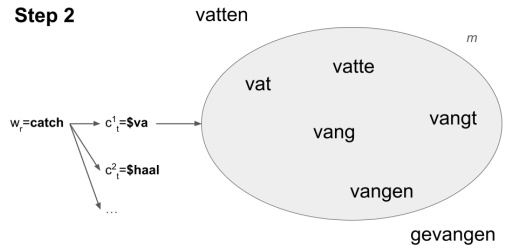
Instead, I use here the forward step of the Conceptualizer model of Liu et al. (2023), which, given a seed word w_r in the reference language r, iteratively finds character substrings $[c_t^1, c_t^2, \ldots c_t^n]$ of words in t whose distribution across the utterances in B is statistically most strongly 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.

Table 1 presents examples of morphological families, using the seed language (Vietnamese) and corpora introduced below. Vietnamese cåm 'feel' has two c_t: $danke$ ($ = word boundary) and fuehl. The morphological family of $danke$ definitionally only contains danke itself, whereas fuehl matches many (related) words in the bitext in which it co-occurs with cåm. Figure 1a presents a morphological family found for an English-to-Dutch mapping, used here as our guiding example.

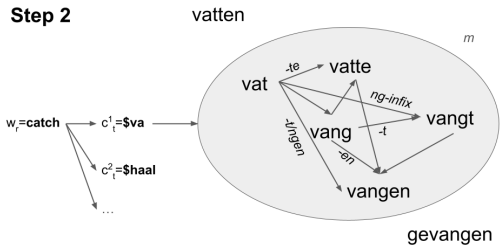
3.2 S2: Learning productive transformations

The morphological families are next used to learn productive transformations in Step 2. This procedure closely follows Morsel (Lignos, 2010). This step starts with initializing a set F of candidate transformations $f_1, f_2, \ldots, f_n$. The procedure iterates over all $m \in M$. For each m, each possible pair $\langle w_t^i, w_t^j\rangle$ in m is considered. All transformations from a set of allowed transformations F_{all} that transform w_t^i into w_t^j are added to F.

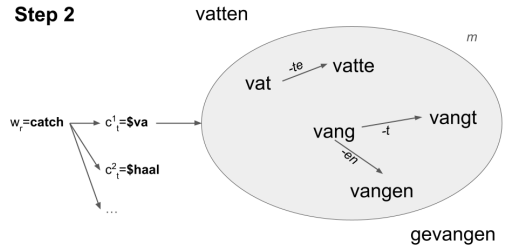
F_{all} is defined to represent the typological diversity of morphological processes. The following are



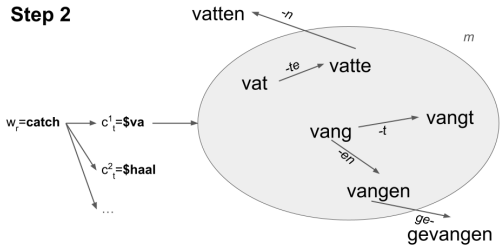
(a) Step 1: Determining morphological families.



(b) Step 2: Retrieving all possible pairwise transformations.



(c) Step 2: Best-first pass determining derivations.



(d) Step 3: Expanding the transformations.

Figure 1: Simplified illustration of the 3 steps of VORM, given English (reference) and Dutch (target).

the allowed types of transformations on the right edge of the string (mirrored transformations are defined for the left edge):

Suffixation: 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: The Fanbyak pair `ini`–`inini` 'to shoot' is modeled by full right reduplication of ∼`ni` (with tildes representing reduplication);

Partial-V right reduplication; all strings of one or more vowels[2] 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 `s`, preceded by `a`.

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 = `a`, given that $w_t^{i'}$=$w_t^{j'}$=`gve`, anchored on a = before-last-consonant-cluster.

Figure 1b illustrates the set of transformations (labels on the directed edges) for the guiding example: the morphological family `$va` reflects two 'real' lemmas: `vang` 'catch [someone]' and `vat` 'catch [a disease]'. Not 'knowing' this, the model tries all possible transformations (as defined below) between any pair of word forms, such as `-te` suffixation between `vat` and `vatte`, but also (incorrectly) `ng`-infixation between `vat` and `vangt`.

Next, a best-first heuristic extracts a set of productive transformations $F_p \subseteq F$. 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 $\langle w_t^i, w_t^j \rangle$ such that there

[2] Vowels are characters that, when stripped of diacritics, are one of {a,e,i,o,u,y}. Any other character is a consonant.

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 $f_{\text{best}} = \arg\max_f |P_f|$.[3] 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 modeled by f_{best}. The procedure is repeated until $|P_{f_{\text{best}}}| < \theta_f$, where θ_f is a pre-set threshold.

Figure 1c illustrates a possible resulting state in our example: common suffixes like `-t` and `-en` are extracted to form derivations between `vang` and `vangt` or `vangen`, while transformations with fewer instances, such as `-t/ngen` modeling the transformation from `vat` to `vangen`, are eliminated at this stage.

The derivations found through the best-first heuristic afford two sources of constraints on the application of F_p in the full vocabulary in the next Step. First, derivations form **chains**: `bookings` may have been derived from `booking` with `-s` suffixation, after which `booking` was derived from `book` through `-ing` suffixation. We denote the chain or derivation d as ⟨`-ing`, `-s`⟩, 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 both models some w_i through d_i and some other w_j through d_j.

An additional 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 ⟨`bauch + _`⟩ and ⟨`_ + gefuehl`⟩, 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] Ties are broken first by morphological type, where the ordering given above is followed, then by affix length (longer affixes are preferred).

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 (such as Dutch `gevangen`, the past participle of `vang` in Figure 1d), 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 $D(s)$ 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 at least half the other words 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 at least half the other words 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. Figure 1d illustrates: the

words vatten and gevangen are not part of the morphological family but can be modeled with productive transforms that form attested chains from words that are in the morphological family.

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). A corpus that can be fill this gap is DoReCo (https://doreco.huma-num.fr/; Seifart et al., 2024), consisting of collections of transcribed fieldwork materials in 52 languages. Much of these materials have interlinear glosses, exemplified in Table 3, where for each word, the morphological analysis is given. Such data allow us to 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 3) 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 other tasks such as interlinear gloss induction, cf. McMillan-Major, 2020). The script also generates a train/development/test split over the data to facilitate experimentation. While the derived data cannot be published under the corpus licence, their generation is exactly reproducible. The datasets used, along with relevant statistics on the derived data, are presented in Table 2 in the Appendices. This table also gives the citation for each individual language, required as part of the user agreement of the corpus.

Morphological profiles of the 37 languages are presented in Appendix A (alongside similar profiles for the MC10 and MC22 data, for comparison). The average number of morphemes ranges between 1.17 and 3.26 per language in the DoReCo-Morph dataset, representing a broad variety of morphological complexity. Moreover, for all languages, there is at least some difference between the canonical forms and the surface strings (cf. Figure 5), suggesting that more than the mere placement of surface string boundaries is necessary to adequately represent the morphological structure of most languages. 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 (22/37) of languages display some form of reduplication, with some languages having over 10% of their word types display reduplication. This underscores the point of Todd et al. (2022) that being able to represent reduplication is necessary for a truly multilingual unsupervised morphological model.

5 Evaluation

Evaluation data First, Vorm is compared with other models on two benchmarks: Morphochallenge 2010 (MC10; Kurimo et al., 2010), with gold data for English, Finnish, Turkish, and German canonical and surface (for all but German) segmentation, and the SIGMORPHON 2022 task on surface segmentation (SGM22; Batsuren et al., 2022) for eight languages. Next, we consider the novel DoReCo-morph dataset of 37 languages.

Evaluation metrics The standard metrics were applied. First, EMMA-2 (Virpioja et al., 2011) measures the precision and recall between the gold (canonical) segmentation and the predicted segmentation by inferring mappings between gold and predicted morpheme labels that optimize Precision and Recall, thus solving the problem of potentially differing morpheme labels. It is applied to both datasets with canonical forms: MC10 and DoReCo-morph. Second, Boundary Precision and Recall (BPR; Batsuren et al., 2022) measures the F1 score of the proportion of predicted boundaries found in the gold data (precision) and conversely the proportion of gold data boundaries predicted (recall) for all datasets. While commonly

language (glottocode; family; area; reference)	language (glottocode; family; area; reference)
Yali (apah1238; Nuclear Trans New Guinea; PNS; Riesberg, 2024)	Nisvai (nisv1234; Austronesian; PNS; Aznar, 2024)
Arapaho (arap1274; Algic; NAM; Cowell, 2024)	N——ng (nngg1234; Tuu; AFR; Güldemann et al., 2024)
Baïnounk Gubëeher (bain1259; Atlantic-Congo; AFR; Cobbinah, 2024)	Northern Kurdish (nort2641; Indo-European; ERS; Haig et al., 2024)
Beja (beja1238; Afro-Asiatic; AFR; Vanhove, 2024)	Northern Alta (nort2875; Austronesian; PNS; Garcia-Laguia, 2024)
Bora (bora1263; Boran; SAM; Seifart, 2024)	Fanbyak (orko1234; Austronesian; PNS; Franjieh, 2024)
Cabécar (cabe1245; Chibchan; NAM; Quesada et al., 2024)	Pnar (pnar1238; Austroasiatic; ERS; Ring, 2024)
Cashinahua (cash1254; Pano-Tacanan; SAM; Reiter, 2024)	Daakie (port1286; Austronesian; PNS; Krifka, 2024)
Dolgan (dolg1241; Turkic; ERS; Däbritz et al., 2024)	Ruuli (ruul1235; Atlantic-Congo; AFR; Witzlack-Makarevich et al., 2024)
Evenki (even1259; Tungusic; ERS; Kazakevich and Klyachko, 2024)	Sanzhi Dargwa (sanz1248; Nakh-Daghestanian; ERS; Forker and Schiborr, 2024)
Goemai (goem1240; Afro-Asiatic; AFR; Hellwig, 2024)	Savosavo (savo1255; Isolate; PNS; Wegener, 2024)
Gorwaa (goro1270; Afro-Asiatic; AFR; Harvey, 2024)	Nafsan (sout2856; Austronesian; PNS; Thieberger, 2024)
Hoocak (hoch1243; Siouan; NAM; Hartmann, 2024)	Sümi (sumi1235; Sino-Tibetan; ERS; Teo, 2024)
Jahai (jeha1242; Austroasiatic; ERS; Burenhult, 2024)	Tabasaran (taba1259; Nakh-Daghestanian; ERS; Bogomolova et al., 2024)
Jejuan (jeju1234; Koreanic; ERS; Kim, 2024)	Teop (teop1238; Austronesian; PNS; Mosel, 2024)
Kakabe (kaka1265; Mande; AFR; Vydrina, 2024)	Texistepec Popoluca (texi1237; Mixe-Zoque; NAM; Wichmann, 2024)
Kamas (kama1351; Uralic; ERS; Gusev et al., 2024)	Mojeño Trinitario (trin1278; Arawakan; SAM; Rose, 2024)
Komnzo (komn1238; Yam; PNS; Döhler, 2024)	Urum (urum1249; Turkic; ERS; Skopeteas et al., 2024)
Movima (movi1243; Isolate; SAM; Haude, 2024)	Vera'a (vera1241; Austronesian; PNS; Schnell, 2024)
Dalabon (ngal1292; Gunwinyguan; AUS; Ponsonnet, 2024)	

Table 2: Languages in the DoReCo-morph dataset. The macroareas are: PNS = Papunesia, NAM = North America, SAM = South America, AFR = Africa, ERS = Eurasia, AUS = Australia.

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 3: Interlinear Gloss; Savosavo (Wegener, 2024)

used, it is a less linguistically insightful metric, as (per Figure 5) non-identity between the canonical morphemes and the surface string is crosslinguistically extremely common.

Training data The bitexts used for MC10 and SGM22 were (up to) a million words of bitext from Opus2018 (Lison and Tiedemann, 2016) subtitles from `www.opensubtitles.org/`. Vietnamese was chosen as the reference language as it has little morphology. Bitexts for German and Turkish were orthographically normalized with the test data. For the DoReCo-morph experiment, bitexts were generated from the corpora, using the **w** and **f** layers (cf. Table 3). Free translations were mostly in English, with some in Malay, Spanish, and others.

Comparison models For the MC10 and SGM22, I compare Vorm against published results, but add Morfessor2 (Virpioja et al., 2013) to the latter as an unsupervised baseline. For DoReCo-morph, I run Morfessor2, ParaMA2 (Xu et al., 2020), and MorphAGram (Eskander et al., 2020) (in the language-independent setting) as unsupervised models, and Chipmunk (Cotterell et al., 2015), as a supervised model. 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).

Tuning Models were tuned on each dataset individually, using the standard splits of MC10 and SGM22, and the proposed split (12% of the data of each language) for DoReCo-morph. To better understand the performance of the Vorm model, an ablation experiment was furthermore run, leaving out Step 1 ('-S1') by instead having one single morphological family containing all vocabulary items, not extracting compounds ('-C'), and leaving out Step 3 ('-S3'). As the optimal hyperparameters for these settings may differ from the unablated version of Vorm, tuning was done on each individual ablation variant. Details and results (hyperparameter settings and accuracy metrics) of the tuning for all models and ablation variants can be found in Appendix B. Below, I will report test data results on the best-tuned model per model/ablation variant.

6 Results

6.1 Results by dataset

MorphoChallenge 2010 results. Table 4 presents the results for MC10. First, we focus on the metric for canonical segmentation, EMMA-2. Across the four languages, Vorm has the highest average F_1 score at 90.0. For the individual languages, we find that Vorm presents a substantial improvement over MorphAGram and Morfessor for Finnish, German, and Turkish, but not for English, where MorphAGram outperforms Vorm. Considering the ablation steps, we find that for some languages not using the translation equivalences in Step 1 ('-S1') or not finding compounds ('-C') improves

	EMMA-2						BPR					
	morf	AG	VORM	-C	-S1	-S3	morf	AG	VORM	-C	-S1	-S3
English	85.9	88.7	84.1	84.1	**91.2**	56.6	75.2	**80.0**	54.0	52.9	43.1	40.6
Finnish	73.4	77.7	94.9	**95.0**	92.9	46.4	62.8	**71.1**	24.8	23.6	25.1	40.6
German	80.9	85.9	93.7	**93.9**	93.2	41.3			n/a			
Turkish	61.3	69.3	**87.3**	86.0	78.9	28.3	64.6	**78.9**	24.2	23.3	23.2	19.0
avg.	75.3	80.4	**90.0**	89.7	89.0	43.1	67.5	**76.7**	34.3	33.3	30.5	24.1

Table 4: Model comparison on the development sets for Morphochallenge 2010 [MC10], comparing Morfessor (Morf) and the best MorphAGram (AG) model against VORM with ablation variants, on EMMA-2 and BPR F_1 scores. The best result per language and per metric is boldfaced.

	DeepSPIN-3	morf	VORM	-C	S1	S3
Czech	93.84	**28.71**	28.18	27.12	25.75	6.34
English	93.63	**49.90**	41.85	33.17	40.63	10.80
French	95.73	**23.63**	20.33	21.67	20.31	3.99
Hungarian	98.72	**34.47**	34.43	33.67	32.44	32.44
Italian	97.43	11.84	11.35	**12.50**	11.40	2.39
Latin	99.39	**17.77**	12.92	13.25	12.98	4.10
Russian	99.35	11.46	15.60	**18.56**	14.42	0.65
Spanish	99.04	9.23	17.99	**19.06**	17.96	1.25
avg.	97.29	**23.38**	20.82	20.67	20.07	7.41

Table 5: Model comparison on the tests sets for the SIGMORPHON 2022 challenge comparing DeepSPIN-3 (supervised) and Morfessor2 against VORM and its ablation variants on the Batsuren et al. (2022) evaluation measure. The best unsupervised result per language is boldfaced; the best result overall underlined.

the quality of the model, suggesting that further development of these components might be necessary. Removing the extension to the full vocabulary ('-S3') is, however, consistently detrimental.

On the surface segmentation measure of BPR, VORM is substantially outperformed by Morfessor and MorphAGram. This effect may be due to the differences between the metrics: EMMA-2 favours canonical morpheme identity, but 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 for VORM.

SIGMORPHON 2022 results. For the SGM22, only surface segmentation is considered, using the metric provided by the task. The results are presented in Table 5. While no unsupervised model performs even close to the supervised models (here, the best-performing supervised model DeepSPIN-3, Peters and Martins, 2022, is given as a reference point), VORM without compounding ('-C') occasionally outranks Morfessor2 in its performance. This further underscores the previous observation that VORM does not excel in surface segmentation.

DORECO-MORPH. Finally, Table 6 present the aggregated results for VORM and comparison models over the 37 DORECO-MORPH languages, with Table 15 in the appendices presenting the EMMA-2 scores per language. For the EMMA-2 scores, unablated VORM outperforms the other unsupervised models for 20/37 languages (32 if considering the ablated variants). MorphAGram is the optimal model for 1 language. Considering average model performance, we find VORM outperforming other unsupervised models by a substantial margin, coming within a 2% range of the supervised Chipmunk model. Notably, the language VORM performs worst on still reaches an EMMA-2 score of 78.0, while Chipmunk only scores 69.9 on its worst case – with Morfessor also performing robustly at 77.1. In the ablation experiment, we find that the effects of leaving out compounding ('-C') are negligible, and that not having Step 1 in many cases *improves* performance (indeed, the worst case *without* Step 1 is slightly better than the worst case of unablated VORM). Omitting Step 3 ('-S3') in all cases leads to a substantial drop in performance.

For surface segmentation (BPR), however, the performance is more mixed: here, Chipmunk is the best overall model with a large margin, with ParaMA2 being the best model for 15 languages, and VORM for 21. Notably, leaving out Step 3 here frequently leads to an improvement for VORM, owing perhaps to the fact that these are small datasets for which the complete vocabulary is captured in Step 1 (as opposed to the MC10 data, where the corpus data only contained a subset of the test data), and that as such extension beyond the morphological families leads to more Precision errors than improvement of Recall.

6.2 Discussion

On the whole, the results suggest that VORM is a competitive model of **canonical segmentation**. On

	EMMA-2								BPR							
	chip	morf	para	AG	vorm	-S1	-C	-S3	chip	morf	para	AG	vorm	-S1	-C	-S3
max?		4		1	**20**	11		1		1	**15**		5	1		**15**
avg.	91.4	86.1	80.3	84.6	**89.9**	88.6	89.9	69.1	86.9	56.9	57.0	34.3	58.8	46.6	58.2	**60.4**
worst	69.9	77.1	69.7	71.3	78.0	**78.3**	78.0	35.3	65.6	31.0	**35.7**	13.3	32.9	29.4	32.7	29.6

Table 6: Aggregated EMMA-2 & BPR F_1 scores for the DoReCo-morph dataset for [chip]munk (supervised), [Morf]essor2, [Para]MA2 , Morph[AG]ram, and VORM with its ablation variants. Best unsupervised results in bold; best overall results underlined.

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

morphologically complex languages like Finnish and Turkish, its improvement over other unsupervised models is substantial. The ablation experiments paint a complicated picture of what leads to these results – the addition of a compounding component, and the 'narrowing' of the hypothesis space through the use of morphological families in Step 1 have only a small, and sometimes even a negative, effect. The Precision-oriented focus of the compounding component may lead to limited extraction of compounds. Step 1 may be redundant with the filtering mechanisms of Step 2: when all words are compared with each other, low-frequency transforms will be eliminated by the frequency threshold, and frequent, but spurious, transforms may be weeded out by being pre-empted by a more frequent transform in the best-first pass. However, for some languages (Turkish in MC10, several DoReCo-morph languages), the omission of Step 1 does come at a cost, suggesting that narrowing by translation equivalence is not always redundant.

On **surface segmentation** VORM does not perform as competitively. This can be attributed to the lower Recall the model achieves here, and its focus on canonical segmentation leading to variable boundaries on the surface string. Importantly, this contrast suggests that canonical and surface segmentation are substantially different tasks.

The examples in Table 7 demonstrate the model's capacity to analyze reduplication. We see that only VORM analyses the forms correctly, 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, Morfessor).

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

7 Conclusion

This paper introduces VORM, a novel unsupervised morphological segmentation system, which uses translation-equivalency to narrow down the set of word pairs on which the inferred morphological transformations are based. Aside from affixation, the model can represent base-modifying transformations and reduplication. Generalizing models are induces 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) in order to maintain high precision.

Further exploration on the DoReCo-morph dataset could identify more specific modeling challenges by breaking down the full dataset into linguistically interesting subsets (cases with reduplication, cases where the canonical form deviates substantially from the surface form through assimilation processes, etc.). 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.

Barend Beekhuizen. 2025. Token-level semantic typology without a massively parallel corpus. In *The 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*.

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.

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.

Gonzalo Navarro. 2001. A guided tour to approximate string matching. *ACM Comput. Surv.*, 33(1):31–88.

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 Morphological profiles of the languages

A.1 Number of morphemes

The number of morphemes, as given in the gold standard datasets is presented in Figures 2 (for DoReCo-Morph), 3 (for MC10) and 4 (for MC22).

A.2 Number of insertions and deletions

Only for the DoReCo-Morph data do we have both the surface forms and the canonical forms; for MC10, despite canonical morphemes being given, the inflectional morphemes are mostly given in a featural notation ('+PL', '+SUP') and as such a canonical string of phonological/orthographic segments cannot be faithfully extracted. For MC22, only suface string segmentations are given.

The difference between the surface form and the canonical form was calculated by running a Wagner-Fisher algorithm (see Navarro, 2001, for the complexity of authorship attribution of this algorithm) on the two strings to compute the path of maximal string overlap, allowing only for character insertion and deletion operations. Any insertions (the canonical form has more characters than the surface form; e.g., `believing` for `believe` + `-ing`) and deletions (the surface form has more characters than the canonical form; e.g., `barring` for `bar` + `-ing`) were counted.

Figure 5 presents the number of insertions and deletion going from the surface form to the canonical form.

A.3 Prevalence of morpheme types

Per language, it was determined heuristically whether a morpheme was free, reduplicative and bound, or affixal and bound. Reduplicative morphemes are consistently tagged with a tilde ('∼') on their left or right edge in the DoReCo-Morph data (and are absent from the MC10 data). Affixal morphemes are marked with capitalized glosses in DoReCo-Morph and either a grammatical feature-style notation starting with '+' (e.g., '+PL') or a grammatical category marking as 'p' (prefix) or 's' (suffix) in the MC10 data. Morpheme types are undefined for the MC22 data.

Figure 6 presents the counts for the DoReCo-Morph data, while Figure 7 presents the counts for the MC10 data.

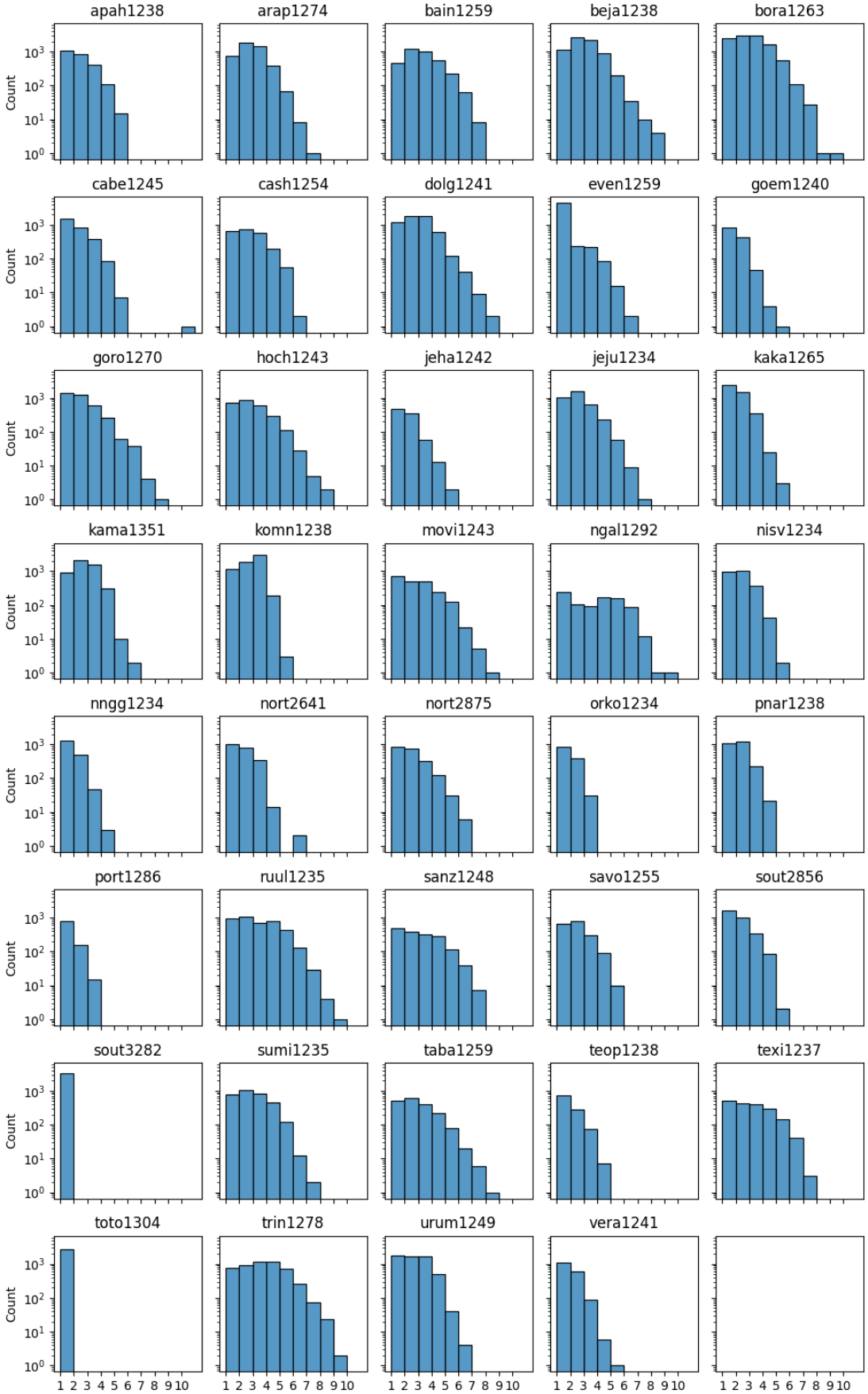


Figure 2: Histogram of the number of morphemes in the DoReCo-Morph data.

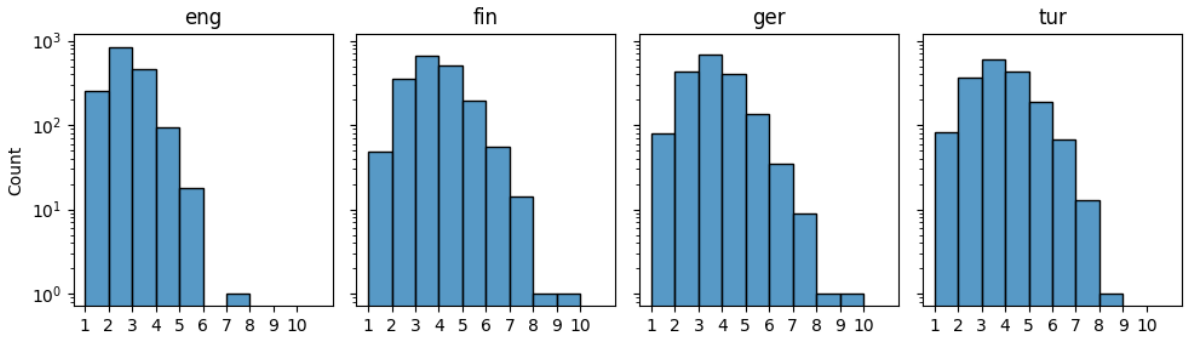


Figure 3: Histogram of the number of morphemes in the MC10 data.

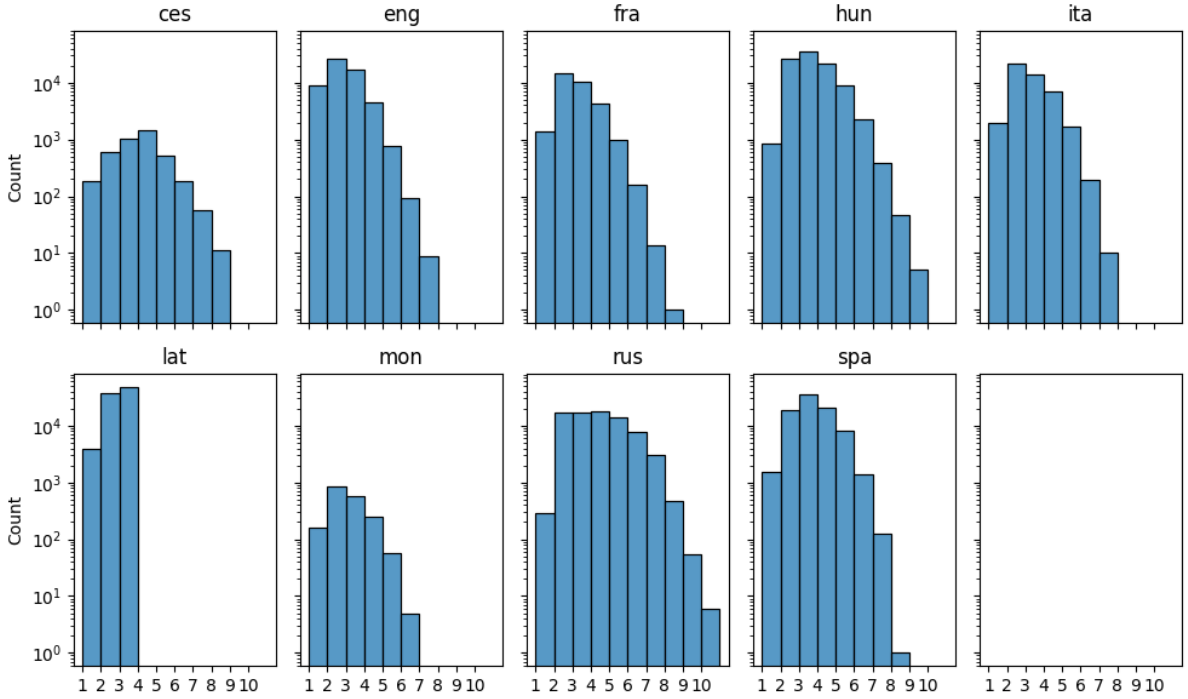


Figure 4: Histogram of the number of morphemes in the MC22 data.

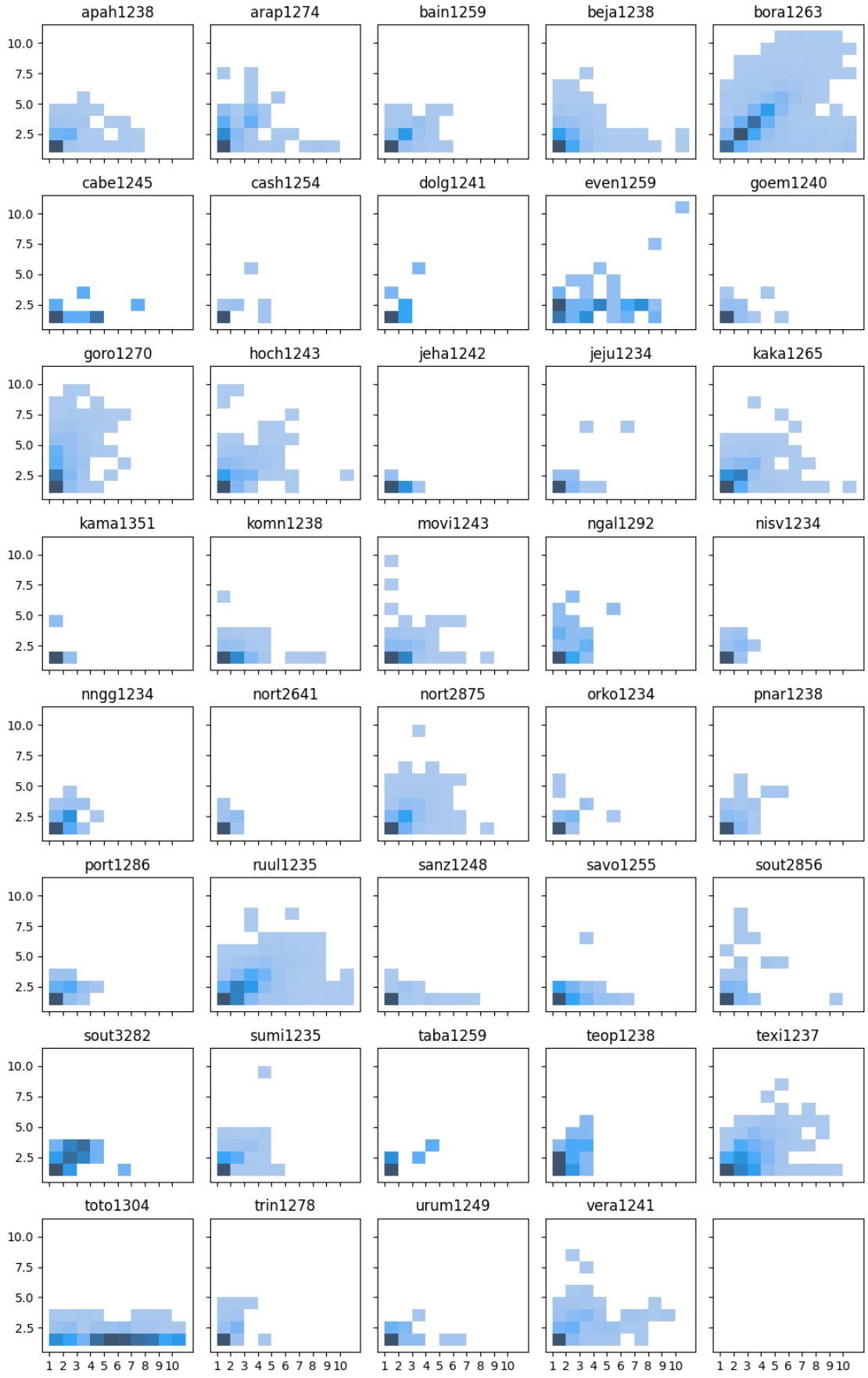


Figure 5: Heatmap of the number of insertions (rows) and deletions (columns) going from the surface form to the canonical form in the DoReCo-Morph data (darker means more instances).

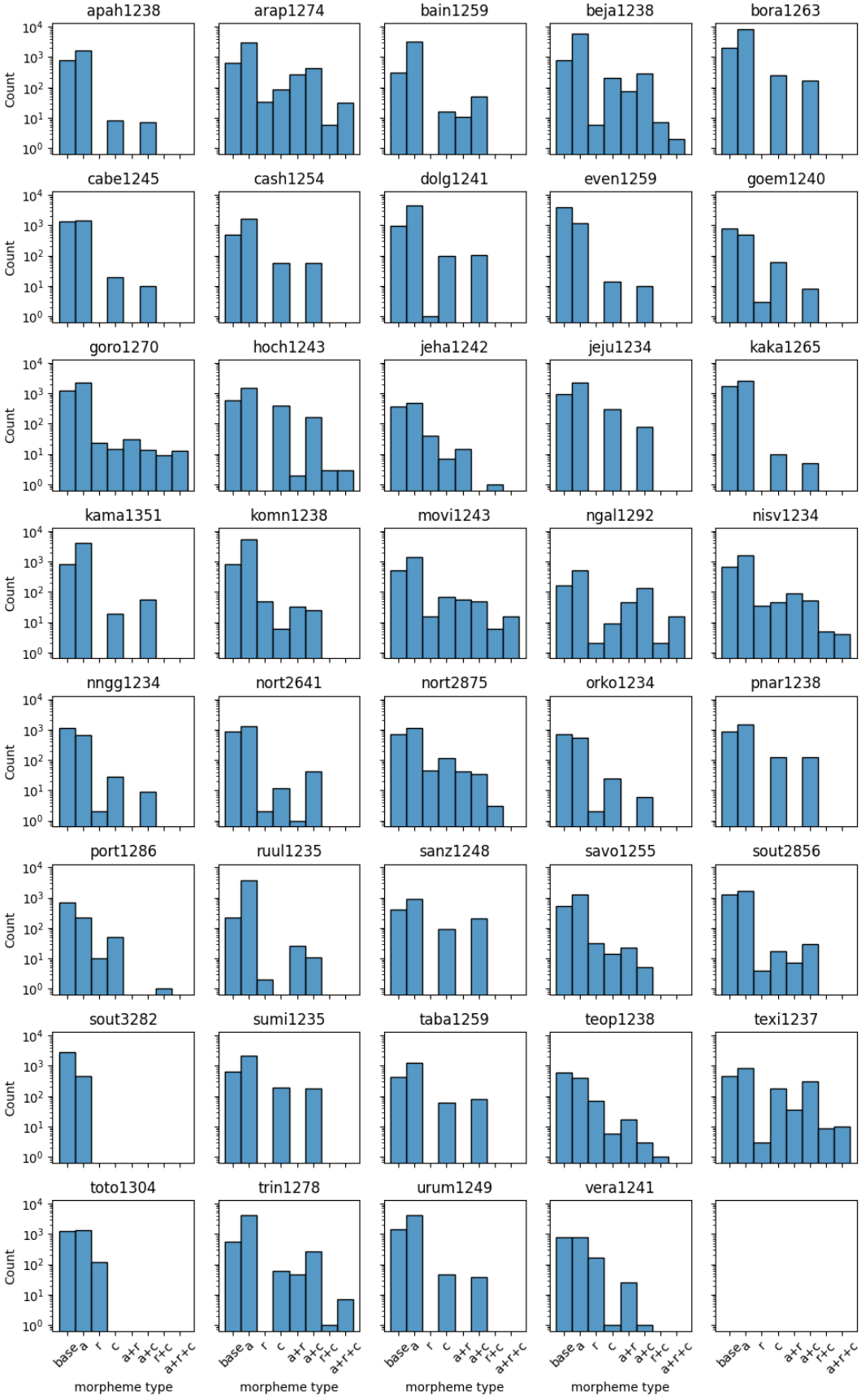


Figure 6: Histogram of the combinations of three morphological types in the lexical items of the DoReCo-Morph languages. 'base' = morphologically symplex, 'a' = affixation; 'r' = reduplication; 'c' = compounding

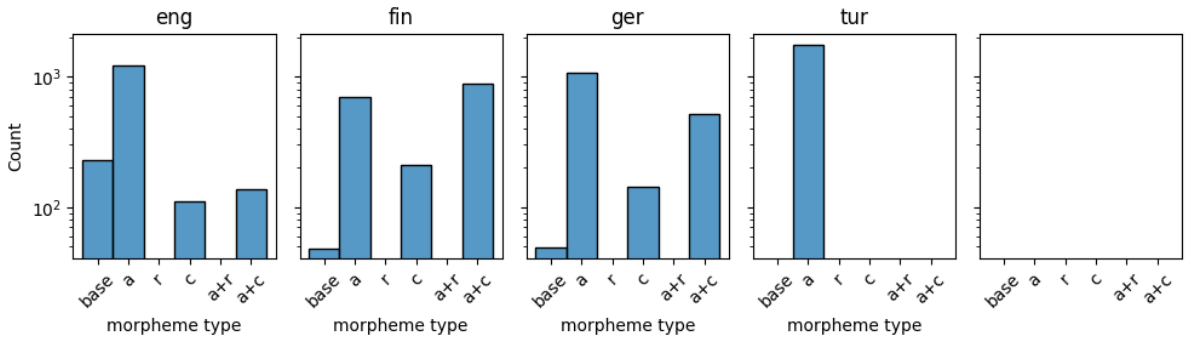


Figure 7: Histogram of the combinations of three morphological types in the lexical items of the DoReCo-Morph languages. 'base' = morphologically symplex, 'a' = affixation; 'r' = reduplication; 'c' = compounding

B Tuning experiments

The model was tuned on the development split (12% of the data for each language) in the DoReCo-morph data, the training split for MC10 and the development split for SGM22. For each task, and for each metric (EMMA-2 or BPR), the best-performing set of hyperparameters of each model (and of each ablation variant of the VORM model) was selected.

B.1 DoReCo-Morph data

For **VORM**, the free parameters $\theta_f \in \{3, 5, 10, 20\}$ (minimum number of word pairs modeled by a transformation in Step 2) and $\theta_c \in \{1, 5, 10, 20\}$ (minimum number of compound template occurrences for it to be used in Step 3) using a grid search over the values. The ablation over model components (-CMPD: no compounding, -S1: no Step 1, i.e. comparing all of the vocabulary in Step 2, -S3: no Step 3) was done simultaneously, as optimal values for θ_f and θ_c can be expected to vary across ablation steps.

Tuning scores are given in Table 8. As the optimal parameter settings do not line up across the two measures, the parameter setting with the highest average across the two scores was selected per ablation setting. In the main text, we report on $\theta_f = 20, \theta_c = 1$ for no ablation and -CMPD, $\theta_f = 20, \theta_c = 20$, for -S1, and $\theta_f = 3, \theta_c = 1$ for -S3.

For **Morfessor**, the model was tuned on the three ways of using token counts (token counts or: 'token', no counts or: 'type', and 'log-counts'). Scores are given in Table 9. Log-counts is on average the best-performing setting.

For **ParaMA** (Xu et al., 2020), I varied the minimum stem length ($\in \{1, 3\}$) and whether the model

ablation	θ_c	θ_f	EMMA-2	BPR
	1	3	90.48	42.48
	1	5	90.60	46.57
	1	10	89.62	51.51
	1	20	85.73	**56.62**
	5	3	90.63	42.43
	5	5	90.80	46.47
	5	10	89.78	51.37
	5	20	85.82	56.42
	10	3	90.65	42.43
	10	5	**90.83**	46.46
	10	10	89.80	51.25
	10	20	85.86	56.29
	20	3	90.65	42.43
	20	5	**90.83**	46.46
	20	10	89.80	51.22
	20	20	85.88	56.15
-CMPD	n/a	3	90.65	42.43
-CMPD	n/a	5	**90.83**	46.46
-CMPD	n/a	10	89.80	51.22
-CMPD	n/a	20	85.88	**56.09**
-S1	1	3	87.25	30.02
-S1	1	5	87.88	30.21
-S1	1	10	88.52	33.21
-S1	1	20	88.71	**40.80**
-S1	5	3	87.42	30.03
-S1	5	5	88.53	30.24
-S1	5	10	89.39	33.23
-S1	5	20	89.21	40.70
-S1	10	3	87.42	30.03
-S1	10	5	88.53	30.24
-S1	10	10	89.69	33.25
-S1	10	20	89.52	40.63
-S1	20	3	87.42	30.03
-S1	20	5	88.53	30.24
-S1	20	10	89.69	33.25
-S1	20	20	**89.77**	40.52
-S3	1	3	**76.71**	**59.74**
-S3	1	5	76.48	59.53
-S3	1	10	76.17	59.06
-S3	1	20	75.55	58.42
-S3	5	3	76.70	59.70
-S3	5	5	76.48	59.45
-S3	5	10	76.17	59.00
-S3	5	20	75.55	58.39
-S3	10	3	76.70	59.69
-S3	10	5	76.47	59.45
-S3	10	10	76.16	58.93
-S3	10	20	75.56	58.33
-S3	20	3	76.70	59.69
-S3	20	5	76.47	59.44
-S3	20	10	76.17	58.92
-S3	20	20	75.55	58.27

Table 8: Average EMMA-2 and BPR scores on DoReCo-Morph tuning data for the VORM model. Best model scores per ablation (none, -CMPD, -S1, -S3) boldfaced, best overall score (per metric) underlined.

parameters	EMMA-2	BPR
token	86.6	**51.5**
type	85.6	33.2
log-counts	**88.2**	48.4

Table 9: Average EMMA-2 and BPR scores on DoReCo-Morph tuning data for the Morfessor model. Best model scores boldfaced

parameters	EMMA-2	BPR
−compound, min stem ≥ 1	78.7	41.5
+compound, min stem ≥ 1	78.7	41.5
−compound, min stem ≥ 3	83.3	53.5
+compound, min stem ≥ 3	**83.3**	**53.5**

Table 10: Average EMMA-2 and BPR scores on DoReCo-Morph tuning data for the ParaMA model. Best model scores boldfaced

parameters	EMMA-2	BPR
default	**86.7**	**23.1**
reported	86.1	22.2
vaguer-G	85.8	22.4
sharper-G	86.2	22.1

Table 11: Average EMMA-2 and BPR scores on DoReCo-Morph tuning data for the MorphAGram model. Best model scores boldfaced

tried to find compounds. Table 10 presents the results. Compounding and a minimal stem length of ≥ 3 leads to the best setting on both metrics.

For **MorphAGram** (Eskander et al., 2020), the primary two settings were the model defaults and the reported values (in which the Gamma parameters of the py-cfg model (Johnson et al., 2006) were set to pyb-gamma-s $= 10$ and pyb-gamma-c $= 0.1$. As these parameters were found to be effective before, tuning was undertaken in both directions, resetting them to the default ('vaguer-G': pyb-gamma-s$= 1$ and pyb-gamma-c$= 1$) and making them more extreme (as suggested in the py-cfg documentation: 'sharper-G': pyb-gamma-s$= 100$ and pyb-gamma-c$= 0.01$). No tuning of the Alpha and Beta parameters of the py-cfg model was done as the optimal tuned in the paper was the default. Table 11 presents the results; the default setting was consistently the optimal one.

For **Chipmunk**, no parameters were found that would lead to differences in model performance.

B.2 MC10

For the MC10 data, only novel results were generated for Vorm, with the other results being cited from other papers. The free parameters $\theta_f \in \{30, 60, 100\}$ (minimum number of word pairs modeled by a transformation in Step 2) and $\theta_c \in \{10, 30, 60\}$ (minimum number of compound template occurrences for it to be used in Step 3) were tuned using a grid search over the values. As with the DoReCo-morph data, the ablation variants were tuned separately. The results are given in Table 12. The best average parameter settings used for the test phase were $\theta_f = 100, \theta_c = 10$ for no-ablation, both metrics, $\theta_f = 100$ for -C, $\theta_f = 100, \theta_c = 100$ for -S1, EMMA-2, and $\theta_f = 100, \theta_c = 30$ for -S1, BPR, and finally $\theta_f = 30, \theta_c = 100$ for -S3, EMMA and $\theta_f = 60, \theta_c = 100$ for -S3, BPR.

B.3 MC22

For the MC22 data, the free parameters of VORM $\theta_f \in \{30, 60, 100\}$ (minimum number of word pairs modeled by a transformation in Step 2) and $\theta_c \in \{10, 30, 100\}$ (minimum number of compound template occurrences for it to be used in Step 3) were tuned using a grid search over the values. The results are given in Table 13. For EMMA-2, values of $\theta_f = 100, \theta_c = 100$ were found to be on average optimal for the no-ablation variant, -S1 and -C, and $\theta_f = 30, \theta_c = 100$ for the

ablation	θ_c	θ_f	EMMA-2 eng	EMMA-2 fin	EMMA-2 ger	EMMA-2 tur	EMMA-2 avg.	BPR eng	BPR fin	BPR tur	BPR avg.
	10	30	**92.93**	96.17	**95.77**	90.48	93.84	42.34	26.09	23.40	30.61
	10	60	92.59	96.26	95.37	92.84	94.26	53.11	**26.95**	22.45	34.17
	10	100	92.22	96.27	94.46	94.16	**94.28**	**55.87**	25.60	**24.05**	**35.17**
	30	30	**92.93**	96.17	**95.77**	90.48	93.84	42.34	26.09	23.40	30.61
	30	60	92.59	96.26	95.37	92.84	94.26	53.11	**26.95**	22.45	34.17
	30	100	92.22	96.27	94.46	94.16	**94.28**	**55.87**	25.60	**24.05**	**35.17**
	100	30	**92.93**	96.17	**95.77**	90.48	93.84	42.34	26.09	23.40	30.61
	100	60	92.59	96.26	95.37	92.84	94.26	53.11	**26.95**	22.45	34.17
	100	100	92.18	**96.28**	94.47	**94.21**	**94.28**	**55.87**	25.60	**24.05**	**35.17**
-C	n/a	30	**92.97**	96.26	**95.86**	90.46	93.89	42.05	25.49	23.21	30.25
-C	n/a	60	92.62	96.26	95.41	92.81	94.28	52.32	**26.14**	21.98	33.48
-C	n/a	100	92.20	**96.36**	94.49	**94.11**	**94.29**	**54.90**	24.66	**23.30**	**34.29**
-S1	10	30	93.87	90.09	91.93	82.60	89.62	34.69	25.55	20.27	26.84
-S1	10	60	**94.29**	94.14	93.55	87.01	92.25	38.83	24.50	21.05	28.13
-S1	10	100	92.90	95.37	95.13	**90.31**	93.43	**43.78**	**25.65**	23.23	**30.89**
-S1	30	30	93.87	90.09	91.93	82.60	89.62	34.69	25.55	20.27	26.84
-S1	30	60	**94.29**	94.14	93.55	87.01	92.25	38.83	24.50	21.05	28.13
-S1	30	100	92.90	95.37	95.13	**90.31**	93.43	**43.78**	**25.65**	23.23	**30.89**
-S1	100	30	93.87	90.09	91.93	82.60	89.62	34.69	25.55	20.27	26.84
-S1	100	60	92.88	94.14	93.55	87.01	91.90	39.23	24.50	21.05	28.26
-S1	100	100	92.89	**95.65**	**95.24**	90.16	**93.48**	**43.78**	25.62	**23.34**	24.48
-S3	10	30	**73.40**	**58.59**	**53.77**	43.00	**57.19**	**39.94**	**14.73**	21.05	25.24
-S3	10	60	73.36	58.58	**53.77**	42.97	57.17	39.69	14.72	**21.39**	25.27
-S3	10	100	73.07	58.58	53.76	43.02	57.11	39.24	14.37	21.12	24.91
-S3	30	30	**73.40**	**58.59**	**53.77**	43.00	**57.19**	**39.94**	**14.73**	21.05	25.24
-S3	30	60	73.36	58.58	**53.77**	42.97	57.17	39.69	14.72	**21.39**	25.27
-S3	30	100	73.07	58.58	53.76	43.02	57.11	39.24	14.37	21.12	24.91
-S3	100	30	**73.40**	**58.59**	**53.77**	43.00	**57.19**	**39.94**	**14.73**	21.05	25.24
-S3	100	60	73.36	58.58	**53.77**	42.97	57.17	39.69	14.72	**21.39**	**25.27**
-S3	100	100	73.11	58.50	53.76	**43.15**	57.13	39.24	14.37	21.06	24.89

Table 12: EMMA-2 and BPR scores on MC10 tuning data for the VORM model. Best model scores per ablation variant and per metric boldfaced

-S3 variant. For BPR, $\theta_f = 100, \theta_c = 100$ was found to be the optimal setting for no-ablation and -C, and $\theta_f = 100, \theta_c = 30$ for -S1 and -S3.

For **Morfessor**, the model was again tuned on the three ways of using token counts (token counts or: 'token', no counts or: 'type', and 'log-counts'). Scores are given in Tables 14. Across languages, the 'type' setting performed the best.

ablation	θ_c	θ_f	ces	eng	fra	hun	ita	lat	rus	spa	avg.
	10	30	27.68	33.61	**21.65**	**35.21**	12.44	**13.22**	**18.95**	**19.14**	22.74
	10	60	**30.93**	35.49	20.78	34.10	**12.47**	12.67	16.30	18.08	22.60
	10	100	28.18	**41.85**	20.33	34.43	11.35	12.92	15.60	17.99	**22.83**
	30	30	27.68	33.61	**21.65**	**35.21**	12.44	**13.22**	**18.95**	**19.14**	22.74
	30	60	**30.93**	35.49	20.78	34.10	**12.47**	12.67	16.30	18.08	22.60
	30	100	28.18	**41.85**	20.33	34.43	11.35	12.92	15.60	17.99	**22.83**
	100	30	27.68	33.61	**21.65**	**35.21**	12.44	**13.22**	**18.95**	**19.14**	22.74
	100	60	**30.93**	35.49	20.78	34.10	**12.47**	12.67	16.30	18.08	22.60
	100	100	28.18	**41.85**	20.33	32.44	11.35	12.92	15.60	17.99	22.58
-CMPD	n/a	30	27.12	33.17	**21.67**	**33.67**	**12.50**	**13.25**	**18.56**	**19.06**	**22.38**
-CMPD	n/a	60	**28.86**	34.24	20.73	32.34	12.48	12.72	15.20	17.95	21.82
-CMPD	n/a	100	25.32	**40.62**	20.34	31.77	10.50	12.89	14.42	17.96	21.73
-S1	10	30	21.55	30.41	17.12	34.06	**16.70**	**26.67**	13.19	13.19	21.61
-S1	10	60	22.92	29.22	16.74	34.67	16.19	23.42	14.82	17.37	21.92
-S1	10	100	24.77	28.24	17.07	**34.97**	16.53	19.51	**17.34**	16.05	21.81
-S1	30	30	21.55	30.41	17.12	34.06	**16.70**	**26.67**	13.19	13.19	21.61
-S1	30	60	22.92	29.22	16.74	34.67	16.19	23.42	14.82	17.37	21.92
-S1	30	100	24.77	28.24	17.07	**34.97**	16.53	19.51	**17.34**	16.05	21.81
-S1	100	30	21.55	30.41	17.12	34.06	**16.70**	**26.67**	13.19	13.19	21.61
-S1	100	60	22.92	29.22	16.74	34.67	16.19	23.42	14.82	17.37	21.92
-S1	100	100	**25.75**	**40.63**	**20.31**	32.44	11.40	12.98	14.42	**17.96**	**21.99**
-S3	10	30	**6.50**	**10.91**	**4.05**	**2.70**	**2.42**	**4.10**	**0.68**	**1.27**	4.08
-S3	10	60	6.45	10.85	4.02	2.69	2.40	4.09	0.67	1.26	4.05
-S3	10	100	6.34	10.80	3.99	2.67	2.39	**4.10**	0.65	1.25	4.02
-S3	30	30	**6.50**	**10.91**	**4.05**	**2.70**	**2.42**	**4.10**	**0.68**	**1.27**	4.08
-S3	30	60	6.45	10.85	4.02	2.69	2.40	4.09	0.67	1.26	4.05
-S3	30	100	6.34	10.80	3.99	2.67	2.39	**4.10**	0.65	1.25	4.02
-S3	100	30	**6.50**	**10.91**	**4.05**	**2.70**	**2.42**	**4.10**	**0.68**	**1.27**	4.08
-S3	100	60	6.45	10.85	4.02	2.69	2.40	4.09	0.67	1.26	4.05
-S3	100	100	6.34	10.80	3.99	32.44	2.39	**4.10**	0.65	1.25	**7.75**

Table 13: BPR scores on MC22 tuning data for the VORM model. Best model scores per language and per ablation boldfaced.

	ces	eng	fra	hun	ita	lat	rus	avg.
morflogtoken	14.79	46.75	23.59	**35.03**	**12.14**	17.70	**12.34**	23.19
morftoken	10.37	40.52	20.94	34.13	11.23	17.68	10.60	20.78
morftype	**28.71**	**49.90**	23.63	**34.47**	11.84	**17.77**	11.46	**25.39**

Table 14: BPR scores on MC22 tuning data for Morfessor. Best model scores boldfaced

C Further quantitative breakdown of results

This Appendix supplements section 6 with the results broken down along several axes.

- Table 15 displays the results on the DoReCoMorph dataset, broken down per language.

	chip	morf	para	AG	vorm	S1	-C	S3
apah1238	90.2	85.8	83.8	80.3	86.6	88.9	86.6	69.0
arap1274	92.9	90.4	70.2	90.6	93.8	90.8	93.8	64.1
bain1259	95.8	77.1	80.3	93.1	92.6	93.8	92.6	53.1
beja1238	89.8	85.0	80.9	89.7	95.0	95.2	95.0	50.8
bora1263	87.3	80.4	67.3	87.2	93.5	94.3	93.5	35.4
cabe1245	94.2	82.6	82.7	77.0	92.5	89.5	92.5	82.5
cash1254	95.0	87.0	81.9	88.1	91.6	91.9	91.6	65.9
dolg1241	95.9	87.2	81.6	88.6	92.2	92.2	92.2	65.9
even1259	69.9	81.5	81.4	79.4	84.8	85.9	84.8	72.2
goem1240	95.5	90.7	84.2	80.5	94.3	87.3	94.3	88.4
goro1270	82.6	82.0	79.4	80.7	88.1	87.6	88.1	71.8
hoch1243	90.4	89.0	79.0	87.1	90.6	86.4	90.6	60.2
jeha1242	93.4	92.8	90.0	85.4	91.7	91.5	91.7	87.1
jeju1234	93.8	86.4	82.0	86.5	90.7	91.0	90.7	65.7
kaka1265	82.7	83.6	82.5	80.0	87.7	88.6	87.7	73.2
kama1351	95.3	90.7	87.5	91.6	95.5	95.3	95.5	68.8
komn1238	92.5	86.7	78.8	91.8	93.5	93.6	93.5	57.7
movi1243	89.8	86.2	76.1	85.1	89.2	87.0	89.2	62.6
ngal1292	94.8	89.2	67.9	87.7	83.6	78.3	83.6	60.5
nisv1234	94.7	89.4	85.7	87.2	90.3	90.1	90.3	75.9
nngg1234	91.6	89.5	80.5	71.3	87.3	82.9	87.3	87.8
nort2641	93.6	85.3	85.5	84.5	91.1	91.0	91.1	79.9
nort2875	86.8	84.2	79.6	84.7	89.8	89.0	89.8	69.0
orko1234	88.9	84.5	78.2	76.2	86.7	82.2	86.7	86.1
pnar1238	95.1	90.4	84.1	85.9	91.3	89.9	91.3	73.4
port1286	90.3	84.8	83.6	73.9	90.0	84.9	90.0	89.9
ruul1235	91.9	86.8	74.0	87.8	90.7	90.1	90.7	53.1
sanz1248	94.4	86.3	76.3	85.1	78.0	80.2	78.0	63.7
savo1255	90.9	88.3	83.8	88.4	90.7	90.5	90.7	70.0
sout2856	92.7	89.1	84.7	84.3	89.3	87.6	89.3	77.8
sumi1235	94.2	87.0	85.1	86.6	92.1	93.7	92.1	57.8
taba1259	91.8	82.5	81.3	86.4	88.2	83.0	88.2	64.6
teop1238	89.5	84.1	77.2	74.0	84.3	81.8	84.3	86.0
texi1237	92.2	80.0	76.2	85.5	87.9	85.8	87.9	62.9
trin1278	96.7	85.4	73.0	91.1	90.0	90.1	90.0	50.5
urum1249	95.8	89.3	86.1	86.4	92.1	92.4	92.1	69.1

Table 15: EMMA-2 results for the DoReCo dataset for Chipmunk (supervised), Morfessor, ParaMA2, MorphAGram, and Vorm (with ablation variants).