Spatial relation marking across languages: extraction, evaluation, analysis

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

Abstract

This paper presents a novel task, detecting Spatial Relation Markers (SRMs, like English ***in*** *the bag*), across languages, alongside a model for this task, RUIMTE . Using a massively parallel corpus of Bible translations, the model is evaluated against existing and baseline models on the basis of a novel evaluation set. The model presents high quality SRM extraction, and an accurate identification of situations where language have zero-marked SRMs.

1 Introduction

Massively parallel corpora, where the same source text has been translated into many different languages, form a unique opportunity to compare how the languages of the world express the same message, allowing for both fine-grained (utterance level) and large-scale comparisons (Mayer and Cysouw, 2012; Wälchli, 2014; Levshina, 2016; Asgari and Schütze, 2017; Liu et al., 2023). While earlier studies (e.g. Wälchli, 2014) applied manual extraction procedures, automated methods for retrieving the translation-equivalent markers across languages, at the level of parallel utterance tokens, are necessary to study crosslinguistic variation at scale. Substantial progress has been made in developing such methods (e.g. Wälchli, 2014; Asgari and Schütze, 2017; Liu et al., 2023; Beekhuizen et al., 2024). However, to date, none of these methods has undergone substantial intrinsic evaluation.

Here, I consider the typologically interesting and well-studied domain of spatial relation marking (i.e. prepositions like *in the cup* and *onto the mountain* in English and their translation equivalents; Levinson et al., 2003; Feist, 2008; Viechnicki et al., 2024). This domain can be expected to be challenging for automated extraction procedures due to the great degree of crosslinguistic variation in how languages divide up spatial relational meanings (Levinson et al., 2003; Feist, 2008; Viechnicki et al., 2024), the diversity of means of expression (as affixes or adpositions), and the extensive presence of zero marking of the spatial relation – marginally present in English expressions like *I'm _ home* and *I'm going _ home*, but prevalent in many other languages (Stolz et al., 2014; Haspelmath, 2019).

doculect	example	SRM(s)
English	**On** their heads were something …	on
Indonesian	… dan **di atas** kepala mereka …	di, atas
Quechua	… kansapa umankuna**pi** kurimanta	-pi
German	… und **auf** ihren köpfen wie …	auf
Kilivila	… leikatububulaisi paila kabilia **o**pwanetasi eisikamsi …	o-
Mixtec	… saá ná'a ña kánóo sini ñii ñii …	∅
Nigerian Pidgin	… war wetin dey dia head bi like gold	∅

Table 1: Equivalent Spatial Relation Markers in Rev. 7:9

This paper presents five contributions, with Materials at https://github.com/dnrb/ruimte.

- a dataset for evaluating the extraction of spatial relation markers (SRMs) in 18 languages;
- the formulation of two novel tasks for assessing the extraction quality (1) of the SRMs themselves, and (2) of the identification of the *absence* of any SRM;
- a novel model, RUIMTE , for extracting SRMs from a massively parallel corpus;
- an evaluation of it against comparable models;
- a brief demonstration of downstream insight of these results for the typology of spatial relation marking.

2 Data and seed set

For typological coverage, I used a corpus of Bible translations. As the Parallel Bible Corpus (Mayer and Cysouw, 2014) is not publicly available, Bible translations were downloaded through the API of faithcomesbyhearing.com (N=1,367), and from aboriginalbibles.org.au (N=14). Given translation availability, only New Testament data were used. Each unique translation (identified with

Proceedings of the 29th Conference on Computational Natural Language Learning, pages 571–585
July 31 - August 1, 2025

the language's 3-character ISO-939 code followed by a 3-character identifier) was considered one 'doculect'.[1] Under copyright agreements, translations cannot be reproduced but the materials contain a list of doculects.

To the best of my knowledge, there are no languages with spatial relation markers (SRMs) that are not polysemous with non-spatial (temporal and metaphorical/abstract) meanings, and as such no surface form of a (set of) markers can be used as a 'clean' seed to the extraction procedure. Instead, I manually created a seed set of exclusively spatial usages of English prepositions, as follows:

All prepositional phrases with nominal complements and one of the prepositions *in, on, at, to, onto, into, from, out, off* were identified in the World English Bible (ENGWEB) translation using the SpaCy dependency parser (Honnibal and Montani, 2017), and subsequently manually annotated for whether (1) whether they involved a spatial relation, (2) the **dynamicity** of the relation: static ('Locative'), dynamic towards a Ground ('Allative') or dynamic away from a ground ('Ablative'; using the terminology of Haspelmath, 2019), (3) the spatial **relation**, using the finest-grained categories of Levinson et al. (2003), and (4) the **ground type**, or the semantic class of the ground, using distinctions made in Haspelmath (2019) alongside several bottom-up identified categories. Examples are given in Appendix A.

3 Extraction methods

The task at hand is to extract, given a spatial relation marker (SRM) token in one doculect, all and only the translation equivalent SRMs in other doculects. Table 1 exemplifies the challenge: some languages have adpositions (English, German), others affixal SRMs (Kilivila, Quechua). Many doculects use one SRM, but Indonesian uses two, and Mixtec and Nigerian Pidgin use no explicit SRM in this example. In this section, I define several components of a model addressing this extraction problem. The critical components of the pipeline are: (1) morphological segmentation, (2) an alignment/extraction heuristic, and (3) the post-processing of allomorphy and complex SRMs.

[1] I adapt the term 'doculect' (Cysouw and Good, 2013) in order to reflect the fact that one 'language' (i.e., an ISO-939 code) may have multiple translations, each representing one 'documented variety', and to stress the somewhat tenuous relation between the documentation of a language through a text with potentially culturally foreign content that is often created (to a large extent) by a non-native speaker.

3.1 Morphological segmentation

Given that SRMs can be affixal, the extraction procedure should be able to consider affixes as candidate extractions. Some extraction procedures (e.g. Liu et al., 2023) already consider sub-word strings, but others might require the space-bound words to be further segmented into the stems and bound morphemes prior to extraction. Given that for the majority of the 1,381 doculects in the corpus no off-the-shelf morphological segmentation procedure is available, we will have to rely on unsupervised segmentation procedures that can be trained on the parallel Bible corpus itself. Here, I consider three unsupervised models.

MORFESSOR (Virpioja et al., 2013), first, can be trained on wordlists derived from the Bible corpus. Word frequencies affect the likelihood of segmentation in the model – I used the three suggested transformations of word frequency: the **type** model assigns a frequency of 1 to each type, **log-token** the log-transformed token frequency and **token** the token frequency. As MORFESSOR does not provide a distinction between stems and affixes, FLATCAT (Grönroos et al., 2014) was applied to determine the morphological status of the segments.

Second, MORSEL (Lignos, 2010) is a precision-oriented unsupervised procedure based on a best-first heuristic processing the space of possible morphological transforms. Like MORFESSOR, it relies on word lists and frequencies. I used the two predefined parameter settings **Aggressive** and **Conservative**, differing only in their approach to the detection of stem compounds.

Finally, VORM (Beekhuizen, 2025), is an unsupervised model that leverages translations to constrain the search space of morphological transforms and follows the intuition of MORSEL in making a best-first pass through the hypothesized morphological transforms. The minimal number of instances of a morphological transform was set to $N = 10$.

3.2 Extraction procedures

The second component of the pipeline is the extraction procedure itself. I will introduce the novel model, after which I present comparison models.

3.2.1 The RUIMTE extraction model

The Ground nouns and their translations form relatively easily identified anchors of spatial relation marking, with SRMs expected to occur close by. The RUIMTE ('**R**etrieval of **U**nique **I**nstances of **M**arkers of **T**opological **E**ssence') model leverages

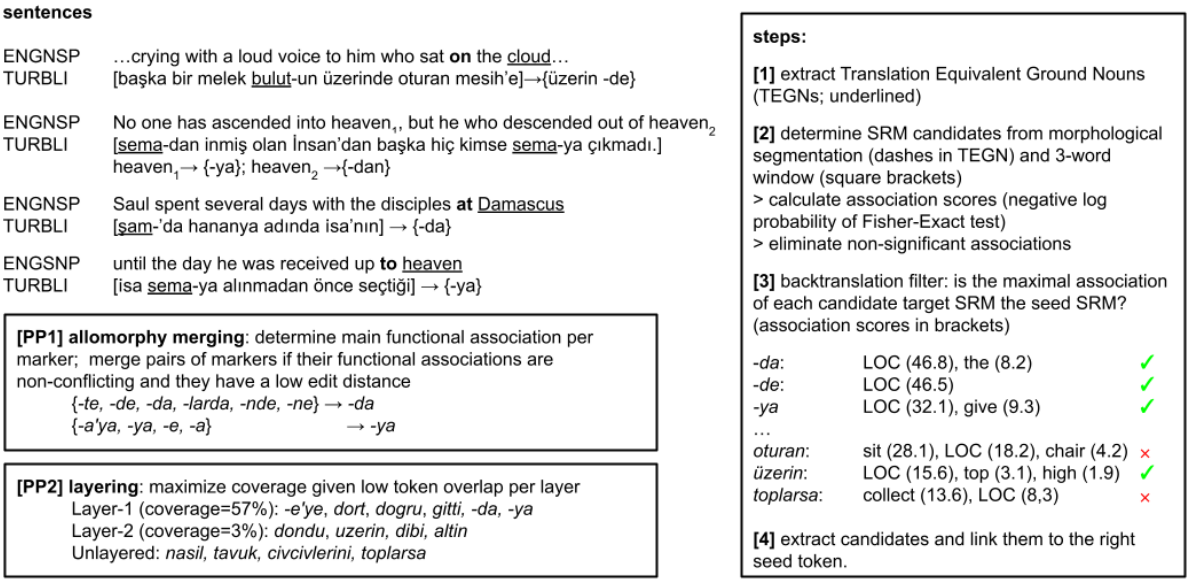

Figure 1: Exemplification of the RUIMTE model: main steps and postprocessing

this fact to identify SRMs, using intuitions similar to the noun-case extraction model of Weissweiler et al. (2022), namely that the overrepresentation inside the window of a target noun is a cue for extraction. Figure 1 provides an example, to follow along with the description of the steps below.

In **Step 1** of the procedure, translations of all Ground nouns are extracted with an adapted forward pass of the Liu et al. (2023) method. This procedure determines, for a seed noun type n, the character string $s_{\max}$ in the target language with the strongest statistical association to n, based on their co-occurrence frequency across Bible verses. More precisely, let U be the set of verses containing n, V the set of verses whose translation contains a character string s, and A the set of all utterances for which translations are available. The association between n and s is then defined as the negative log probability of a one-tailed Fisher Exact test over the following 2×2 table:

$$\begin{array}{ll} |U \wedge V| & |U \setminus V| \\ |V \setminus U| & |A \setminus (U \vee V)| \end{array}$$

Two constraints on $s_{\max}$ eliminate spurious associations, namely that $|U \wedge V| > 0.10 \times |U|$ and $|U \wedge V| > 0.10 \times |V|$. Next, any space-bound strings in the translations of utterances in U containing $s_{\max}$ are extracted as translation-equivalent ground noun tokens (TEGNs) of n, and the utterances in $|U \wedge V|$ are removed from U, A, and V. The procedure is repeated until no more valid candidates can be extracted.

Step 2, next, determines SRM candidates. For each seed item, all target language words in a 3-word window around each of the item's TEGN tokens (including the TEGN token) are retrieved and morphologically segmented using a morphological model. Each segment, combined with its position (whether it occurs before the TEGN, after it, or is part of it), forms a candidate SRM. Keeping track of the position is informative for the model, as adpositions typically occur on one side of the head noun. Given the extracted candidate SRMs, we then calculate the association of each candidate SRM type to the full set or a subset of the seed items. The association score of a candidate SRM is defined as the maximal negative log probability of a one-tailed Fisher Exact test applied to the following 2×2 table, maximizing U_{x} from the set of U and any U_{f} used:

$$\begin{array}{ll} |U_{\text{x}} \wedge C| & |N \wedge C| \\ |U_{\text{x}} \wedge D| & |N \wedge D| \end{array}$$

where:

- U is the set of verses containing a seed item,
- U_{f} is the set of verses containing seed items with a specific annotated feature value or combination of feature-values,
- N is the set of verses containing any seed noun in a non-prepositional (and therefore almost certainly non-locative) context,
- C is the set of verses whose translation contains the candidate SRM,
- D is the set of verses whose translation does not contain the candidate SRM.

We can define U_{f} variably. First, the spatial relation itself ('rel') can be used, defining three

nodes in the taxonomy of Levinson et al. (2003): $U_{\text{containment}}$ (annotated instances of 'IN-2D' and 'IN-3D'), U_{support} (instances of 'ON', 'ON-TOP', and 'ATTACHMENT'), and $U_{\text{colocation}}$ (instances of 'COLOCATION'). This allows us to find associations with SRMs that are exclusively used for one but not the other relation, effectively introducing a prior from the typological literature on what languages frequently do. Another such prior comes from the dynamicity ('dtc') of the spatial relation, defining three seed sets U_{static}, U_{goal}, and U_{source} (cf. Haspelmath, 2019). A third option is to combine them ('dtc&rel'). Fourth, we can use no U_{f}s ('all'), and finally, we can use the English prepositions ('prp') as an easily accessible seed type. We call these five settings the **seed types**.

SRMs with scores $< -\log 1^{e-6}$ are omitted. The resulting set still contains spurious markers. To remove these, **Step 3** implements a backtranslation filter that eliminates candidate target SRMs that are more strongly associated with frequently co-occurring context words of the seed SRMs (e.g. *go* in the context of *go into their house*). To do so, the forward-pass of the Liu et al. model is applied to each seed language word occurring with a frequency ≥ 10 in a 3-word window around any seed item noun (excluding the prepositions). This procedure gives us a statistical association score between the seed language word and the candidate SRMs that is comparable to the association score retrieved in the previous step. If a context word has a stronger association with a candidate SRM than the seed, that candidate SRM is deleted, as it is more likely a translation of a frequent context word of a spatial relation marker in the seed language.

Finally, **Step 4** takes one seed item token at a time, and finds the TEGN token whose candidate SRMs are most strongly associated with the U_{f} of the seed item. This ensures the correct extraction for sentences with multiple seed SRMs (e.g. the second sentence in Figure 1. A further constraint imposed is that only the highest-ranked affix is extracted, as spatial relations are not expected to be expressed through multiple affixes the same noun.

3.2.2 CONCEPTUALIZER

A first comparable model is the procedure of Liu et al. (2023), CONCEPTUALIZER, which, given a set of seed utterances in which a particular marker occurs, iteratively finds the substrings in a target language that are statistically most strongly associated with that set of seed utterances. As with the RUIMTE model, we can use various U_{f} independently as seeds and concatenate the results. In particular, I define the same five seed types as for RUIMTE. I further used the parameter settings cited in the paper ('original') as well as a loosening of some of the stricter settings ('bare': allowing up to 30 iterations, and only considering target-language substrings occurring in $\geq 0.1\%$ of U_{f}).

3.2.3 Alignment-based baseline models

Both the CONCEPTUALIZER and RUIMTE models are designed for the task of marker extraction in massively parallel corpora. As informed baseline models, I consider models based on unsupervised word alignment. Word alignment models allow us to create a bigraph between the seed language utterances and the (morphologically segmented) target language utterances. Extracting all aligned segments (words and affixes) to the seed SRMs can be expected to perform reasonably well as an extraction procedure. I apply two alignment procedures, EFLOMAL (Östling and Tiedemann, 2016) and FASTALIGN (Dyer et al., 2013), to the bitext between the seed language and each target language, retrieving any alignment to seed item prepositions.

As in previous models, we can vary the seed types, replacing the tokens of seed item prepositions in the bitext by a string identifying their feature representation. Similarly, we vary the morphological segmentation procedure used to preprocess the target language. Two final model parameters for aligner-based extraction procedures are symmetrization heuristics and a frequency filter (as proposed by Liu et al., 2023). Alongside the seed-to-target alignments ('fwd') and target-to-seed alignments ('rev'), we can consider their union and intersection, as well as three symmetrization heuristics that add and remove further alignments, namely 'diag-grow', 'diag-grow-final' and 'diag-grow-final-and' (implemented in `atools` Dyer et al., 2013). For the frequency filter, I consider no filter ('> 0'), an expected frequency of an extracted SRM given a seed type of more than 1 ('> 1'), or of more than 1% of the size of the set of seed types it was aligned to ('$> 1\%$').

3.3 Postprocessing

Two postprocessing steps were found to improve extraction quality on the development doculects (see Section 4.2). They are modular steps that can be applied to the outputs of any of the extraction procedures defined above. First, many lan-

guages display allopmorphy, either in their adpositions (German *in*, *im*, *ins*, 'in') or affixes (Turkish *-da/de/ta/te* 'locative case'). I define a simple heuristic to automatically **merge** these:

An agenda is initialized with all extracted SRMs per doculect ranked by frequency. Starting with the most frequent SRM, all remaining SRMs in the agenda are considered in turn, merging them with the current target SRM and its already-merged other SRMs if either they are formally near-identical (i.e., string identity after stripping diacritics and ignoring whether it is an affix vs. an adposition and preposition vs. postposition) or if they are formally possible allomorphs (i.e., having a low string edit distance) *and* have functionally similar patterns. Such allomorphs are then removed from the ranked list and the next marker is considered. At the end, all instances of allomorphs are replaced by the most frequent allomorph.

For the 'functional similarity' constraint, we consider per SRM (or cluster of already merged SRMs) which feature-value combination (from among whichever features are used in the seed type for that language; defaulting to both dynamicity and relation if the seed type was 'all' or 'prp') leads to the greatest Information Gain in classifying whether a seed item is translated with that SRM (of: one SRM from that SRM cluster) or not. If the values overlap for at least one feature (i.e., one SRM has 'containment' and 'colocation' for 'relation' as its feature-values optimally discriminating it, while the other has just 'containment') and do not contrast (i.e., the same two SRMs do not have non-overlapping values for the other feature, 'dynamicity' – e.g., 'static' for the first SRM and 'goal' for the latter), the two SRMs are considered functionally similar.

Second, complex adpositions were identified through **layering**: first, the set of SRMs that (1) minimally overlap with each other w.r.t. the TEGNs they occur with and (2) jointly cover the largest set of tokens is extracted as Layer-1, after which the procedure is repeated on the remaining markers to find a possible Layer-2. Affixal markers are eliminated from the second layer, as (obligatory) locative case marking should take place on the layer with the greatest coverage, and any affixes found on Layer 2 in the development doculects were false positives. Any unlayered markers are eliminated. Note that this step may aid in quality but was initially conceived for analytic purposes. Many languages have complex adpositions (e.g., Indonesian in Table 1) and being able to determine which adpositions form a paradigm is an important step in characterizing SRM systems.

4 Experimental set-up

4.1 Preprocessing

As not all Bibles come in Roman script, and as several morphological models depend on ASCII encoding, I transduced the text with an isomorphic mapping into ASCII for each doculect that at least partially used Roman characters (e.g. Vietnamese), and applied `unidecode` to transliterate the unicode characters into ASCII in other cases (e.g. Persian).

4.2 Annotation

Comparably little structured evaluation on the extraction of translation equivalent linguistic elements from massively parallel corpora has been carried out. Here, I introduce a dataset of 180 seed items for which I manually extracted (using grammars and dictionaries, alongside Google Translate) the SRMs in each of a typologically diverse set of 18 doculects. The 180 items were randomly sampled by selecting 60 seed items of each dynamicity value (locative, allative, ablative).

Zero coding was decided as follows: if no translation of the ground noun was found, the category 'noTEGN' was assigned. If the verbal predicate entailed the relation between the subject and the ground noun (like *enter* or *ascend*), or if some other non-spatial relation (commitatives like *with* and partitives like *of* are common) was marked, 'nonSpatialRelation' was assigned. Finally, if there was a translation-equivalent ground noun and a spatial relation to some verbal or nominal head that did not entail the relation, but no overt marking, 'trueZero' was assigned.

Table 5 in Appendix C presents the doculects, along with their top-3 markers and their proportion of zeros (both nonSpatialRelation and trueZero). I split the data into a development set of the top 9 doculects and a test set of the last 9, further only considering the even items of the development set when developing and tuning the various models.

4.3 Evaluation procedure

With these data, we can define an evaluation procedure. I formulate the extraction of SRMs and the correct identification of zero marking as separate tasks, given the interest in the typological literature in zero marking.

SRM extraction is a multi-class classification problem (multiple SRNs may simultaneously apply). Moreover, the strings in the extracted data may not match the annotated data exactly, due to variation in extraction and allomorphy. The proposed metric of evaluation is able to work with these constraints. First, given a set of extracted SRMs $E = \{e_1, e_2, \ldots, e_n\}$ and a set of annotated SRMs $A = \{a_1, a_2, \ldots, a_n\}$, each defining a set of seed tokens $U(x)$, where x is an SRM from A or E, we find the injective mapping M between A and E that maximizes the model's extraction accuracy, by maximizing the sum of the cardinalities of the intersections of $U(a_i)$ and $U(M(a_i))$, or: the seed items in which a_i occurs resp. the seed items in which some e_j in E, mapped to by $M(a_i)$ occurs.

With this mapping, we can determine, for each a_i mapping to some $e_j = M(a_i)$, how many True Positives ($|U(a_i) \cap U(e_j)|$), False Positives ($|U(e_j) \setminus U(a_i)|$) and False Negatives ($|U(a_i) \setminus U(e_j)|$) it has, and sum those across all $a_i \rightarrow M(a_i)$ mappings. The token count of any unmapped annotated items is added to the False Negatives, while unmapped extracted items are added to the False Positives. This allows us to define, for each doculect, the Precision, Recall, and F_1-score.

For the **evaluation of zero extraction** Precision, Recall, and F_1-score were defined as usual for a binary categorization problem. Predictions of zeros were compared against annotated cases, and counted as correct if 'trueZero' or 'nonSpatialRelation' was annotated; instances of 'noTEGN' were left out of consideration.

5 Results

5.1 Basic pipelines

I first consider the basic pipelines without postprocessing. The two alignment procedures have further hyperparameters like alignment symmetrization and frequency filtering that multiply out to a large number (1730) of unique models. The Materials present a full spreadsheet with performance per doculect for each unique model.

To narrow down the scope I consider only model components that perform substantially better than others; Figure 4 in Appendix D presents these comparisons. I only keep models that use (if applicable) the 'forward' symmetrization heuristic (which provides a good balance on the performance on both SRM extraction and zero extraction), and a frequency filter of $\geq 1\%$, which performs better than the other two filters on both tasks. The MORFESSOR 'token' and 'logtoken' models never performed as well as the 'type' model and were eliminated from consideration; similarly, using no morphological segmentation performed consistently worse. Between the alignment models, FASTALIGN had consistently lower scores than EFLOMAL and was not considered further.

Table 2 presents the performance of the remaining models on both SRM extraction (left columns) and zero extraction (right columns), reporting F_1-scores averaged over 18 doculects or over the 5 'zero doculects', i.e., doculects with $\geq 10\%$ zeros. The best-performing model on both tasks ($F_1 = 66.4$ for SRM extraction and $F_1 = 77.4$ for zero extraction) is the RUIMTE model using MORSEL-Aggressive. Notably, the optimal seed differs between the tasks: English prepositions ('prp') are optimal for SRM extraction while seeding on any spatial relation ('all') works better for zero extraction, due to the lower numbers of spurious markers extracted in the latter case. The VORM morphological model performs almost as well on both tasks at $F_1 = 66.1$ resp. $F_1 = 77.2$.

5.2 Effects of post-processing

Next, I assess the effect of the **postprocessing steps**, merging and layering, on the extraction of SRMs and zeros, considering the same models as in the last section. Table 3 presents the results, narrowing the seed types down to only 'all' and 'prp' as no optimal performances were found among the semantic-feature based seed types. We find that in particular the merging step has a positive influence on extraction. The best SRM-extraction models, i.e., VORM with a 'dtc' or 'prp' seed and using the merging step ($F_1 = 70.6$ resp. $F_1 = 69.6$), outperform the counterpart without merging (the 'basic' model) ($F_1 = 66.1$ resp. $F_1 = 64.2$) by > 4 points. Layering, however, does not appear to have the anticipated impact, with scores similar or somewhat lower than the basic model.

5.3 Performance analysis

What models by and large get right, are the most frequent SRMs per doculect, for which the statistical association is unequivocally strong. The remainder of this section considers where variation between the models was found and what the varying performance could be attributed to.

Looking at **individual doculects** might lead to insight in the variation. Figure 2 presents the Preci-

aligner; morphological model	SRM extraction					Zero extraction				
	all	dtc	dtc&rel	prp	rel	all	dtc	dtc&rel	prp	rel
CONCEPTUALIZER-bare	17.4	17.2	16.5	16.7	17.3	33.3	36.7	33.7	30.7	32.1
CONCEPTUALIZER-original	10.2	14.9	17.3	17.6	12.8	44.2	43.2	44.2	41.3	44.7
EFLOMAL; MORFESSOR-type	48.7	53.6	53.5	49.7	49.3	48.9	48.0	48.9	56.4	48.2
EFLOMAL; MORSEL-agg.	50.9	55.5	55.3	52.2	51.6	45.5	49.3	46.3	51.8	49.7
EFLOMAL; MORSEL-con.	50.1	54.3	54.4	50.4	50.6	45.1	47.6	49.7	53.6	47.4
EFLOMAL; VORM	49.7	55.2	55.4	50.5	51.5	42.3	44.4	45.2	50.7	45.8
RUIMTE; MORFESSOR-type	57.8	60.2	60.1	60.6	59.1	**76.9**	69.4	67.4	70.4	74.4
RUIMTE; MORSEL-agg.	62.5	65.0	65.2	**66.4**	62.6	**77.4**	65.0	64.1	66.8	72.4
RUIMTE; MORSEL-con.	61.6	64.1	64.4	**65.8**	63.0	76.3	65.6	64.9	67.6	72.5
RUIMTE; VORM	63.6	64.2	64.3	**66.1**	63.2	**77.2**	67.2	65.2	70.4	73.0

Table 2: **Basic models.** Mean F_1-scores for SRM extraction (left) and zero extraction (right) per combination of aligner and morphological model (rows) and seed type (columns) for the best-performing model components.

morphological model; seed	SRM extraction				Zero extraction			
	basic	merge	layer	both	basic	merge	layer	both
MORFESSOR-type; seed = all	57.8	58.4	55.1	55.7	**76.9**	**76.9**	75.8	75.8
MORFESSOR-type; seed = prp	60.6	61.5	59.8	59.3	70.4	70.4	71.1	70.1
MORSEL-aggressive; seed = all	62.5	64.7	58.4	60.6	**77.4**	**77.4**	75.9	75.9
MORSEL-aggressive; seed = prp	66.4	69.4	65.2	65.5	66.8	66.8	70.3	66.6
MORSEL-conservative; seed = all	61.6	63.9	58.1	60.4	76.3	76.3	75.7	75.7
MORSEL-conservative; seed = prp	65.8	68.1	64.9	64.9	67.6	67.6	71.0	67.4
VORM; seed = all	63.6	65.0	61.8	63.3	**77.2**	**77.2**	**76.6**	**76.6**
VORM; seed = prp	66.1	**70.6**	65.7	69.0	70.4	70.4	71.0	71.6

Table 3: **Postprocessing steps.** Mean F_1-scores for SRM extraction (left) and zero extraction (right) per combination of aligner and morphological model (rows) and combination of postprocessing steps (columns).

sion and Recall, per doculect, for a select number of models: the best-performing CONCEPTUALIZER, FASTALIGN, and EFLOMAL models (prior to post-processing) alongside the top-7 best performing RUIMTE models. For SRM extraction, the results for 5 illustrative doculects are shown in Figure 2 with the full set in Figure 5 in Appendix E. There are two types of doculects: those where the two alignment-based models (EFLOMAL and FASTALIGN) perform more like RUIMTE and those where they perform more like CONCEPTUALIZER.

This distinction seems to line up with the morphological expression of the SRMs: affixal SRMs, like in Cree and Kilivila, are not as well extracted with alignment-based extraction as with RUIMTE, while for doculects with primarily adpositional SRMs, like German and Vietnamese, the differences between alignment-based models and the RUIMTE are smaller. It is possible that he increased space of possible alignments for morphologically complex doculects decreases the alignment quality. This explanation is supported by the finding that the most-frequent SRM in Cree, *-ihk*, is in alignment-based models aligned to the seed item in only a subset of the cases for which it is annotated, suppressing the Recall. Moreover, 20+ unique non-spatial markers are (spuriously) aligned to the seed items, suppressing the Precision.

For zero marking (See Figure 6 in App. E), we notice, first, that the RUIMTE-based models achieve near-perfect Recall across the 5 zero-marking doculects, meaning that most annotated zeros are indeed extracted as zeros. The challenge, however, is Precision, i.e. : the model detecting a zero SRM where there is a non-zero SRM present. This effect is particularly strong for Bambara and Somali. For both, Precision is poor due to the high degree of polysemy of the SRMs (as noun classifiers in the former and highly general verbal particles in the latter) which leads to their spurious

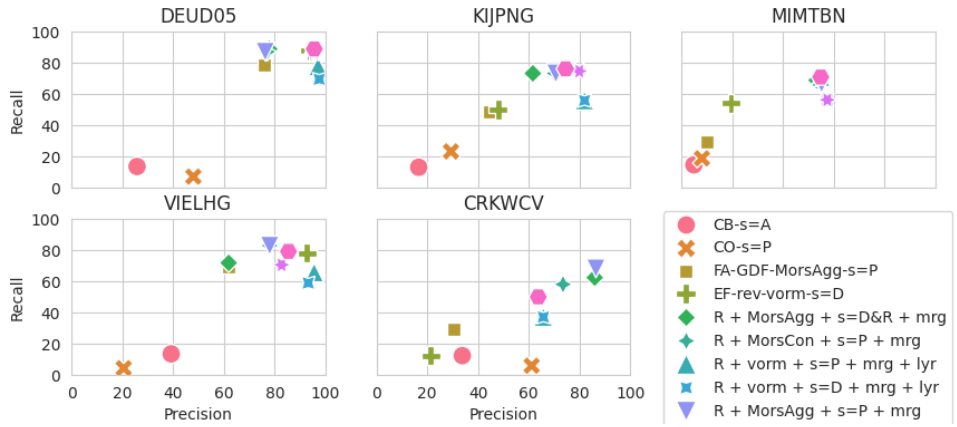

Figure 2: Precision and Recall on SRM extraction in 9 select models (see text) for 5 doculects.

presence in windows where the true marker is zero.

Considering doculects that show low performance, we find Mixtec, Finnish, and Malayalam, all with $F_1 < 70$ in any model. Weak performance for Mixtec can be attributed to the doculect's frequent use of zero marking. Lower-frequency prepositions (*sata, nandoso, ndaa*) did not reach the significance threshold of Step 2, and several spurious markers were instead extracted where zeros should be predicted. For Finnish and Malayalam, the morphological complexity might lead to an increased number of missegmented cases. Further, allomorphy (for Finnish) presents challenges. While allomorphy is generally resolved correctly, some non-allomorphs that are formally and functionally similar, such as *-lla/-llä* 'adessive' and *-lle* 'allative', tend to be merged erroneously. For Malayalam, a final challenge consists of its stacked locative cases, which the model cannot extract given the 'one-affix' constraint.

When considering the doculects for which the model performs well ($F_1 > 80$ on the top model, i.e., German, Indonesian, Dutch, Vietnamese, and Bambara), we find that the most common SRMs are correctly identified in a vast majority of cases (often well over 90%). Three sources of remaining errors can be identified. First, there are instances where the target-language SRM falls outside of the 3-word window around the Translation-Equivalent Seed Noun and is thus not extracted. Second, the 3-word window may contain spurious, but more strongly associated markers (e.g., when two adpositional phrases occur closeby to each other). Third, we find cases of failure to extract SRMs when they are either of low frequency or have a more frequent homonymous meaning (e.g. Dutch *te* 'to, at' is also the infinitive marker, like English *to*). In both scenarios, the association score with the seed SRMs in Step 2 is suppressed, leading to non-extraction.

5.4 Discussion

Reasonably good performance was achieved on the tasks of extracting SRMs and zeros across 18 doculects. Components of the best performing models included the novel RUIMTE extraction procedure and Precision-oriented morphological segmentation (MORSEL and the novel VORM model), as well as using the English prepositions as seed items. The latter was particularly surprising, given that most doculects do not encode spatial relations exactly along the lines of English SRMs.

Among the extraction procedures, CONCEPTUALIZER performed remarkably poorly, in contrast with its compelling performance as reported by Liu et al. (2023), as well as its reliability as a component in the RUIMTE model for extracting TEGNs. This suggests that CONCEPTUALIZER works well for lexical, open-class items, but not so much for more closed-class ones. Nonetheless, its components and general intuitions (regarding the use of co-occurrence statistics) translate well to this domain and form the engine of the RUIMTE model. Furthermore, Precision-oriented segmentation procedures such as MORSEL and VORM outperform the MORFESSOR baseline substantially for SRM extraction, suggesting that oversegmentation is harmful to the extraction, likely because it introduces noisy candidate SRMs.

6 Applications

To study potential use for typology, I briefly explore the best-performing model, RUIMTE +VORM,

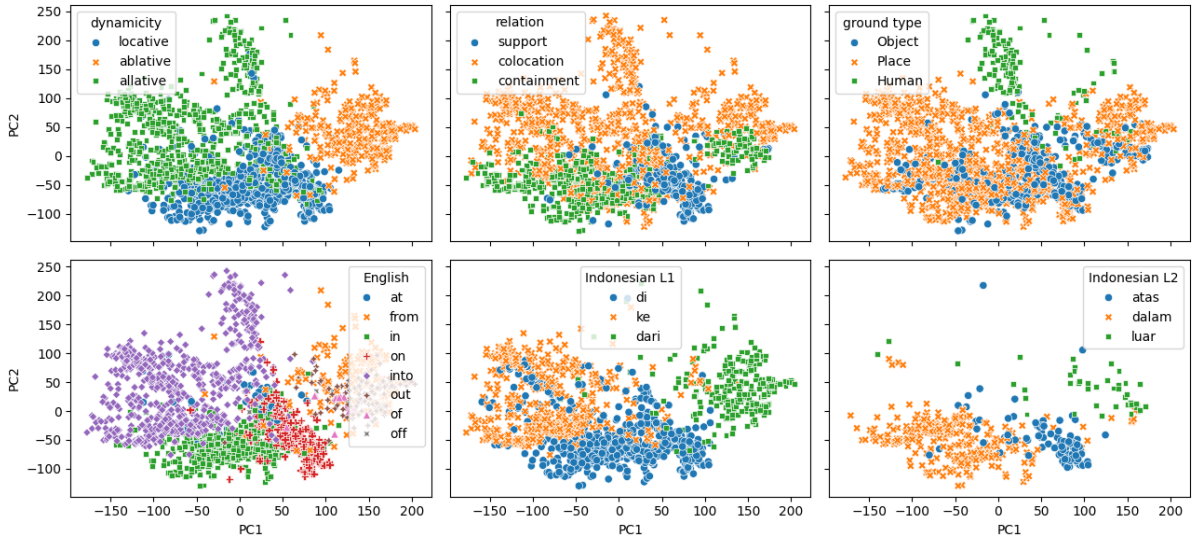

Figure 3: Exponential PCA space

with 'prp' seed type and merging. Markers in 1161 doculects were extracted. All seed items for which $> 66.7\%$ of doculects had no TEGN and subsequently all doculects for which $> 66.7\%$ of the seed items had no TEGN were eliminated, leaving a 1966×912 (tokens by doculects) table.

These extractions allow us to compare doculects on a token level. The main axes of crosslinguistic variation can be visualized with dimensionality reduction, here using Exponential PCA (Collins et al., 2001). Figure 3 shows 6 different colour codings of the same space, with each marker representing one seed item: the top three panels display the three annotated features – dynamicity, relation, and ground type. Notably, the values of all three features are reasonably separable in the 2-dimensional PCA space, meaning that there are doculects drawing SRM contrasts on the basis of each of these features. Indeed, we see some of those patterns play out in the two doculects in the right panels: while the English prepositions align with all three features, Indonesian neatly splits out dynamicity and relation between two sets (layers) of prepositions.

Second, these data allow us to assess typological claims concerning zero marking. Based on a survey of grammars, Stolz et al. (2014) conclude that zero marking is crosslinguistically more common in locatives and allatives than in ablatives. My data supports this finding. Per item, the proportion of zero-extractions across doculects was calculated. Aggregating those proportions, we find that the median locative item has zero marking in 34% (IQR: 27-41%) of doculects, the median allative in 31% (IQR: 25-38%), and the median ablative in 25% (IQR: 20-34%), with the differences between each pair significant (independent t-test; $P < .001$). It further suggests that locatives may be more commonly zero-marked than the allatives. While a fuller considerations is beyond the scope of this paper, these initial explorations demonstrate the potential for the typology of SRMs.

7 Conclusion

This paper introduces a novel procedure for extracting Spatial Relation Markers (SRMs) across languages, and evaluates it on a novel dataset. The method is shown to have good potential for answering typological questions about SRMs.

Interestingly, the components of my pipeline, as well as others working on similar tasks (Wälchli, 2014; Weissweiler et al., 2022; Liu et al., 2023), are mostly using simple statistics and best-first extraction procedures, rather than more complex Machine Learning techniques (e.g., seq2seq models) using global optimization. This is not a coincidence: the size of the data, combined with the nature of the hypothesis space (the number of SRMs – 0, 1, or 2; affixes vs. adpositions), makes extraction procedures that explicitly constrain the search space on priorly motivated grounds more successful.

This paper intends to contribute to the growing body of work on computational semantic typology with this paper, by introducing more rigorous evaluation techniques, providing annotated seed and evaluation data, and suggesting novel ways that spatial relation markers can be extracted.

Limitations

The work presented here was run on a corpus of Bible translations. The question whether the same methodology works well on other parallel corpora in different genres and dealing with different topics has not been positively answered, thus potentially constituting a limitation of the method that future work would have to settle.

Acknowledgments

This work has been over a decade in the making since its first conception, with the publication of Liu et al. (2023) and the subsequent invitation to talk about it at the CogSci 2023 symposium *Space in Context* (Grigoroglou et al., 2023) forming a major impetus to its completion. I am indebted to Suzanne Stevenson and Bernhard Wälchli for intermittent discussion over the years, and to Kit Donohue for contributing to the annotations of the seed set. Any inaccuracies and imperfections are, however, solely mine. I would further like to express gratitude to the three anonymous CoNLL reviewers who each provided helpful feedback.

References

Ehsaneddin Asgari and Hinrich Schütze. 2017. Past, present, future: A computational investigation of the typology of tense in 1000 languages. *arXiv preprint arXiv:1704.08914*.

Barend Beekhuizen. 2025. VORM: Translations and a constrained hypothesis space support unsupervised morphological segmentation across languages. In *29th Conference on Computational Natural Language Learning (CoNLL 2025)*.

Barend Beekhuizen, Maya Blumenthal, Lee Jiang, Anna Pyrtchenkov, and Jana Savevska. 2024. Truth be told: a corpus-based study of the cross-linguistic colexification of representational and (inter) subjective meanings. *Corpus Linguistics and Linguistic Theory*, 20(2):433–459.

Michael Collins, Sanjoy Dasgupta, and Robert E Schapire. 2001. A generalization of principal components analysis to the exponential family. In *Advances in Neural Information Processing Systems*, volume 14.

Michael Cysouw and Jeff Good. 2013. Languoid, doculect and glossonym: Formalizing the notion 'language'. *Language Documentation & Conservation*, 7.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Ramy Eskander, Smaranda Muresan, and Michael Collins. 2020. Unsupervised cross-lingual part-of-speech tagging for truly low-resource scenarios. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4820–4831, Online. Association for Computational Linguistics.

Michele I Feist. 2008. Space between languages. *Cognitive science*, 32(7):1177–1199.

Myrto Grigoroglou, Barbara Landau, Anna Papafragou, Ercenur Ünal, Kevser Kırbaşoğlu, Dilay Karadoller, Beyza Sumer, Asli Ozyurek, Barend Beekhuizen, Kenny R Coventry, Piotr J. Barc, Lucy-Amber Roberts, and Harmen Gudde. 2023. Space in context: Communicative factors shape spatial language. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

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

Martin Haspelmath. 2019. Differential place marking and differential object marking. *STUF-Language Typology and Universals*, 72(3):313–334.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Jane Klavan, Maarja-Liisa Pilvik, and Kristel Uiboaed. 2015. The use of multivariate statistical classification models for predicting constructional choice in spoken, non-standard varieties of estonian. *SKY Journal of Linguistics*, 28.

Stephen Levinson, Sérgio Meira, The Language, and Cognition Group. 2003. 'Natural concepts' in the spatial topological domain-adpositional meanings in crosslinguistic perspective: An exercise in semantic typology. *Language*, pages 485–516.

Natalia Levshina. 2016. Why we need a token-based typology: A case study of analytic and lexical causatives in fifteen European languages. *Folia Linguistica*, 50(2):507–542.

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.

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, Toronto, Canada. Association for Computational Linguistics.

Thomas Mayer and Michael Cysouw. 2012. Language comparison through sparse multilingual word alignment. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 54–62.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. *Oceania*, 135(273):40.

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.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Patrick Schone and Dan Jurafsky. 2000. Knowledge-free induction of morphology using latent semantic analysis. In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.

Thomas Stolz, Sander Lestrade, and Christel Stolz. 2014. *The crosslinguistics of zero-marking of spatial relations*, volume 15. Walter de Gruyter GmbH & Co KG.

Thomas Stolz, Nataliya Levkovych, and Aina Urdze. 2017. When zero is just enough... in support of a special toponymic grammar in Maltese. *Folia Linguistica*, 51:453–482.

Willy Van Langendonck. 2007. *Theory and typology of proper names*. Mouton de Gruyter.

Peter Viechnicki, Kevin Duh, Anthony Kostacos, and Barbara Landau. 2024. Large-scale bitext corpora provide new evidence for cognitive representations of spatial terms. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1089–1099.

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

Bernhard Wälchli. 2014. Algorithmic typology and going from known to similar unknown categories within and across languages. In Benedikt Szmrecsanyi and Bernhard Wälchli, editors, *Aggregating Dialectology, Typology, and Register Analysis: Linguistic Variation in Text and Speech*, pages 355–393. Walter de Gruyter.

Leonie Weissweiler, Valentin Hofmann, Masoud Jalili Sabet, and Hinrich Schuetze. 2022. CaMEL: Case Marker Extraction without Labels. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5506–5516, Dublin, Ireland. Association for Computational Linguistics.

A Seed annotation methods

The basis of the features used throughout this paper involves a token-level manual annotation of several dimensions of the spatial relation. While manual (semantic) annotation is admittedly prone to challengeable decisions and reductively frames the spatial relations as mutually exclusive, the paper hopefully demonstrates that it leads to useful and interpretable results.

The decision procedure for annotating the 'Relation' feature was based on the finest-grained foci (relations) of Levinson et al. (2003):

- 'Containment' was assigned if the boundaries of the Ground could be conceptualized as restricting the movement of the Figure in two ('In-2D'; in a boat, into his ear) or three ('In-3D'; in his house, into the prison) dimensions;
- 'Support' was assigned if either the Figure was higher than the Ground and not touching it ('Over'; [hang s/t] above his head) or if the surface contact with the Ground could be conceptualized as restricting the movement of the Figure – specifically, the relation annotated was:
  - 'On-Top' if contact was with the highest vertical region of the Ground (on the housetop, [put s/t] on his head);
  - 'Attachment' if contact was with a non-vertical region of the Ground and was maintained mechanically through an external source of Force (pinning, adhesion, clamping; on the stake; [sew s/t] on a garment),
  - 'On' otherwise (on the throne, [pour s/t] on his face)
- 'Under' was annotated if English uses under (under his feet, [take s/o] under her wing). Being a very small category, Under was excluded from the analysis.
- the residual category 'At3' breaks down into 'Near' relations, assigned whenever English uses near, at or by and no Figure-Ground contact is entailed (e.g., at the fire, [let s/o down] by the wall), and 'Colocation' otherwise. Notably, Colocation contains all cases where the Ground denotes a 'place', a (named) symbolically defined region (at Jerusalem, in Judea, to the country, in heaven), following Van Langendonck (2007) in considering their dimensionalities as irrelevant.

For the Dynamicity feature, 'Allative' was assigned whenever a (caused) motion predicate was present with a preposition marking a Ground that is a Goal of the motion predicate (*to, in(to), on(to)*), 'Ablative' if there was a (caused) motion predicate with a Source-marking preposition (*from*, *out of*, *off of*), and 'Static' otherwise.

For types of Grounds ('Ground Type'), I developed a categorization schema based on observations that Places, Named Places, and Human Grounds are occasionally marked differently from regular object-denoting Grounds (Stolz et al., 2017; Haspelmath, 2019), the existence of aquatic adpositions (Levinson et al., 2003), and that the mobility of the Ground affects the lexical choice (Klavan et al., 2015), as well as a bottom-up categorization of prevalent ground types in the corpus, distinguishing:

- Places: a region that is not easily conceptualized as a 'thing' but rather as inherently a 'location' of something else (to the place, in heaven), including Toponyms (in Asia, to Mount Sinai) and Buildings – an Object with unique relevance to humans as shelter, dwelling (in the temple, into the house);
- Object: a bound, countable physical, natural or artificial, entity (to the ship, in his hand);
- Human (bring him to the high priest)

B The intuition of the VORM segmentation model

The VORM model ('**V**ertaling **O**ndersteunt **R**edelijke **M**orfologie'; Dutch for 'Translations support reasonable morphology') is an unsupervised morphological segmentation procedure. Here, I present the intuition briefly; for a complete treatment see Beekhuizen (2025). Like MORSEL, VORM first finds recurrent character string transformations between pairs of words and makes a best-first pass through the word list to obtain derivations based on such transformations. However, only those pairs of words are inspected for the presence of potential transformations that are translation equivalents of the same word in a reference language (here: the seed translation ENGWEB). Distributional semantic information has long been used to bootstrap

doculect	Turkish	Finnish
word	*sofradakiler*	*polveutuu*
meaning	'them at the table'	'descends'
gold	sofra -da -ki -ler	polvi -ua -tuu
MORFESSOR-type	sofrada -kiler	polveutu -u
MORFESSOR-logtok.	sofrada -kiler	polve -utuu
MORFESSOR-token	sofrada -kiler	polveutuu
MORSEL-aggressive	sofradakiler	polveutu -u
MORSEL-cons.	sofradakiler	polveutu -u
VORM	sofra -da -ki -ler	polvi -i/ea -a/utuu

Table 4: Examples of the morphological models

morphological segmentation (Schone and Jurafsky, 2000; Narasimhan et al., 2015), but the proposal here is that translation is similarly a strong signal to constrain the unsupervised learning of morphological segmentation, as has been argued for other tasks, like PoS tagging (Eskander et al., 2020).

While this initial step provides a high-precision inventory of (sequences of) morphological transformations, many morphologically related words do not map onto the same translation equivalent in other languages and are as such not yet linked to each other. In a second step, all possible derivations of all words are generated, on the basis of the set of transformation sequences found in the first step. An agenda with all words is initialized, after which a best-first procedure finds the stem that has the largest morphological family size (i.e., occurs in the candidate derivations of the most words). The modeled words are removed from the agenda and the procedure is repeated until the agenda is empty.

Table 4 presents examples for the morphological extraction procedure applied to two words from Turkish and Finnish and compares it to the other models.

C Information on the annotated doculects

See Table 5.

D Model component comparison

See Figure 4 for an aggregated comparison of the model components.

E Doculect-level performance analysis

Figures 5 and 6 present the Precision and Recall per doculect, for all doculects on both tasks.

name (iso); affiliation	macroarea	most common three markers	% zeros
Plains Cree (CRKWCV)	Algic, North-America	*-ihk* (141) *oci* (39) *isi* (10)	7.3
German (DEUD05)	Indo-European, Eurasia	*in* (51) *aus* (34) *auf* (32)	1.2
Finnish (FINELC)	Finno-Ugric, Eurasia	*-an* (51) *-sta* (46) *-ssa* (24)	6.1
Indonesian (INDNTV)	Austronesian, Oceania	*di* (61) *dari* (52) *ke* (40)	3.9
Kilivila (KIJPNG)	Austronesian, Oceania	*o-* (67) *wa* (33) *metoya* (29)	28.5
Mixtec (MIMTBN)	Mixe-Zoque, North-Am.	*noo* (33) *ini* (7) *ndaa* (2)	62.0
Nigerian Pidgin (PCMTSC)	Creole, Africa	*for* (47), *from* (47) *inside* (7)	26.2
Somali (SOMSIM)	Afro-Asiatic, Africa	*ku* (59) *ka* (52) *soo* (35)	17.9
Turkish (TURBLI)	Turkic, Eurasia	*-ya* (57) *-dan* (53) *-da* (51)	3.4
Bambara (BAMLSB)	Mande, Africa	*la* (63) *bo* (41) *kono* (26)	23.2
Basque (EUSNLT)	isolate, Eurasia	*-an* (61) *-tik* (58) *-ra* (49)	1.7
Malayalam (MALNIB)	Dravidian, Eurasia	*-il* (101) *-kku* (34) *ninnu* (32)	10.1
Dutch (NLDDSV)	Indo-European, Eurasia	*in* (43) *uit* (34) *op* (31)	0.0
Persian (PESTPV)	Indo-European, Eurasia	*az* (49) *bah* (40) *dar* (38)	9.0
San Martín Quechua (QVSTBL)	Quechuan, South-Am.	*-pi* (77), *-manta* (41), *-man* (29)	0.0
Rundi (RUNBSB)	Niger-Congo, Africa	*mu* (83) *i* (35) *ku* (25)	9.5
Spanish (SPABDA)	Indo-European, Eurasia	*en* (59) *de* (44) *a* (34)	6.7
Vietnamese (VIELHG)	Austroasiatic, Eurasia	*tu* (35) *trên* (19) *o* (18)	9.0

Table 5: The 18 annotated doculects. The top nine are development doculects; bottom nine test doculects

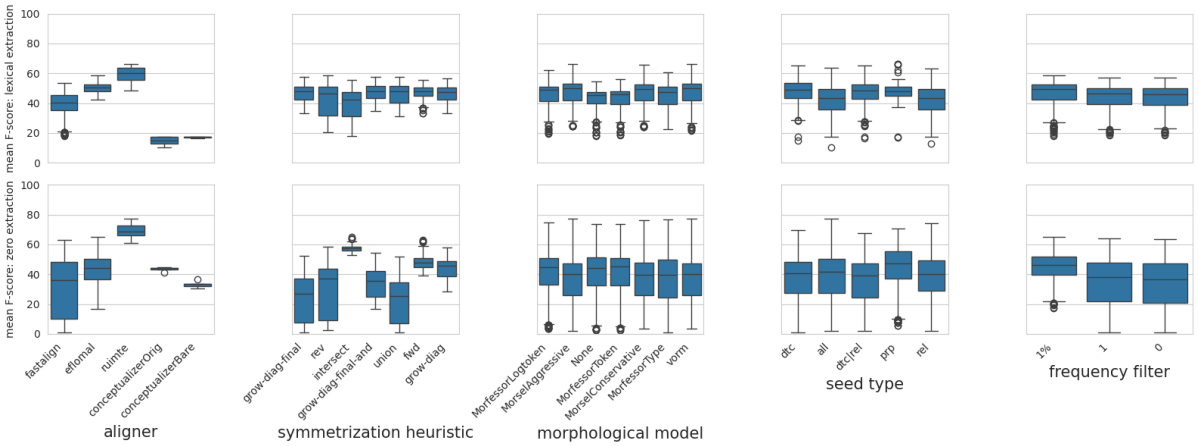

Figure 4: Comparison of model components (columns) on the performance (F-score, averaged across 18 doculects) of SRM extraction (top row) and zero extraction (bottom row).

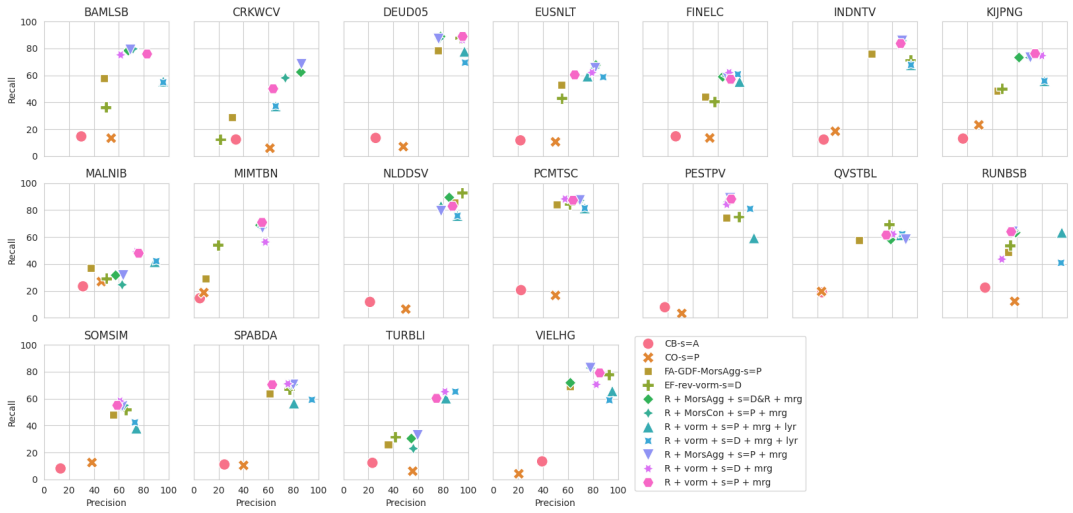

Figure 5: Precision and Recall on SRM extraction per doculect for the best models per aligner and the top-7 models of RUIMTE

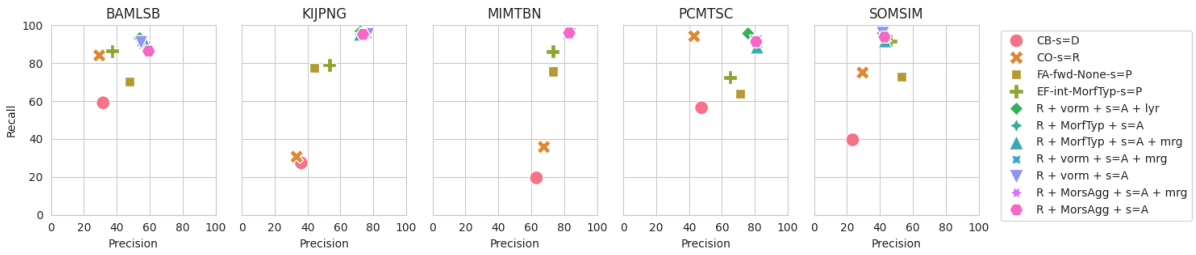

Figure 6: Precision and Recall on zero-marker extraction per zero-marking doculect for the best models per aligner and the top-7 models of RUIMTE.