A discovery procedure for synlexification patterns in the world's languages

Hannah S. Rognan
University of Toronto
Department of Linguistics
hannah.rognan@mail.utoronto.ca

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

Abstract

Synlexification is the pattern of crosslinguistic lexical semantic variation whereby what is expressed in a single word in one language, is expressed in multiple words in another (e.g., French *monter* vs. English *go+up*). We introduce a computational method for automatically extracting instances of synlexification from a parallel corpus at a large scale (many languages, many domains). The method involves debiasing the seed language by splitting up synlexifications in the seed language where other languages consistently split them. The method was applied to a massively parallel corpus of 198 Bible translations. We validate it on a broad sample of cases, and demonstrate its potential for typological research.

1 Introduction

Languages vary in how they 'package' the same conceptual content in words. Variation in colexification – a word in one language having two or more (partial) translation equivalents in another (e.g., English *blue* translating to Russian *sinij* 'dark blue' and *goluboj* 'light blue'), has been widely studied (François, 2008; Östling, 2016; Kemp et al., 2018). Another kind of variation occurs when a word in one language is, on the same occasion, translated as two or more words in another language. For example, French *monter* translates to English *go* and *up*. Here, the complex concept expressed by a single lexical item in one language is split into two constituent concepts in another language – i.c. English *go* expressing 'motion', and *up* the 'vertically elevated' nature of the goal location. While this kind of variation has been studied for individual cases, its generalization was only recently explicated by Haspelmath (2023), who dubbed the phenomenon 'synlexification', and its inverse 'circumlexification' (e.g., French *monter* synlexifies what English *go+up* circumlexifies).

Parallel corpora have been successfully used to investigate crosslinguistic patterns of colexification (Wälchli, 2014; Liu et al., 2023; Beekhuizen et al., 2024). However, extant computational approaches are by design unable to find cases of synlexification. Furthermore, existing corpus-based studies for individual cases do not allow for general discovery across semantic domains, which would be desirable to better understand the determinants of the typological variation in synlexification patterns. Our procedure aims to overcome these challenges.

In this paper, we first review corpus-based studies of synlexification across several semantic domains, motivating a more systematic approach. We then introduce a two-step model for automatically extracting synlexification patterns from parallel corpora. We validate the extracted patterns through comparison with documentary resources (grammars and dictionaries), and show that our method captures both many known and novel cases of synlexification. Finally, we present an initial exploration of the typological variation. Code and (shareable) data are available through https://github.com/dnrb/synlexification.

2 Background

2.1 Synlexifciation across domains

Motion verbs provide a well-established domain for studying synlexification, as languages vary in how they encode the manner of motion ('walking', 'rolling', 'going') and the path ('up', 'out', 'back'). Central here is Talmy (1991)'s distinction between satellite-framed and verb-framed languages. In the former (e.g., Germanic), manner is expressed through the verb and path through a particle, while verb-framed languages (e.g., Romance) encode the path directly in the verb, such as French *monter*, corresponding to *go* and *up* in English. Verkerk (2013) used a parallel corpus of Indo-European languages to examine crosslinguistic variation in

motion event expression.

Causatives are another domain in which typological differences in synlexification are prevalent (Levshina, 2015). Languages vary in whether lexicalize caused events as single verbs (e.g., *show* 'cause to see') or express them analytically (e.g., with one element expressing 'cause' and another 'see'). It has been found, using parallel and comparable corpora, that there is variation in the degree to which languages express different types of causation (e.g., 'making' vs 'letting'; Levshina, 2016) and different kinds of events (Haspelmath et al., 2014).

Light verbs form a third domain. Samardžić and Merlo (2010) use parallel corpora and word alignment procedures to investigate how English light verb constructions (e.g., *have a laugh*) align with single verbs in German (e.g., *lachen* 'laugh'). Their results reveal that such English constructions frequently map to one-word expressions in German. Nagy T. et al. (2020) extend this approach by automatically detecting cross-linguistic equivalents of light verb constructions in 4 languages. Both papers demonstrate that parallel corpora and word alignment techniques with automated decision procedures can highlight systematic variation in synlexification patterns across languages.

Negative verbs Different strategies for expressing negation have been found in the world's languages (Miestamo, 2007). One way to express negation is to combine a verb with a separate negative marker (e.g., *not+know*), another is to incorporate the negative meaning in a single word such as Tundra Nenets *yexara-* 'not know' (Nikolaeva, 2014, p. 285), and some words are inherently negative like *lack* and *refuse* (Miestamo, 2007). Some languages have been noted to deploy such synlexifying forms more than others (e.g., Ainu; Kwong, 2017), and some semantic domains are more likely to have synlexifying negative verbs (e.g. existentials; Veselinova, 2013).

Compounding, finally, is the morphological strategy of forming new lexical items from other lexical items. Languages vary in the extent to which they apply this strategy or instead choose to 'label' the concept (Štekauer et al., 2012), thus synlexifying what the compounding language circumlexifies. A notable case studied through parallel corpora are co-compounds, which consist of nouns that frequently occur in similar contexts (Wälchli, 2005, 2007), such as *hand-foot* meaning *limbs*.

These studies demonstrate the prevalence of synlexification across domains and languages, and validate the use of parallel corpus methods for identifying such patterns. However, these approaches focused on specific constructions or lexical domains, with top-down methods for detecting instances of the variation. Our approach proposes a bottom-up, scalable extraction method that identifies synlexification patterns across many languages and domains simultaneously, enabling both replication of known patterns and discovery of novel ones.

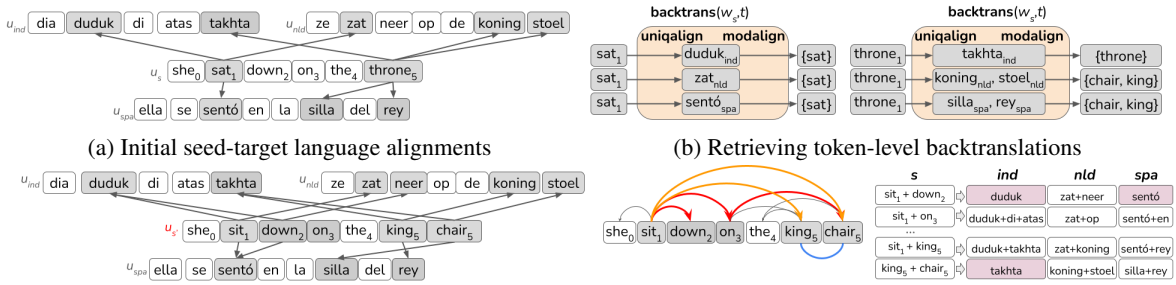
2.2 Explanations of synlexification patterns

Although explanations of the cross-linguistic variation in synlexification patterns has not been studied systematically, Haspelmath (2023) suggests Mańczak (1966)'s law of differentiation as a candidate explanation. This 'law' states that more frequently used meanings are more likely to be differentiated. The intuition is that more frequent groups of concepts are more likely to be synlexified, while less frequent groups of concepts are expected to remain circumlexified. While colexification patterns have been studied along these lines (e.g., Kemp et al., 2018), only initial evidence for the application of this idea to synlexification has been found in the form of the lexical vs. analytic causatives (Haspelmath et al., 2014).

Synlexification patterns are also expected to vary between languages. Ullmann (1966) notes that German tends to use more circumlexified forms than English or French. Aranovich and Wong (2023) distinguish between 'lexicological languages', such as Chinese, which tend to use more lexical items to express complex concepts, and 'grammatical languages', such as Sanskrit, which rely more on grammatical constructions. Seiler (1975) presents a similar distinction, but draws attention to the nature of the semantic operation, with some languages 'describing' (circumlexifying) complex concepts (e.g., Swedish *morbror* 'mother brother' and *farbror* 'father brother') and others 'labelling' (synlexifying) them (e.g. English *uncle*). Our approach can shed light on the extent to which languages as a whole tend to follow certain strategies.

2.3 Goals

To study patterns of synlexification at scale (many languages, many lexical fields), an automated extraction procedure is necessary. Existing automated procedures, all focussing on colexification patterns, include Wälchli (2014); Liu et al. (2023); Viechnicki et al. (2024) and Beekhuizen et al. (2024).



(a) Initial seed-target language alignments

(b) Retrieving token-level backtranslations

(c) Alignments with the granularized seed language

(d) Identifying synlexifications (purple)

Figure 1: Schematic illustration of the synlexification detection model

However, presumably as a means to restrict the search space, none of these procedures consider the alignment of one element in one language with multiple elements in another language, which is the goal of the current study. As such, this paper presents a novel approach that allows for the detection of instances of synlexification in a massively parallel corpus, i.e., a corpus where one text is translated into many languages.

By operationalizing a typological insight about the variable expression of 'the same' meanings through formal means, this paper further aims to contribute to the emerging field of (corpus-based) algorithmic typology (Wälchli, 2014; Wälchli and Sjöberg, 2024), in which typological concepts are investigated through formalization and quantification. This paper explores the notion of synlexification as proposed by Haspelmath (2023) by looking at the linguistic patterns that emerge when it is fairly directly applied to translated text.

After validating the output of this method, we present initial explorations of the kinds of insights this method can lead to: the nature of the variation of the occurrence of synlexification across lexical domains and languages, the discovery of novel domains of synlexification, and the functional determinants of the likelihood of synlexification.

3 A synlexification detection model

To find instances of variation in the lexification of the same concepts, we first have to identify such concepts. Given that translations of the same message can be expected to express more or less the same lexical-semantic content, we can use a bitext B_t between a seed language s and a target language t to provide us with **comparison concepts** (i.e., analytic concepts that allow us to compare languages without making a claim as to their language-specific validity; Haspelmath, 2018), and apply word alignment techniques (Tiedemann, 2011) to determine if there are recurrent many-to-one mappings between multiples of words in s and singleton words in t. This, then allows us to compare across target languages if the same multiples are aligned to singletons across target languages. Step 2 of the method describes this procedure.

However, a seed language may synlexify what target languages circumlexify, and one-to-many alignments from a seed language to a target language are not comparable: if English *enter* aligns with Dutch *ga+binnen* and German *tritt+ein*, we have no way of knowing whether Dutch and German circumlexify the complex concept expressed by *enter* similarly. To maintain the set-up of Step 2 (i.e., finding many-to-one s-to-t mappings), Step 1 creates a synthesized version of the seed language corpus in which seed language words are artificially circumlexified if other languages reliably do so.

3.1 Preliminaries

Several definitions will be used throughout. We define a **bitext** to a target language t as $B_t = [\langle u_s^1, u_t^1\rangle, \langle u_s^2, u_t^2\rangle, \ldots, \langle u_s^n, u_t^n\rangle]$, where u_s^i is a seed language utterance and u_t^i a sentence-aligned target language utterance. **Word types** in a (seed or target) language l are denoted $v_s \in V_s$ while **word tokens** are denoted $w_s \in W_s$. (Types and tokens need to be kept separate for several definitions). The function $\mathbf{type}(w^i)$ retrieves the type v associated with a word token w^i.

Given a set of word alignments between tokens of s and t, derived through some alignment procedure, the function $\mathbf{align}(w_l, l')$, then, retrieves the word alignments between a token $w_l \in u_l$ and a set of tokens $\{w_{l'}^i, w_{l'}^j, \ldots, w_{l'}^n\} \subseteq u_l'$ for a language pair $\langle l, l'\rangle$. A further function, $\mathbf{uniqalign}(w_l, l', Q)$ retrieves tokens aligned to w_l that themselves align exclusively to words in some set Q consisting of word tokens of l.

$\lvert\{w_s \lvert \textbf{type}(w_s) = v_s \wedge P \subseteq \textbf{backtrans}(w_s, t)\}\rvert$	$\lvert\{w_s \lvert \textbf{type}(w_s) = v_s \wedge P \not\subseteq \textbf{backtrans}(w_s, t)\}\rvert$
$\lvert\{w_s \lvert \textbf{type}(w_s) \neq v_s \wedge P \subseteq \textbf{backtrans}(w_s, t)\}\rvert$	$\lvert\{w_s \lvert \textbf{type}(w_s) \neq v_s \wedge P \not\subseteq \textbf{backtrans}(w_s, t)\}\rvert$

Table 1: Four quantities going into the Fisher Exact test to determine $\textbf{association}(v_s, P, t)$.

Formally, $\textbf{uniqalign}(w_l, l', Q) = \{w_{l'} | w_{l'} \in \textbf{align}(w, l') \wedge \textbf{align}(w_{l'}, l)/Q = \emptyset\}$. Two further functions build on the alignment functions. First, $\textbf{modalign}(v_l, l')$ ('modal alignments') returns the word type in l' that is most commonly (modally) aligned with v_l. Formally, $\textbf{modalign}(v_l, l') =$

$$\arg\max_{v_{l'} \in V_{l'}} |\{w_l | \textbf{type}(w_l) = v_l \wedge \exists w_{l'}.\textbf{type}(w_{l'}) = v_{l'} \wedge w_{l'} \in \textbf{uniqalign}(w_l, l')\}|.$$

Second, $\textbf{backtrans}(w_l, l')$ returns the set of modal alignments of the word tokens $\{w^i_{l'}, \ldots, w^n_{l'}\}$ given by $\textbf{uniqalign}(w_l, l', \{w_l\})$, or: the most-common backtranslations into l of w_l, given its alignments to tokens in l' that only align to w_l themselves. To exemplify, given a set of s-t alignments (Figure 1a), the $\textbf{backtrans}$ function (Figure 1b) retrieves the most-commonly aligned word of the target word tokens aligned with each of the seed language tokens.

3.2 Synthesizing a circumlexified seed corpus

We propose that cross-linguistically recurrent, and statistically reliable one-to-many alignments between the seed language s and the various target languages $t \in T$ allow us to replace synlexified concepts in s by synthesized circumlexifications. This procedure requires us to define what alignments are reliable and how to determine which tokens are replaced by a circumlexification.

First, for every seed language word type $v_s \in V_s$ with lexical meaning (here: nouns, adjectives and verbs) we retrieve all significantly associated potential circumlexifications, or: **paraphrases**, where a paraphrase $P = \{v^i_s, \ldots, v^n_s\}$, that is: a set of seed language word types (possibly including v_s itself), requiring $|P| \geq 2$ so that the paraphrase is into *more* words than the original. A paraphrase P is significantly associated with v_s given t, or: $\textbf{association}(P, v_s, t) = \top$, if a Fisher-Exact test over the 2×2 table in Table 1 yields a p-value below a pre-set threshold $\theta_{fe} \in (0, 1)$, and $\textbf{association}(P, v_s, t) = \bot$ otherwise.

Concretely, the Fisher Exact test assesses whether the association between v_s and P given a target language t is significant if the number of tokens of v_s whose modal backtranslations into s include P (top-left cell) is higher than expected by chance, that is: compared to (1) the set of tokens of v_s whose backtranslations do *not* include P (bottom-left cell), and (2) the set of tokens of other types whose backtranslations *do* include P (top-right cell). Following the example of Figure 1, if for Spanish the backtranslation *{king,chair}* occurs across many tokens of English *throne*, and *{king,chair}* infrequently occurs as the backtranslation of other English word types, the association between v_s = *throne* and P = *{king,chair}* is likely significant. Note that we use the inclusion of P in the backtranslation of w_s rather than the identity of P and $\textbf{backtrans}(w_s, t)$, because spurious backtranslations may occur in noisy alignments, thus weakening the $\langle v_s, P\rangle$ associations.

Retrieving significant $\langle v_s, P\rangle$ associations across target languages allows us, then, to leverage the crosslinguistic frequency of such association. If many target languages circumlexify v_s *in the same way* (i.e., backtranslating to the same paraphrase P), we have evidence that relevant tokens of v_s should be replaced by P, so that we would be able to identify that those languages circumlexify what other languages (including the seed language) synlexify. We approach this issue by iteratively replacing the seed language tokens whose types show significant v_s, P associations across the greatest number of target languages, as follows.

First, let $T_{\langle v_s, P\rangle}$ be the set of target languages for which $\textbf{association}(v_s, P, t) = \top$. We define the best word type and paraphrase pair $\langle v^{\max}_s, P^{\max}\rangle = \arg\max_{\langle v_s, P\rangle} |T_{v_s,P}|$, that is: the pair with the greatest number of languages for which it is significant. (Ties between $\langle v_s, P\rangle$ pairs are broken by the average p-value of the Fisher-Exact tests given the languages in $T_{v_s,P}$, prioritizing lower p-values).

Next, given $\langle v^{\max}_s, P^{\max}\rangle$, the set of **replaced tokens** is defined as $\{w_s | \textbf{type}(w_s) = v^{\max}_s \wedge \exists t.(t \in T_{\langle v^{\max}_s, P^{\max}\rangle} \wedge P \subseteq \textbf{backtranslate}(w_s, t))\}$, or: the set of tokens of v_s that backtranslate for at least one target language $t \in T_{\langle v^{\max}_s, P^{\max}\rangle}$ to a set of seed language types that include $P^{\max}$. These tokens are then replaced by $P^{\max}$ in a new corpus of the granularized seed language s', and removed from

W_s, after which the **association** mappings are recomputed. The procedure then repeats, calculating a novel $\langle v_s^{\max}, P^{\max}\rangle$, until $|T_{\langle v_s^{\max}, P^{\max}\rangle}| < \theta_{bt}$, i.e., the set of languages for which the $\langle v_l^{\max}, P^{\max}\rangle$ association is significant is smaller than some pre-set threshold $\theta_{bt} \in [1..\infty]$.

3.3 Finding reliable synlexification patterns

Next, the circumlexified seed language is word-aligned with each of the target languages (Figure 1c), and Step 2 involves finding reliable alignments between word pairs in s and words in t. To constrain the search space, we only consider pairs of seed language word tokens that meet two requirements. First, the pair contains one member with a lexical part of speech (here: nouns, adjectives, verbs, adpositions) and one member with either such a part of speech or a contentful satellite element (derivational affix, adverb, particle, proper noun). Second, the pair consists of elements of the same paraphrase P (blue line in Figure 1d for *king* and *chair*), or stand in a head-dependent relation to each other in a dependency parse (red lines; e.g., *sit* and *down*), including a second-order relation linking heads to the nominal dependents of any adpositions headed by the heads (orange lines; e.g., *sit* and *chair*). The first criterion restricts the search space to only parts of speech expressing lexical content, which is what we are centrally interested in. Second, all attested cases of synlexification (cf. §2) involve elements in a grammatical head-dependency relation to each other in circumlexifying languages, suggesting that this is a reasonable restriction of the search space.

For each pair of word tokens in the granularized seed language $\langle w^i_{s'}, w^j_{s'}\rangle$ meeting these criteria, we now retrieve the alignments in t using $\mathbf{uniqalign}(w^i_{s'}, l', \{w^i_{s'}, w^j_{s'}\})$. The **lexification** function, defined formally as

$$\mathbf{lexification}(w^i_{s'}, w^j_{s'}) =$$

$$\begin{cases} \text{synlexified} & \text{if } \mathbf{ua}(w^i_{s'}, t, S) \cap \mathbf{ua}(w^j_{s'}, t, S) \neq \emptyset, \\ \text{unlexified} & \text{if } \mathbf{ua}(w^i_{s'}, t, S) = \emptyset \wedge \mathbf{ua}(w^j_{s'}, t, S) = \emptyset, \\ i\text{-lexified} & \text{if } \mathbf{ua}(w^i_{s'}, t, S) = \emptyset \wedge \mathbf{ua}(w^j_{s'}, t, S) \neq \emptyset, \\ j\text{-lexified} & \text{if } \mathbf{ua}(w^i_{s'}, t, S) \neq \emptyset \wedge \mathbf{ua}(w^j_{s'}, t, S) = \emptyset, \\ \text{circumlex.} & \text{otherwise,} \end{cases}$$

(where $\mathbf{ua} = \mathbf{uniqalign}$, $S = \{w^i_{s'}, w^j_{s'}\}$ and 'circumlex.' is short for 'circumlexified') determines the lexification category. A pair of tokens is said to be **synlexified** if both tokens are **uniqalign**-ed to the same token(s) in t, **unlexified** if both tokens **uniqalign** to no words in t, i-**lexified** if $w^j_{s'}$ has no alignments but $w^i_{s'}$ does, j-**lexified** if, conversely, $w^i_{s'}$ has no alignments but $w^j_{s'}$ does, and **circumlexified** otherwise (i.e., both tokens have alignments in u_t but these sets do not overlap).

4 Experimental set-up

Corpus We test our model on a corpus of Bible translations gathered through the `bible.is` API. While this corpus has issues of ecological validity owing to the nature of the concepts expressed (being exogenous to many cultures) and the frequent production of these texts by non-native speakers (Pinhanez et al., 2023; Domingues et al., 2024), it has been used extensively in successfully identifying patterns of crosslinguistic lexical semantic variation that align with observations based on other data sources (Wälchli, 2014; Asgari and Schütze, 2017; Liu et al., 2023). Recognizing the non-identity between the translated, religion-oriented variety of a language and other, more ecologically valid, genres, we use the term **doculect** (Cysouw and Good, 2013) to refer to the variety of a language documented through translation. With these caveats, we treat the results as a lower-bound estimate of the real variation.

Preprocessing A sample of 198 doculects was derived through diversity sampling (Miestamo et al., 2016) ensuring areal and genetic diversity (see Appendix A for a list of doculects). The seed doculect, not part of the sample, was set to be the World English Bible translation. The text of this doculect was preprocessed by lemmatizing, PoS-tagging, and dependency-parsing it with SpaCy (Honnibal and Montani, 2017) and subsequently splitting derivationally complex words (e.g. *un-believe-able*) through CELEX2 (Baayen et al., 1996), as these reflect complex meanings that may be synlexified in other doculects. Target doculects were preprocessed by removing punctuation and segmented with VORM, a state-of-the-art unsupervised canonical morphological segmentation model (Beekhuizen, 2025b), which segments words into stems and affixes. Alignments for both s and s' were subsequently derived with Eflomal (Östling and Tiedemann, 2016), using the 'grow-diag-final-and' heuristic. The alignment in Step 1 was done with stems in the target doculects only (as inclusion of affixes at this state led to noise in the procedure), whereas the alignment for Step 2 also included affixes.

Parameters The significance threshold θ_{fe} was

set at 1^{e-6} and the minimum number of languages for which a the $\langle v_s, P\rangle$ association was significant was set as $\theta_{bt} = 3$. Both values were based on post-hoc assessment of the extraction quality and more complete parameter tuning on a benchmark set will be left for future research. Among the valid circumlexified seed language word pairs, only those that occurred ≥ 10 times throughout the circumlexified seed data were kept for further analysis.

5 Validating the model

Step 1 of the method splits 896 vocabulary types in the seed language, including some cases that are very frequently split among the target doculects, such as *answer=say+answer* (100 doculects) and *smoke=smoke+fire* (60 doculects). Notably, most splits involve splitting a word type into the word type itself and an additional element, though cases like *sail=go+boat* (34 doculects) are found as well. A larger sample is presented in Appendix B with the full data being available in the repository.

Based on the circumlexified seed doculect corpus, a total of $2,563$ comparison meaning pairs with a frequency ≥ 10 were found, and alignment patterns into the 198 target doculects were extracted with Step 2 of the extraction algorithm. While the next section demonstrates what can be done with these data, here we first provide a post-hoc validation of the model.

5.1 Validating extracted synlexifications

Given that no evaluation set is available, we validate the model by inspecting its extractions for several well-known cases, alongside several hand-picked ones representing frequently and infrequently synlexified meanings.[1] For each case, we selected one doculect whose predicted most-common strategy was to synlexify the meanings, one that most commonly circumlexified them, and one that most commonly left one meaning underspecified (i or j-lexified). For each doculect, we compared the extracted markers against grammars and dictionaries, referring to the translation tokens to validate. The fields and a qualitative description of the assessment can be found in Appendix C, while Table 2 summarizes the results. Although we had equal numbers of each predicted strategy initially, we conducted the evaluation multiple times as we implemented improvements to the model, which lead to a new strategy prediction for some cases. In addition, the model labeled $3/47$ inspected pairs as dominantly 'unlexified', which means no prediction regarding the modal type could be made. Overall, 80% of the 41 cases that were determined with certainty were correctly labeled by the model.

predicted strategy	correct	uncert.	incorrect
Synlexified (13)	0.85	0.15	0.00
Circumlexified (15)	0.80	0.00	0.20
x-lexified (16)	0.63	0.06	0.31

Table 2: Results of manual validation. 'uncert.'=uncertain; (N) = number of cases.

Among the accurate cases, we find the pair *enter+in*, synlexified as *natt* in Fulfulde (cf. McIntosh, 1984, p. 125: *natt-ay* 'enter'), circumlexified as *go+iin* in Jamaican English (cf. Bailey, 1968, p. 227), and j-lexified in Karkar as *mek* (cf. Rigden, n.d., p. 112, 116: *mek* 'in'). The majority of errors were i or j-lexifications which should have been labeled as cases of circumlexification. For instance, the pair *to+world* was labeled as underspecified in Bora because the pair frequently aligns only to *ííñujɨ́* (*land*) and in other cases to *-vu* (a spatial goal marker; Thiesen and Weber, 2012, p. 156), but the pair should have been aligned to both of these words to indicate circumlexification – an error attributable to the strictness of the **uniqalign** procedure, as many instances of these markers were found to have spurious alignments. Another type of error involved doculects synlexifying a concept but predicted to circumlexify. For instance, Ndyuka synlexifies the pair *un-+clean* with the word *takuu* meaning 'evil' (Huttar and Huttar, 1994, p. 62), but the model defines the tokens as circumlexified because in many instances, *takuu* is aligned with only one member of the pair ⟨*un-*, *clean*⟩, and other Ndyuka words with the other member.

6 Exploring synlexification patterns

The validation suggests that the method is a reasonable first attempt at extracting patterns of synlexification at a lexicon-wide scale and for a typologically diverse sample of doculects. Next, we explore applications of the extracted data, to demonstrate the linguistic use of the method.

[1] Notably, this validation step provides more evidence for the quality of the extraction than most other computational methods (e.g. Liu et al., 2023; Beekhuizen et al., 2024), though see Beekhuizen (2025a) and Beekhuizen (2025c) for thoroughly evaluated extraction algorithms.

part of speech pair	N	% syn.	top-3 most frequently synlexified (N doculects)
Adposition+Noun	461	18%	mountain+on (116) before+foot (80) in+peace (72)
Adposition+Verb	460	19%	rise+up (107) get+up (104) down+fall (95)
Noun+Verb	408	38%	bread+eat (173) law+write (164) apostle+send (163)
Verb+Verb	245	20%	deceive+lie (162) suffer+torment (146) persecute+suffer (120)
Noun+Noun	198	81%	boat+ship (186) boat+sea (186) horse+soldier (184)
Adjective+Noun	87	68%	blind+eye (179) blood+dead (178) famine+hungry (172)
Affix+Verb	86	90%	teach+-er (142) serve+-ant (105) pray+-er (105)
Proper Noun+Verb	84%	10%	Peter+answer (2) Jesus+answer (1) Christ+die (1)

Table 3: Synlexification across PoS pairs. N= number of pairs, % syn.= % of pairs synlexified in ≥ 1 doculect.

Distribution across the lexicon. Most (1715/2563, or 67%) comparison concept pairs are not dominantly lexified in any doculect (where 'dominant' means 'applied in $\geq 50\%$ of the tokens of that pair'). This suggests that synlexification happens in select areas of the lexicon. Breaking down the pairs by their grammatical categories (Table 3; a larger sample is given in Table 10 in App. D), we find substantial variation: combinations of proper nouns and verbs are for instance rarely synlexified (9%). Conversely, many of the noun+noun, adjective+noun, and affix+noun pair have at least one doculect synlexifying them, possibly due to such combinations building complex categories that can variably be 'described' or 'labelled' (cf. Seiler, 1975) for communication to succeed.

Motion verb synlexification, part of the preposition+verb combinations, can be found among the most frequently dominantly synlexified preposition+verb pairs – e.g., *rise+up* or *down+fall*, but other preposition+verb combinations reflect more 'accidental' combinations of verbs and prepositions, making preposition+verb pairs have a low number of synlexifying doculects. While looking at the level of grammatical categories is likely too coarse a subdivision, the variation across grammatical categories suggests that some of the uneven distribution across the lexicon may be related to the types of concepts they denote.

Areal distribution. Secondly, not all doculects are equally likely to synlexify, as discussed in §2. There are substantial areal patterns, with the average number of comparison meaning pairs dominantly synlexified ranging from 122 (Australian doculects), and 150 (South-America), over 173 (North-America) and 190 (Papunesia), to 215 (Africa) and 228 (Eurasian). Figure 2 plots the number of dominantly synlexified pairs across the 198 doculects. These areal patterns are open to multiple interpretations. The high numbers for the European doculects (Basque, Dutch, Finnish, Hungarian, Greek) might reflect the extended exposure of these cultures to the cultural concepts of Christianity ('pray', 'temple', 'prophet', ...), leading to short, synlexifying forms. However, not all variation can be attributed to cultural factors, as there is substantial variation between other macro-areas where the dissemination of these religious concepts is more recent. Moreover, the clearly religious concepts form only a small subset of all variably synlexified concepts.

Potential for case studies. Synlexification patterns have mostly been studied for specific semantic domains (cf. §2). The proposed procedure allows us to study such cases by retrieving matching comparison concept pairs. The well-studied case of **motion events** can for instance be studied by looking for motion verbs (*go, fall, sit, put, ...*) and particles (*in, out, up, down, ...*). For most such pairs, which are presented in Table 11 in Appendix E, doculects do not synlexify. Most frequently synlexified are five pairs of motion along the vertical axis: *rise+up* ($N = 107$), *get+up* ($N = 104$), *fall + down* ($N = 95$), *sit + down* ($N = 81$), and *stand + up* ($N = 67$). Notably, in some of these cases the direction of movement is already implicated by the manner of motion verb. These cases raises interesting questions about the concept of synlexification per se. If one language l circumlexifies this complex concept into a pair of lexical items $v_l^{\text{fall}}, v_l^{\text{down}}$, aligning with 'fall' and 'down', and another language l' synlexifies them with one lexical item $v_{l'}$, which dictionaries define as 'fall', does this mean that the meanings of v_l^{fall} and $v_{l'}$ are different with the former underspecifying the 'down' component? This seems counterintuitive: after all, even in a synlexifying language like English, *He fell* (as opposed to *He fell down*) at least implicates and perhaps entails 'down'.

Conversely, several of the cases for which sub-

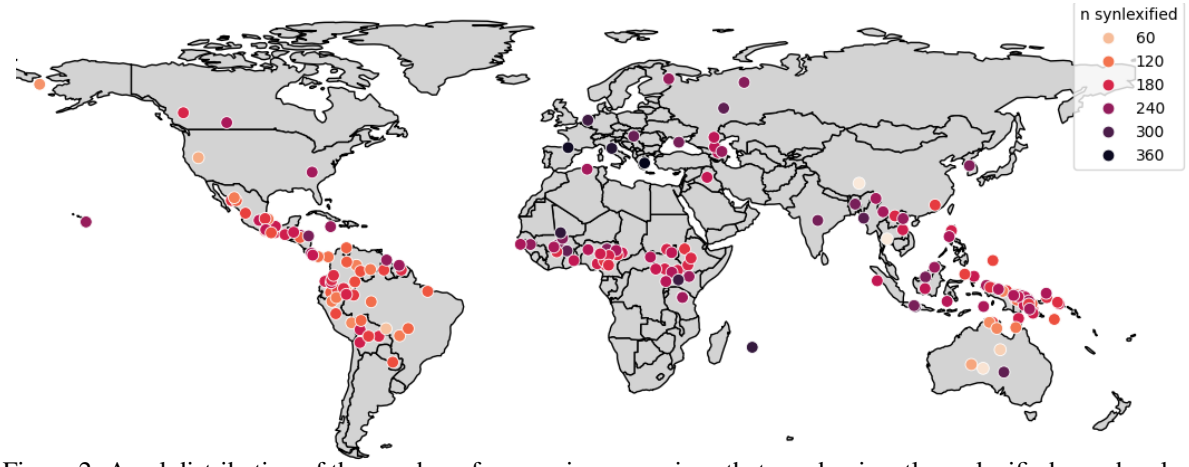


Figure 2: Areal distribution of the number of comparison meanings that are dominantly synlexified, per doculect

stantial typological variation is expected (*enter* + *in* and *go* + *out*) were dominantly synlexified only rarely across doculects ($N = 3$ resp. $N = 2$). While further validation and linguistic analysis is necessary, these data suggest that matters are more complex than the simple path vs. satellite-framing typology lets on.

The bottom-up discovery procedure further allows us to explore new domains involving variation in the synlexification patterns. Speech events form one such domain: **manner of speech verbs**, such as *promise*, *lie*, *answer* and *ask* are often found synlexified, like in English, but more frequently (across doculects) circumlexified into an element translating to English *say* and another to the manner (i.e., *promise*, *false*, *answer*, *ask*). Table 12 in Appendix E presents an overview. While typological observations about speech verbs have been made for small sets of languages (Caballero and Paradis, 2017), the method presented here supports a larger-scale typological comparison.

Mańczak's law of differentiation. Finally, we explore the hypothesis that more frequent meaning pairs are more likely to be synlexified, due to communicative efficiency. We evaluate this hypothesis with the following logistic regression model:

$$\begin{aligned}\text{synlexified} \sim\ & \log \text{pair.frequency} + \text{pos} + \\ & \text{macroarea} + (1|\text{doculect}) + (1|\text{pair})\end{aligned}$$

That is: for each doculect and for each pair, we predict whether the doculect dominantly synlexifies the pair on the basis of the log-frequency of the comparison meaning pair, as derived in Step 2, the part of speech ('pos'; dummy-coded for the 5 most frequent parts of speech pairs, with other pos-pairs coded as 'other'), and the macroarea (dummy-coded). Random intercepts were added for doculects and pairs, reflecting biases of individual doculects or pairs that should be included to constrain the inferred effects of the target variables.

Table 13 in App. F presents full regression results. Critically, over and above significant effects of 'pos' and 'macroarea', the frequency of the meaning pair significantly predicts the likelihood of that pair being synlexified, with the positive direction being in line with Mańczak's law of differentiation. The effect size is furthermore substantial: the observed log Odds Ratio of 1.203 means that for every unit increase in log frequency (e.g, going from $\log N = 3$ to $\log N = 4$, or: $N \approx 20$ to $N \approx 54$), the the likelihood of synlexifying the pair increases more than threefold ($\exp 1.203 \approx 3.330$). Two concerns here are whether the variably-lexified comparison concepts have enough ecological validity and whether the counts of the English-based comparison meaning pairs are a valid measure of meaning frequency. Addressing these would be paramount to further research.

7 Conclusion

This paper introduced a novel method for extracting patterns of synlexification from a parallel corpus at the scale of 198 languages and the full lexicon and validated it on over 40 cases. While the model performed generally well, substantial room for improvement remains. First, replacing seed language words by other seed language words in Step 1 means that the (co)lexification pattern of the seed language still affects what alignments are likely to be made. Explorations of methods that infer latent discrete n-tuples (e.g., through topic modelling, cf. Blei and Lafferty, 2009) prove difficult to tune to yield desired results. In future work, we hope to develop such improvements, create more rigorous methods of evaluation, and apply the method to more ecologically valid corpora.

References

Faruk Abu-Chacra. 2007. *Arabic: An Essential Grammar*. Routledge, Milton Park, Abingdon, Oxon; New York, NY.

Raúl Aranovich and Alan Wong. 2023. Saussure's cours and the monosyllabic myth: the perception of chinese in early linguistic theory. *Language & History*, 66(1):59–79.

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*.

Gorka Aulestia. 1989. *Basque-English dictionary*. Basque series. University of Nevada Press, Reno.

Yves Avril. 2006. *Parlons Komi*. Parlons. L'Harmattan, Paris.

Nicholas Awde and Muhammad Galaev. 2014. *Chechen-English, English-Chechen Dictionary and Phrasebook*. Routledge, Abingdon, Oxon and New York, NY. Originally published by Curzon Press Ltd, 1997.

R Harald Baayen, Richard Piepenbrock, and Leon Gulikers. 1996. The celex lexical database (cd-rom).

Beryl Bailey. 1968. *Jamaican Creole Language Course*. Peace Corps, Washington, D.C.

Barend Beekhuizen. 2025a. Spatial relation marking across languages: extraction, evaluation, analysis. In *29th Conference on Computational Natural Language Learning (CoNLL 2025)*.

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

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.

Keith Berry and Christine Berry. 1999. *A Description of Abun: A West Papuan Language of Irian Jaya*, volume 115 of *Pacific Linguistics: Series B*. Research School of Pacific and Asian Studies, Australian National University, Canberra.

David M Blei and John D Lafferty. 2009. Topic models. In *Text mining*, pages 101–124. Chapman and Hall/CRC.

Ross Bowden. 1997. *A Dictionary of Kwoma: A Papuan Language of North-East New Guinea*, volume 134 of *Pacific Linguistics: Series C*. Research School of Pacific and Asian Studies, Australian National University, Canberra.

David Briley. 1997. Four grammatical marking systems in bauzi. In Karl J. Franklin, editor, *Papers in Papuan Linguistics No. 2*, pages 1–131. Research School of Pacific and Asian Studies, Australian National University, Canberra.

Rosario Caballero and Carita Paradis. 2017. Verbs in speech framing expressions. *Journal of Linguistics*, 22(1):1–40.

Eugene H. Casad. 2012. Cora–spanish lexical database (draft). Manuscript. Unfinished lexical database, posted as is without peer review.

Rodolfo Cerrón-Palomino. 2006. *El Chipaya o Lengua de los Hombres del Agua*, 1 edition. Fondo Editorial, Pontificia Universidad Católica del Perú, Lima.

Michael Cysouw and Jeff Good. 2013. Languoid, doculect and glossonym: Formalizing the notion 'language'. *LANGUAGE DOCUMENTATION & CONSERVATION*, 7.

Rudolf Pieter Gerardus de Rijk. 2008. *Standard Basque: A Progressive Grammar*. MIT Press, Cambridge, MA.

René Dirven, Louis Goossens, Yvan Putseys, and Emma Vorlat. 1982. *The scene of linguistic action and its perspectivization by speak, talk, say and tell*. John Benjamins.

Pedro Henrique Domingues, Claudio Santos Pinhanez, Paulo Cavalin, and Julio Nogima. 2024. Quantifying the ethical dilemma of using culturally toxic training data in AI tools for indigenous languages. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 283–293, Torino, Italia. ELRA and ICCL.

Norm Draper and Sheila Draper. 2002. *Dictionary of Kyaka Enga: Papua New Guinea*, volume 532 of *Pacific Linguistics*. Pacific Linguistics, Canberra.

Tom E. Dutton and Dicks Thomas. 1994. *A new course in Tok Pisin (New Guinea Pidgin)*, volume 67 of *Pacific Linguistics : Series D, Special Publications*. Australian National University, Canberra.

Irwin Firchow. 1974. Rotokas grammar. Manuscript. Accessed via SIL International.

Alexandre François. 2008. Semantic maps an the typology of colexifications: Intertwining polysemous networks across languages. In Martine Vanhove, editor, *From polysemy to semantic change: Towards a typology of lexical semantic associations*, pages 163–216. John Benjamins, Amsterdam.

Martin Haspelmath. 2018. How comparative concepts and descriptive linguistic categories are different. In Daniël Olmen, Tanja Mortelmans, and Frank Brisard, editors, *Aspects of linguistic variation*, pages 83–114. De Gruyter Mouton.

Martin Haspelmath. 2023. Coexpression and synexpression patterns across languages: comparative concepts and possible explanations. *Frontiers in Psychology*, 14.

Martin Haspelmath, Andreea Calude, Michael Spagnol, Heiko Narrog, and Elif Bamyaci. 2014. Coding causal–noncausal verb alternations: A form–frequency correspondence explanation. *Journal of linguistics*, 50(3):587–625.

Steffen Haurholm-Larsen. 2016. *A Grammar of Garifuna*. Ph.D. thesis, Universität Zürich, Bern.

Jeffrey Heath. 1999. *A Grammar of Koyra Chiini: The Songhay of Timbuktu*. Number 19 in Mouton Grammar Library. Mouton de Gruyter, Berlin and New York.

Jeffrey Heath. 2014. A grammar of yorno-so. Draft manuscript, November 2014.

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

George L. Huttar and Mary L. Huttar. 1994. *Ndyuka*. Descriptive Grammars Series. Routledge, London and New York.

Charles Kemp, Yang Xu, and Terry Regier. 2018. Semantic typology and efficient communication. *Annual Review of Linguistics*, 4(1):109–128.

Yan Kit Kwong. 2017. Lexical negative verbs in ainu language. *International Journal of Humanities and Social Science*, 7(2):166–170.

Natalia Levshina. 2015. European analytic causatives as a comparative concept: Evidence from a parallel corpus of film subtitles. *Folia Linguistica*, 49(2):487–520.

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.

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.

Arnold Lock. 2011. *Abau Grammar*, volume 57 of *Data Papers on Papua New Guinea Languages*. SIL-PNG Academic Publications, Papua New Guinea. Summer Institute of Linguistics.

Martin Maiden and Cecilia. Robustelli. 2013. *A reference grammar of modern Italian*, 2nd ed. edition. HRG. Routledge, London ;. Includes bibliographical references and index.

Witold Mańczak. 1966. La nature du supplétivisme. *Linguistics*, 4(28):82–89.

Mary McIntosh. 1984. *Fulfulde Syntax and Verbal Morphology*. KPI, in association with University of Port Harcourt Press, London and Boston.

Matti Miestamo. 2007. Negation – an overview of typological research. *Language and Linguistics Compass*, 1(5):552–570.

Matti Miestamo, Dik Bakker, and Antti Arppe. 2016. Sampling for variety. *Linguistic Typology*, 20(2):233–296.

István Nagy T., Anita Rácz, and Veronika Vincze. 2020. Detecting light verb constructions across languages. *Natural Language Engineering*, 26(3):319–348.

Irina Nikolaeva. 2014. *A Grammar of Tundra Nenets*.

Colleen Alena O'Brien. 2018. *A Grammatical Description of Kamsá: A Language Isolate of Colombia*. Ph.D. thesis, University of Hawai'i at Mānoa, Honolulu.

Asmah Haji Omar. 1969. *The Iban language of Sarawak: a grammatical description*. Ph.D. thesis, University of London.

Robert Östling. 2016. Studying colexification through massively parallell corpora. In Paeivi Juvonen Maria Koptjevskaja-Tamm, editor, *The lexical typology of semantic shifts*, chapter 6, pages 157–176. De Gruyter Mouton.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with markov chain monte carlo. *The Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Helma Pasch. 2007. Grammar of location and motion in zande. *APAL (Annual Publication in African Linguistics)*, 5.

Claudio S. Pinhanez, Paulo Cavalin, Marisa Vasconcelos, and Julio Nogima. 2023. Balancing social impact, opportunities, and ethical constraints of using ai in the documentation and vitalization of indigenous languages. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 6174–6182. International Joint Conferences on Artificial Intelligence Organization. AI for Good.

David J. Prentice. 1971. *The Murut Languages of Sabah*. Number 18 in Pacific Linguistics: Series C. Research School of Pacific and Asian Studies, Australian National University, Canberra.

Veda Rigden. n.d. Karkar grammar essentials. Unpublished manuscript, SIL.

Danilo Salamanca. 1988. *Elementos de gramática del Miskito*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA. Ph.D. dissertation, 380pp.

Tanja Samardžić and Paola Merlo. 2010. Cross-lingual variation of light verb constructions: Using parallel corpora and automatic alignment for linguistic research. In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*, pages 52–60, Uppsala, Sweden. Association for Computational Linguistics.

Arden G. Sanders and Joy Sanders. 1994. Kamasau (wand tuan) grammar: Morpheme to discourse. Unpublished manuscript. Available at: `http://www.sil.org/pacific/png/abstract.asp?id=47683`.

Hansjakob Seiler. 1975. Die prinzipien der deskriptiven und etikettierenden benennung. In Hansjakob Seiler, editor, *I. I. I. Linguistic Workshop*, pages 2–57. Fink, München.

H. L. Shorto. 2013. *Wa-Praok Vocabulary*, volume 6 of *Asia-Pacific Linguistics*. Asia-Pacific Linguistics, Canberra.

James Neil Sneddon, Alexander Adelaar, Dwi Noverini Djenar, and Michael C. Ewing. 2010. *Indonesian Reference Grammar*, 2 edition. London: Allen Unwin, St Leonards.

Alan M. Stevens and A. Ed Schmidgall-Tellings. 2010. *A Comprehensive Indonesian–English Dictionary*, 2nd edition. Ohio University Press, Athens, Ohio.

Leonard Talmy. 1991. *Path to realization: A typology of event conflation*. Berkeley Linguistics Society, Berkeley, California. 2010.

Wesley Thiesen and David Weber. 2012. *A Grammar of Bora with Special Attention to Tone*. Number 148 in SIL International Publications in Linguistics. SIL International, Dallas, Texas.

Jörg Tiedemann. 2011. *Bitext alignment*, volume 4. Morgan & Claypool Publishers.

Stephen Ullmann. 1966. Semantic universals. In Joseph H. Greenberg, editor, *Universals of Language*, 2nd edition, pages 217–262. MIT Press, Cambridge, MA.

Annemarie Verkerk. 2013. Scramble, scurry and dash: The correlation between motion event encoding and manner verb lexicon size in indo-european. *Language Dynamics and Change*, 3(2):169 – 217.

Ljuba Veselinova. 2013. Negative existentials: A cross-linguistic study. *Rivista di Linguistica*, 25(1):107–145.

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, St. Julian's, Malta. Association for Computational Linguistics.

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.

Earle Waugh, Arok Wolvengrey, Loretta Pete, Ramona Washburn, and Intellinet Team. 2025. Online cree dictionary. Developed by Miyo Wahkohtowin Community Education Authority. Includes lesson plan builder and syllabic content contributions by Ramona Washburn. Accessed: 2025-04-18.

Wedhawati, Marsono, Edi Setiyanto, Dirgo Sabariyanto, Syamsul Arifin, Sumadi, Restu Sukesti, Herawati, Sri Nardiati, Laginem, and Wiwin Erni Siti Nurlina. 2001. *Tata Bahasa Jawa Mutakhir*. Pusat Bahasa Departemen Pendidikan Nasional, Jakarta.

Bernhard Wälchli. 2005. *Co-compounds and Natural Coordination*. Oxford Studies in Typology and Linguistic Theory. Oxford University Press, Oxford.

Bernhard Wälchli. 2007. *Lexical classes: A functional approach to "word formation"*, pages 153–178. De Gruyter Mouton, Berlin, New York.

Bernhard Wälchli and Anna Sjöberg. 2024. A law of meaning. *Linguistic Typology at the Crossroads*, 4(2):1–71.

Pavol Štekauer, Salvador Valera, and Lívia Kőrtvélyessy. 2012. *Word-formation in the world's languages: A typological survey*. Cambridge University Press.

A Data overview

Tables 4-7 present the 198 doculects, along with their affiliation and macro-area.

B Results from Step 1

Table 8 presents a selection form the output of Step 1 of the model as applied to the sample of Bible data. The first 20 extractions ($\langle v_s, P \rangle$ pairs) and every 30th extraction are printed, along with the number of doculects for which this pair was found to be significantly associated ($\mathbf{association}(v_s, P, t)$), their average p-value (negative-$\log_n$ transformed), the number of tokens in the corpus this applies to, and the proportion of all tokens of v_s this number makes up.

C Detailed validation

This section includes details of all the extracted cases of synlexification, circumlexification, and underspecification that we inspected manually using dictionaries and grammars. Table 9 shows each of the pairs of seed words that we looked at, and the strategy that the model predicts for each language, along with the most frequently extracted tokens for the pair and the glosses.

D Fuller table with extracted cases

Table 10 presents a larger set of examples of synlexifications of the different pairs of grammatical categories.

E Typological frequencies for two semantic domains

This section reports on the frequency with which doculects dominantly synlexify sets of comparison meaning pairs. Table 11 shows instances of (caused) motion events. These were based on all cases where one of the verbs *get*, *rise*, *sit*, *go*, *come*, *enter*, *put*, *throw*, *stand*, *depart*, *ascent*, *fall*, *cast*, or *pour* was combined with an adposition/particle from among *in, out, on, off, from, to, back, up, down*. Table 12 shows instances of speech events, based on all pairs where one of the four main speech verbs (Dirven et al., 1982) *say, tell, speak, talk* was combined with any other element. Note that the $N - 175$ instances of comparison meaning pairs including one of those verbs but not synlexified in any language were omitted from the table.

F Regression analysis of synlexification

This section provides further experimental detail on the logistic regression reported in Section 6. Table 13 presents the output of a mixed effects logistic regression (using the `glmer` library in `R`).

ISO 639-3	name	family	macroarea
AAUWBT	Abau	Sepik	Papunesia
ACMAS3	Gilit Mesopotamian Arabic	Afro-Asiatic	Eurasia
ACUTBL	Achuar-Shiwiar	Chicham	South America
AGGPNG	Angor	Senagi	Papunesia
AGMWBT	Angaataha	Angan	Papunesia
ALYXXX	Alyawarr	Pama-Nyungan	Australia
AMFSIM	Hamer-Banna	South Omotic	Africa
AMKWBT	Ambai	Austronesian	Papunesia
AMMWBT	Ama (Papua New Guinea)	Left May	Papunesia
AMNPNG	Amanab	Border	Papunesia
AMPWBT	Alamblak	Sepik	Papunesia
AMRTBL	Amarakaeri	Harakmbut	South America
AMUMVR	Guerrero Amuzgo	Otomanguean	North America
AOJFIL	Mufian	Nuclear Torricelli	Papunesia
ARLTBL	Arabela	Zaparoan	South America
AVAANT	Avar	Nakh-Daghestanian	Eurasia
AVTWBT	Au	Nuclear Torricelli	Papunesia
AZZTBL	Highland Puebla Nahuatl	Uto-Aztecan	North America
BBOBSM	Northern Bobo Madaré	Mande	Africa
BDHWBT	Baka (South Sudan)	Central Sudanic	Africa
BFDWBT	Bafut	Atlantic-Congo	Africa
BIBWBT	Bissa	Mande	Africa
BKLLAI	Berik	Tor-Orya	Papunesia
BOATBL	Bora	Boran	South America
BORWYI	Bororo	Bororoan	South America
BRUNXB	Eastern Bru	Austroasiatic	Eurasia
BSCWBT	Bassari-Tanda	Atlantic-Congo	Africa
BVRXXX	Burarra	Maningrida	Australia
BVZYSS	Bauzi	Geelvink Bay	Papunesia
BYRWBT	Baruya	Angan	Papunesia
BYXWBT	Qaqet	Baining	Papunesia
CABNVS	Garifuna	Arawakan	North America
CAPSBB	Chipaya	Uru-Chipaya	South America
CASNTM	Mosetén-Chimané	isolate	South America
CAXSBB	Lomeriano-Ignaciano Chiquitano	Chiquitano	South America
CBITBL	Cha'palaa	Barbacoan	South America
CBTTBL	Shawi	Cahuapanan	South America
CCOTBL	Comaltepec Chinantec	Otomanguean	North America
CHEIBT	Chechen	Nakh-Daghestanian	Eurasia
CHRPDV	Cherokee	Iroquoian	North America
CJPTJV	Cabécar	Chibchan	North America
CMEWBT	Cerma	Atlantic-Congo	Africa
CONWBT	Cofán	isolate	South America
CRHIBT	Crimean Tatar	Turkic	Eurasia
CRKWCV	Plains Cree	Algic	North America
CRNWBT	El Nayar Cora	Uto-Aztecan	North America
CRXWYI	Central Carrier	Athabaskan-Eyak-Tlingit	North America
CSKATB	Jola-Esulalu	Atlantic-Congo	Africa

Table 4: Overview of doculects used, along with their affiliation and macro-area (Table 1/4).

ISO 639-3	name	family	macroarea
CTGBSB	Chittagonian	Indo-European	Eurasia
DESWBT	Desano	Tucanoan	South America
DIDWBT	Didinga	Surmic	Africa
DIFXXX	Dieri	Pama-Nyungan	Australia
DJKWBT	Aukan	Indo-European	South America
DTSABM	Toro So Dogon	Dogon	Africa
DUDWYI	Hun-Saare	Atlantic-Congo	Africa
ELLELL	Modern Greek	Indo-European	Eurasia
ESEE06	Ese Ejja	Pano-Tacanan	South America
ESSWYI	Central Siberian Yupik	Eskimo-Aleut	Eurasia
EUSNLT	Basque	isolate	Eurasia
FRDWBT	Fordata	Austronesian	Papunesia
FUVLTBL	Hausa States Fulfulde	Atlantic-Congo	Africa
GAHPNG	Alekano	Nuclear Trans New Guinea	Papunesia
GBILAI	Galela	North Halmahera	Papunesia
GHSPNG	Guhu-Samane	Nuclear Trans New Guinea	Papunesia
GRTBBS	Garo	Sino-Tibetan	Eurasia
GUCTBL	Wayuu	Arawakan	South America
GUHWBT	Sikuani	Guahiboan	South America
GUKBSE	Northern Gumuz	Gumuz	Africa
GUPXXX	Bininj Kun-Wok	Gunwinyguan	Australia
HADLAI	Hatam	Hatam-Mansim	Papunesia
HAKTHV	Hakka Chinese	Sino-Tibetan	Eurasia
HTOWBT	Minica Huitoto	Huitotoan	South America
HUNK90	Hungarian	Uralic	Eurasia
HUVTBL	San Mateo del Mar Huave	Huavean	North America
HWCWYI	Hawai'i Creole English	Indo-European	Papunesia
IANPNG	Iatmul	Ndu	Papunesia
IBATIV	Iban	Austronesian	Papunesia
IFBTBL	Batad Ifugao	Austronesian	Papunesia
INDASV	Standard Indonesian	Austronesian	Papunesia
IRKBST	Iraqw	Afro-Asiatic	Africa
ITAR27	Italian	Indo-European	Eurasia
IZZTBL	Izi	Atlantic-Congo	Africa
JAMBSW	Jamaican Creole English	Indo-European	North America
JAVNRF	Javanese	Austronesian	Papunesia
JBUIBS	Jukun Takum	Atlantic-Congo	Africa
JICWBT	Tol	Jicaquean	North America
KABCEB	Kabyle	Afro-Asiatic	Africa
KBHWBT	Camsá	isolate	South America
KERABT	Kera	Afro-Asiatic	Africa
KFBNTA	Northwestern Kolami	Dravidian	Eurasia
KGRLAI	Abun	isolate	Papunesia
KHGNTV	Khams Tibetan	Sino-Tibetan	Eurasia
KHQBIV	Koyra Chiini Songhay	Songhay	Africa
KIAWBT	Kim	Atlantic-Congo	Africa
KMOWBT	Kwoma	Sepik	Papunesia
KMSPNG	Kamasau	Nuclear Torricelli	Papunesia
KNJSBI	Akateko	Mayan	North America
KORSYS	Korean	Koreanic	Eurasia

Table 5: Overview of doculects used, along with their affiliation and macro-area (Table 2/4).

ISO 639-3	name	family	macroarea
KPVIBT	Komi-Zyrian	Uralic	Eurasia
KPWPNG	Kobon	Nuclear Trans New Guinea	Papunesia
KRLNEW	Karelian	Uralic	Eurasia
KRSWYI	Kresh-Woro	Kresh-Aja	Africa
KTOWBT	Kuot	isolate	Papunesia
KYCPNG	Kyaka	Nuclear Trans New Guinea	Papunesia
LEFTBL	Lelemi	Atlantic-Congo	Africa
LMEABT	Peve	Afro-Asiatic	Africa
MAKLAI	Makasar	Austronesian	Papunesia
MBCWBT	Macushi	Cariban	South America
MCAWBT	Maca	Matacoan	South America
MDYBSE	Male (Ethiopia)	Ta-Ne-Omotic	Africa
MEJTBL	Meyah	East Bird's Head	Papunesia
MFEBSM	Morisyen	Indo-European	Africa
MFYWBT	Mayo	Uto-Aztecan	North America
MHIBSU	Ma'di	Central Sudanic	Africa
MHRIBT	Eastern Mari	Uralic	Eurasia
MIFWBT	Mofu-Gudur	Afro-Asiatic	Africa
MILTBL	Peñoles Mixtec	Otomanguean	North America
MIQSBN	Mískito	Misumalpan	North America
MLPTBL	Bargam	Nuclear Trans New Guinea	Papunesia
MOPWBT	Mopán Maya	Mayan	North America
MORBSS	Moro	Heibanic	Africa
MPMTBL	Yosondúa Mixtec	Otomanguean	North America
MPTWBT	Mian	Nuclear Trans New Guinea	Papunesia
MSYPNG	Aruamu	Ramu	Papunesia
MTOTBL	Totontepec Mixe	Mixe-Zoque	North America
MWWHDV	Hmong Daw	Hmong-Mien	Eurasia
MZMWBT	Mumuye	Atlantic-Congo	Africa
NABWBT	Southern Nambikuára	Nambiquaran	South America
NAFWBT	Nabak	Nuclear Trans New Guinea	Papunesia
NASPNG	Naasioi	South Bougainville	Papunesia
NHXNFB	Isthmus-Mecayapan Nahuatl	Uto-Aztecan	North America
NIAIBS	Nias	Austronesian	Papunesia
NIJLAI	Ngaju	Austronesian	Papunesia
NLDHSV	Dutch	Indo-European	Eurasia
NOAWBT	Woun Meu	Chocoan	South America
NTJXXX	Ngaanyatjarra	Pama-Nyungan	Australia
NTPTBL	Northern Tepehuan	Uto-Aztecan	North America
NUYXXX	Wubuy	Gunwinyguan	Australia
OPMTBL	Oksapmin	Nuclear Trans New Guinea	Papunesia
OTQTBL	Querétaro Otomi	Otomanguean	North America
PADWBT	Paumari	Arawan	South America
PAMPBS	Pampanga	Austronesian	Papunesia
PAONAB	Northern Paiute	Uto-Aztecan	North America
PAUPAL	Palauan	Austronesian	Papunesia
PBBDYU	Páez	isolate	South America
PJTXXX	Pitjantjatjara	Pama-Nyungan	Australia
POEWBT	San Juan Atzingo Popoloca	Otomanguean	North America
POIWBT	Highland Popoluca	Mixe-Zoque	North America

Table 6: Overview of doculects used, along with their affiliation and macro-area (Table 3/4).

ISO 639-3	name	family	macroarea
PPOWBT	Folopa	Teberan	Papunesia
PRKBSM	South Wa	Austroasiatic	Eurasia
PUIABC	Puinave	isolate	South America
QUBPBS	Huallaga Huánuco Quechua	Quechuan	South America
RAWBIB	Rawang	Sino-Tibetan	Eurasia
RELBTL	Rendille	Afro-Asiatic	Africa
ROOWBT	Rotokas	North Bougainville	Papunesia
SABWBT	Buglere	Chibchan	North America
SGWBSE	Sebat Bet Gurage	Afro-Asiatic	Africa
SHKBSS	Shilluk	Nilotic	Africa
SPPTBL	Supyire Senoufo	Atlantic-Congo	Africa
SRNBSS	Sranan Tongo	Indo-European	South America
SSDWBT	Siroi	Nuclear Trans New Guinea	Papunesia
SURIBS	Mwaghavul	Afro-Asiatic	Africa
SXNLAI	Sangir	Austronesian	Papunesia
TABIBT	Tabasaran	Nakh-Daghestanian	Eurasia
TACPBC	Western Tarahumara	Uto-Aztecan	North America
TBGWBT	North Tairora	Nuclear Trans New Guinea	Papunesia
TCATBL	Ticuna	Ticuna-Yuri	South America
TCCBST	Barabayiiga-Gisamjanga	Nilotic	Africa
TCSWYI	Torres Strait-Lockhart River Creole	Indo-European	Australia
TEETBL	Huehuetla Tepehua	Totonacan	North America
TEOBSU	Teso	Nilotic	Africa
TFRWBT	Teribe	Chibchan	North America
THATSV	Thai	Tai-Kadai	Eurasia
TIHBSM	Timugon Murut	Austronesian	Papunesia
TIKWYI	Tikar	Atlantic-Congo	Africa
TLJWBT	Talinga-Bwisi	Atlantic-Congo	Africa
TOPTBL	Papantla Totonac	Totonacan	North America
TPIPNG	Tok Pisin	Indo-European	Papunesia
TPTTBL	Tlachichilco Tepehua	Totonacan	North America
TQOTQO	Toaripi	Eleman	Papunesia
TRCWBT	Copala Triqui	Otomanguean	North America
TUFWYI	Central Tunebo	Chibchan	South America
URATBL	Urarina	isolate	South America
URBWBT	Urubú-Kaapor	Tupian	South America
VIELHG	Vietnamese	Austroasiatic	Eurasia
WBABIV	Warao	isolate	South America
WIMWYI	Wik-Mungkan	Pama-Nyungan	Australia
XALIBT	Oirad-Kalmyk-Darkhat	Mongolic-Khitan	Eurasia
XAVTBL	Xavánte	Nuclear-Macro-Je	South America
XSUMEV	Sanumá	Yanomamic	South America
YADTBL	Yagua	Peba-Yagua	South America
YLEWBT	Yele	isolate	Papunesia
YSSYYV	Yessan-Mayo	Sepik	Papunesia
YUJWBT	Karkar-Yuri	Pauwasi	Papunesia
YUZNTM	Yuracaré	isolate	South America
YVATBL	Yawa	Yawa-Saweru	Papunesia
ZNEZNE	Zande	Atlantic-Congo	Africa
ZPMTBL	Mixtepec Zapotec	Otomanguean	North America

Table 7: Overview of doculects used, along with their affiliation and macro-area (Table 4/4).

rank	v_s	P	N doculects	avg. -log p	N tokens	token coverage
1	write	say+write	110	122.27	196	0.92
2	answer	answer+say	100	inf	240	0.97
3	heal	heal+sick	100	39.65	65	0.81
4	scribe	law+scribe	91	282.37	66	1.00
5	forgive	sin+forgive	91	56.45	64	0.98
6	repent	sin+repent	83	88.26	57	1.00
7	vinegar	wine+vinegar	81	30.01	6	1.00
8	widow	widow+woman	80	56.93	27	0.96
9	raise	dead+raise	78	53.38	79	0.84
10	faith	believe+faith	77	inf	279	0.91
11	come	to+come	75	64.33	802	0.66
12	loaf	bread+loaf	75	42.05	27	1.00
13	prostitute	prostitute+woman	67	36.82	13	1.00
14	bread	eat+bread	65	39.65	73	0.85
15	cup	wine+cup	65	31.98	20	0.61
16	drink	wine+drink	63	31.06	61	0.61
17	prophet	write+prophet	60	47.11	125	0.70
18	read	read+write	60	36.69	28	0.88
19	knock	knock+door	60	33.37	9	1.00
20	smoke	smoke+fire	60	29.15	13	1.00
30	life	life+eternal	54	56.87	142	0.68
60	branch	tree+branch	40	29.66	18	0.90
90	silver	money+silver	32	25.06	17	0.81
120	language	language+word	27	49.97	40	1.00
150	endure	suffer+endure	23	28.48	30	0.73
180	barrack	house+soldier	20	31.32	6	1.00
210	milk	child+milk	18	21.75	4	0.80
240	n't	n't+not	16	19.94	54	0.22
270	naked	garment+naked	14	23.81	14	0.78
300	key	key+door	13	23.05	6	1.00
330	tithe	priest+tithe	12	18.70	6	0.67
360	roll	tomb+roll	11	17.41	6	0.46
390	tax	money+tax	9	40.09	22	0.81
420	hypocrite	hypocrite+good	8	56.52	20	0.67
450	wave	wave+water	8	20.06	6	0.67
480	gentle	gentle+peace	7	29.58	14	0.78
510	hospitality	receive+house	7	20.28	3	1.00
540	doctrine	teach+true	6	24.29	9	0.56
570	reconcile	peace+with	5	40.68	10	0.62
600	muzzle	bind+mouth	5	20.65	2	1.00
630	divide	divide+self	4	36.37	14	0.41
660	star	star+heaven	4	23.53	11	0.38
690	ring	finger+ring	4	19.53	2	0.67
720	married	marry+married	4	16.43	4	0.40
750	summer	new+summer	3	26.52	3	1.00
780	spring	spring+water	3	21.59	6	0.12
810	laugh	ridicule+laugh	3	20.05	2	1.00
840	slaughter	bring+kill	3	18.35	3	0.60
870	resist	resist+write	3	16.74	3	0.33

Table 8: Select output of Step 1 (top-20 extractions and every 30th extraction after)

Pair	Language	Predicted Strategy	Verdict	Aligned tokens	Gloss	Source
wife+woman	Basque	synlexified	correct	emazte+emazte	*emazte* (’wife’)	de Rijk (2008, p. 961)
wife+woman	Bora	unlexified	unlexified	méwakye+méwakye	*méwá* (’wife’)	Thiesen and Weber (2012, p. 25)
wife+woman	Mískito	circumlexified	incorrect	maî+maî	*maia* (’spouse’)	Salamanca (1988, p. 228)
dead+die	Arabic	synlexified	correct	مَت+نَمُت	نَمُت (’we die’)	Abu-Chacra (2007, p. 49)
dead+die	Chechen	synlexified	correct	вела+вела	*vella* (’died’)	Awde and Galaev (2014, p. 57)
dead+die	Indonesian	circumlexified	incorrect	mat+i	*mati* (’die’)	Sneddon et al. (2010, p. 75)
king+throne	Indonesian	synlexified	correct	takhta+takhta	*takhta* (’throne’)	Stevens and Schmidgall-Tellings (2010, p. 987)
king+throne	Kamasau	circumlexified	correct	king+sia	*sia* (’chair’)	Sanders and Sanders (1994, p. 61)
go+way	Abau	underspecified	correct	ley+	*ley* (’go’)	Lock (2011, p. 25)
go+way	Bora	underspecified	correct	péé+	*pééhi* (’go’)	Thiesen and Weber (2012, p. 50)
go+way	Italian	underspecified	correct	va+	*andare* (’go’)	Maiden and Robustelli (2013, p. 222)
go+way	Tok Pisin	circumlexified	correct	go+i	*go*(’go’)	Dutton and Thomas (1994, p. 364)
go+out	Basque	synlexified	correct	ilki+ilki	*ilki* (’go out’)	Aulestia (1989, p. 302)
go+out	Zande	underspecified	correct	ndu+	*ndu* (‘go, walk’)	Pasch (2007, 172,173)
go+out	Indonesian	underspecified	correct	pergi+	*pergi* (’go’)	Sneddon et al. (2010, p. 165)
take+way	Ndyuka	underspecified	correct	teke+	*teke* (’take’)	Huttar and Huttar (1994, p. 10)
take+way	Abun	underspecified	correct	nai+	*nai* (’took’)	Berry and Berry (1999, p. 67, 56)
take+way	Kyaka Enga	unlexified	unlexified	la+	*nyii* (’take’)	Draper and Draper (2002)
take+way	Jamaican English	circumlexified	correct	tek+we	*tek* (’take’), *we* (’away’)	Bailey (1968, p. 378)
door+open	Kwoma	circumlexified	correct	nubureja+tagwa	*tagwa* (’open’), *nubereja* (’door’)	Bowden (1997, p. 208, 157)
door+open	Kamsá	synlexified	uncertain	atëfjna+atëfjna	*bësasa* (‘door’)	O’Brien (2018, p.178)
door+open	Cora	unlexified	unlexified	antácuunyaraca+antácuunyaraca	*cuuna* (’open a door’)	Casad (2012, p. 91)
door+open	Iban	circumlexified	correct	pintu+muka	*pintu* (’door’), *muka* (’open’)	Omar (1969, p. 16, 228)
clean+un	Ndyuka	circumlexified	incorrect	takuu+takuu	*takuu* (’evil’)	Huttar and Huttar (1994, p. 62)
clean+un	Javanese	synlexified	correct	jahat+jahat	*jahat* (’evil’)	Wedhawati et al. (2001, p. 155)
clean+un	Cree	circumlexified	correct	kisîpêkisitêw+ekâ	*kisîpêkisitêw* a(wash), *ekâ*,(’un-’)	Waugh et al. (2025)
whole+world	Bauza	underspecified	correct	+bak	*bak* (‘ground’)	Briley (1997, p. 6)
whole+world	Yorno So	underspecified	uncertain	puu+	*puu* (’all’)	Heath (2014, p. 280)
whole+world	Indonesian	underspecified	incorrect	seluruh+	*seluruh* (’whole’), *dunia* (’world’)	Sneddon et al. (2010, p. 41, 56)
to+world	Bora	underspecified	incorrect	vú+	*vu* (goal marker)	Thiesen and Weber (2012, p. 156)
to+world	Ndyuka	underspecified	incorrect	+goontapu	*goontapu* (’world’)	Huttar and Huttar (1994, p. 62)
to+world	Basque	underspecified	incorrect	ra+mundu	*-ra* (allative marker)	de Rijk (2008, p. 50)
from+go	Koyra Chiini Songhay	underspecified	incorrect	koy+	*hun* (‘leave’)	Heath (1999, p. 80)
from+go	Italian	circumlexified	correct	di+partì	*partì* (’depart’), *di* (’from’)	Maiden and Robustelli (2013, p. 366)
from+go	Komi	circumlexified	correct	ысь+мун	*muniy* (’depart’), *-is* (ablative affix)	Avril (2006, p. 90, 242)
go+up	Rotokas	synlexified	correct	ipa+ipa	*ipa* (‘ascended’)	Firchow (1974, p. 40)
go+up	Komi	synlexified	correct	кыпӧдчы+кыпӧдчы	*kaniy* (’ascend’)	Avril (2006, p. 216, 90)
go+up	Timugon Murut	underspecified	correct	minongoi+	*ongoi* (’go’)	Prentice (1971, p. 47)
enter+in	Fulfulde	synlexified	correct	natt+natt	*nattay* (‘enter’)	McIntosh (1984, p. 125)
enter+in	Jamaican English	circumlexified	correct	go+iin	*go* (’go’), *iin* (in)	Bailey (1968, p. 227)
enter+in	Karkar	underspecified	correct	+mek	*mik* (‘in’)	Rigden (n.d., p. 112)
answer+say	Tok Pisin	circumlexified	correct	bek+tok	*tok* (’say’), *bek* (’back’)	Dutton and Thomas (1994, p. 5)
answer+say	Iban	synlexified	correct	nyaut+nyaut	*naut* (’answer’)	Omar (1969, p. 228)
fish+net	Chipaya	synlexified	uncertain	anś+chíś	*ch’iz* (’fish’)	Cerrón-Palomino (2006, p. 194)
fish+net	Kyaka Enga	circumlexified	correct	oma+nyuu	*oma* (’fish’), *nyuu* (’bag’)	Draper and Draper (2002, p. 229, 291)
blind+eye	South Wa	circumlexified	correct	dug+ngai	*ngai* (’eye’), *duk* (’blind’)	Shorto (2013, p. 19)
blind+eye	Garifuna	synlexified	correct	marihin+marihin	*maríhin* (’not see’)	Haurholm-Larsen (2016, p. 201)

Table 9: Aligned tokens to seed pairs that were manually inspected and given a verdict with dictionaries and grammars

PoS pair	least and most often modally synlexified (N languages per pair)
Adposition+Noun (N=461; 18%)	**bottom** = about+thing (1) accord+with (1) voice+with (1) before+man (1) book+of (1) of+rich (1) of+star (1) of+sign (1) country+of (1) demon+of (1) **top** = mountain+on (116) before+foot (80) in+peace (72) disciple+of (46) of+son (41) demon+in (27) of+woe (26) gold+of (18) in+world (17) city+in (14)
Adposition+Verb (N=460; 19%)	**bottom** = before+fall (1) before+go (1) before+set (1) beg+to (1) owe+to (1) over+throw (1) belong+to (1) believe+in (1) bring+up (1) bind+with (1) **top** = rise+up (107) get+up (104) down+fall (95) cry+out (85) before+defile (84) down+sit (81) stand+up (67) out+release (60) at+marvel (49) cut+off (42)
Noun+Verb (N=408; 37%)	**bottom** = understand+word (1) bear+tree (1) sit+throne (1) language+speak (1) say+woman (1) fear+speak (1) man+name (1) man+right (1) enter+place (1) eye+open (1) **top** = bread+eat (173) law+write (164) apostle+send (163) eat+food (161) steal+thief (154) prophet+write (153) glory+worship (151) joy+rejoice (150) bondservant+serve (150) fish+take (146)
Verb+Verb (N=245; 20%)	**bottom** = become+know (1) beg+say (1) bear+give (1) come+touch (1) cry+say (1) hear+let (1) lead+stray (1) command+say (3) go+set (3) look+see (3) **top** = deceive+lie (162) suffer+torment (146) persecute+suffer (120) know+understand (109) hear+marvel (108) greet+kiss (106) eat+reap (95) come+send (95) know+see (93) find+see (92)
Noun+Noun (N=198; 80%)	**bottom** = beast+thing (1) sin+thing (1) house+master (1) fruit+wine (1) gift+sacrifice (1) thing+work (1) man+woman (1) news+word (1) brother+mother (2) bread+piece (3) **top** = boat+ship (186) boat+sea (186) horse+soldier (184) money+stone (182) demon+devil (181) month+moon (181) bird+dove (176) guard+soldier (175) fire+light (174) cloak+garment (172)
Adjective+Noun (N=87; 67%)	**bottom** = day+first (1) day+many (1) new+wine (1) sharp+sword (3) blind+man (4) such+thing (4) body+whole (4) certain+man (6) great+multitude (7) many+people (8) **top** = blind+eye (179) blood+dead (178) famine+hungry (172) afraid+fear (167) dead+tomb (166) angry+wrath (164) eternal+life (164) money+poor (163) parable+word (152) garment+naked (149)
Affix+Verb (N=86; 89%)	**bottom** = ation+save (1) believe+ful (1) ent+excel (2) ent+hear (2) ant+know (2) ful+write (3) ion+suffer (3) ion+relate (4) ance+repent (4) dom+know (4) **top** = er+teach (142) ant+serve (105) er+pray (105) re+turn (105) beware+self (102) appoint+dis (100) care+ful (90) be-ed+love (90) ion+oppress (87) er+sin (82)
Proper Noun+Verb (N=84; 9%)	**bottom** = Christ+die (1) God+worship (1) Isaiah+say (1) Jesus+answer (1) Passover+eat (1) Paul+say (1) Peter+say (1) Peter+answer (2) **top** = Peter+answer (2) Jesus+answer (1) Christ+die (1) Isaiah+say (1) God+worship (1) Passover+eat (1) Paul+say (1) Peter+say (1)
AFX+Noun (N=82; 86%)	bottom = ual+woman (1) body+ion (1) ion+sin (1) ence+word (1) ent+word (1) ly+word (1) ness+thing (1) ness+sin (1) s+side (1) flesh+ly (2) top = et+trump (178) enemy+st (166) a-ed+shame (144) st+war (136) com+passion (111) author+ity (93) cy+prophet (88) out+side (83) age+bond (80) fool+ish (74)
Particle+Verb (N=79; 20%)	**bottom** = come+to (1) crow+not (1) destroy+to (1) enter+to (1) hear+not (1) heal+to (1) release+to (1) love+not (1) stumble+to (1) stand+to (1) **top** = to+want (25) lose+not (8) circumcise+not (4) teach+to (2) not+want (2) crow+not (1) hear+not (1) release+to (1) come+to (1) stumble+to (1)

Table 10: Examples of most and least synlexified granularized seed doculect word pairs per part of speech (PoS) pair. Numbers in parentheses in the first column represent the total number of pairs and the proportion of pairs for which at least one doculect dominantly synlexifies that pair ('% synlex'). PoS abbreviations are [n]oun, [a]djective, [v]erb, [p]reposition, affi[x], proper na[m]e, par[t]icle

pair	N doc.	frequency pair
rise + up	107	29
get + up	104	19
down + fall	95	83
down + sit	81	42
stand + up	67	21
cast + out	33	39
out + pour	29	21
depart + from	29	11
come + down	9	69
go + up	8	61
down + throw	4	15
down + go	3	29
enter + in	3	191
out + throw	3	22
get + in	3	10
on + stand	3	17
go + out	2	117
from + rise	2	31
on + sit	2	66
come + out	1	132
come + from	0	107
fall + on	0	37
fall + in	0	29
fall + from	0	12
enter + to	0	175
come + to	0	1173
come + on	0	88
come + up	0	31
depart + to	0	11
come + in	0	312
cast + in	0	52
cast + to	0	38
go + in	0	211
go + on	0	68
get + to	0	20
fall + to	0	42
from + go	0	109
in + stand	0	51
in + sit	0	75
in + rise	0	12
in + put	0	52
in + pour	0	11
go + to	0	657
on + put	0	40
on + pour	0	13
out + put	0	10
in + throw	0	31
put + to	0	59
pour + to	0	10
rise + to	0	25
sit + to	0	29
stand + to	0	17
throw + to	0	70

Table 11: (Caused) motion events, their cross-doculectal frequency of being dominantly synlexified (N doc.), and their corpus frequency.

pair	N doc.	frequency pair
promise + say	63	57
false + say	54	23
answer + say	42	279
say + thunder	27	11
ask + say	27	185
confess + say	13	16
say + speak	12	329
sin + speak	12	32
ar + say	9	10
say + write	6	266
lie + say	4	30
command + say	3	135
speak + still	2	10
say + word	2	452
prophet + speak	1	13
language + speak	1	20
fear + speak	1	10
among + say	1	10
Isaiah + say	1	11
say + to	1	1675
say + woman	1	16
Peter + say	1	48
cry + say	1	75
say + still	1	10
beg + say	1	14
Paul + say	1	23

Table 12: Speech events, their cross-doculectal frequency of being dominantly synlexified (N doc.), and their corpus frequency.

	Estimate	Std.	Error	z value	Pr(>|z|)
(Intercept)	-14.57443	0.55514	-26.254	< 2e-16	***
log.pair.freq	1.20256	0.14200	8.469	< 2e-16	***
pos.type=ADP+VERB	0.55582	0.34499	1.611	0.107154	
pos.type=NOUN+NOUN	9.23598	0.41556	22.225	< 2e-16	***
pos.type=NOUN+VERB	3.37687	0.36528	9.245	< 2e-16	***
pos.type=other	4.26841	0.31352	13.615	< 2e-16	***
pos.type=VERB+VERB	1.34335	0.41376	3.247	0.001168	**
macroarea=Australia	-1.06364	0.21781	-4.883	1.04e-06	***
macroarea=Eurasia	-0.02568	0.14583	-0.176	0.860197	
macroarea=North America	-0.51993	0.13806	-3.766	0.000166	***
macroarea=Papunesia	-0.31067	0.12234	-2.539	0.011101	*
macroarea=South America	-0.68934	0.13487	-5.111	3.20e-07	***
AIC	118693.8				
residual degrees of freedom	506471				

Table 13: Detailed results of the logistic regression predicting synlexification.