Token-level semantic typology without a massively parallel corpus

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

Abstract

This paper presents a computational method for token-level lexical semantic comparative research in an original text setting, as opposed to the more common massively parallel setting. Given a set of (non-massively parallel) bitexts, the method consists of leveraging pre-trained contextual vectors in a reference language to induce, for a token in one target language, the lexical items that all other target languages would have used, thus simulating a massively parallel set-up. The method is evaluated on its extraction and induction quality, and the use of the method for lexical semantic typological research is demonstrated.

1 Introduction

Lexical semantic typology has benefited immensely from the availability of massively parallel corpora. Having the same message translated from a reference language into many different target languages affords linguists a basis to study variation in word meanings (cf. Haspelmath, 2018) at the fine-grained level of corpus tokens (Levshina, 2016).

The massively parallel set-up (Figure 1, top) allows us to determine, for instance, that Spanish and German both split the meaning of English *know* similarly into 'know someone' (*conozco, kenne*), and 'know something' (*sabe, weiss*). Studies using massively parallel corpora have challenged prior conceptions of semantic variation, showing that languages vary continuously rather than discretely in where they mark lexical boundaries (e.g., Verkerk, 2014 for motion events) and revealing novel factors explaining such lexical boundaries (e.g.,Wälchli, 2016 for verbs of visual perception).

Massively parallel corpora do, however, have methodological downsides (see Levshina, 2021 for a review). They tend to reflect literary genres and have conceptual content that may be foreign to the culture into whose language the text is translated (Domingues et al., 2024; Pinhanez et al., 2023).

Figure 1: Massively parallel corpora vs. original corpora

Moreover, languages differ in how they habitually formulate, i.e., what conceptual contents speakers typically bring up when they engage in 'the same' linguistic activities, such as telling a story (Tannen, 1980) or making a request (Terkourafi, 2011). Finally, translated text displays transfer effects, where properties of a source language are transferred to a target language, thus making the target language look more like the source language (Johansson and Hofland, 1994, though see Levshina, 2017).

These issues could in part be circumvented by using original text corpora with translations *from* the (untranslated) target languages *into* a shared reference language, as in Figure 1, bottom panel. In comparative linguistics, such corpora are commonly used (McEnery and Xiao, 2007; Enghels et al., 2020). However, under this set-up, the translations are not massively parallel. The loss of massive parallelism impacts the comparability: without it, we can, for instance, no longer directly infer that Spanish *conozco* covers (approximately) the same meanings as German *kenne*, and ditto for *sabe* and *weiss*. Moreover, it affects the downstream analytic techniques we can use: many studies rely on dimensionality reduction over the translations of the seed language tokens into *all* target languages, a situation unavailable with an original text set-up.

The use of original text data thus calls for a

Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2025), pages 165–176

method to make these data comparable at the token level. Multilingual contextualized representation spaces, such as Multilingual BERT (Devlin et al., 2019) could be considered here, but given the small amounts of data available for most languages, as well as challenges to the 'true' multilingual nature of Multilingual BERT (Pires et al., 2019), this approach does not seem feasible.

Instead, this paper proposes a method that leverages pre-trained contextual vectors for the reference language to induce, for a token in one target language, the lexical items that all other target languages *would have* used. Doing so, the proposed method simulates a massively parallel set-up. Pre-trained vectors for one language in a translation pair have been succesfully used to improve word alignment quality in bitexts (Dou and Neubig, 2021), and as such we can expect further translation-oriented applications to similarly benefit from the more substantial training data available for resource-rich languages like English.

After describing the method (§2) and introducing the corpora (§3), I will report on three experiments validating the method (§4), and showcase the use of the method for lexical semantic typological research (§5). Code is available at `https://github.com/dnrb/no-parallel-corpus`.

2 A method for inferring lexification

Here, I propose a method to simulate a massively parallel setting when such parallelism is not available. The method takes as its input raw bitexts with translations from a target language t into a reference language r for which contextual vectors are available or can feasibly be trained. The method uses these contextual vectors to induce a classification model predicting lexical choice in t. This model can then be applied to translations of *another* target language t' into r, to infer the lexical choice that t would have made if asked to translate a word token in t'. Doing so for all languages lets us to infer a token-by-language table like those used in studies based on massively parallel corpora.

2.1 Alignment step

To induce a lexical classification model in a target language t on the basis of contextualized vectors in r, we need to know, for a particular token w_t of t, which token w_r of r is translation equivalent, so that an association between w_t and the contextualized vector of w_r can be learned. At the same time, lexical semantic typology tends to be interested in the lexical choice of lemmas (e.g., *believe*) rather than the inflected forms (e.g., *believe, believes, believing, believed*), and as such, the alignment procedure would ideally also identify the shared lemma form in the target languages.

An approach affording both at the same time is Liu et al. (2023)'s Conceptualizer model. Assuming a bitext U, consisting of paired utterances $\langle u_r, u_t \rangle$ in the reference and target language, we define $U_v \subseteq U$ as the set of bitext utterances in which reference language word type v occurs. Given a reference language seed word v, the procedure then considers each possible substring l in t, and retrieves the set of bitext utterances $U_l \in U$ in which l occurs. The most strongly associated substring l_{max} is the substring whose Fisher Exact score over the following 2×2 table has the lowest p-value:

$$\begin{array}{ll} |U_v \cap U_l| & |U_v / U_l| \\ |U_l / U_v| & |U / (U_v \cup U_l)| \end{array}$$

Intuitively, l_{max} is a substring of target language words that frequently occur in the same utterances as v and infrequently occur in utterances where v is not present. The search space over all possible l is further reduced by assuming that $\frac{|U_v \cap U_l|}{|U_v|} \geq \theta_t$ and that $\frac{|U_v \cap U_l|}{|U_l|} \geq \theta_b$, with $\theta_t = 0.01$ and $\theta_b = 0.10$, i.e. that the union of utterances containing v and l should make up 1% or more or all utterances containing v and that the same union should make up 10% or more of all utterances containing l.

When l_{max} is found, $U_{l_{\text{max}}}$ is removed from U_v, and the process is repeated on the updated set U_v, until a pre-set threshold of coverage over the tokens of v is reached (here: $0.95 \times |U_v|$).

In subsequent steps, the model will need to retrieve the word tokens associated with l_{max}. It does so through the function **tokens**(l_{max}), which goes through all $u \in U_v \cap U_{l_{\text{max}}}$ and retrieves, per u, the target language word token that contains l_{max}. If multiple tokens in some u_t contain l_{max}, the one that occurs in the largest number of utterances in $U_v \cap U_{l_{\text{max}}}$ is selected.

2.2 Lemma merger step

Exploration reveals that the Liu et al. (2023) procedure often extracts spurious unique lemmas for a seed word. For instance, both *ˆsepara* and *ˆsepare* (carets denote the start of a string) might be extracted in Spanish as target language lemmas given the seed word *separate*. These are obvious variants of the same lemma (*separar*). Similarly, identical

language (glottocode, family, area: reference)	n tokens	language (glottocode, family, area: reference)	n tokens
Anal (anal1239, Sino-Tibetan, Eurasia: Ozerov)	14026	Nllng (nngg1234, Tuu, Africa: Güldemann et al., 2024)	27035
Yali (Apahapsili) (apah1238, Nuclear Trans New Guinea, Papunesia: Riesberg, 2024)	15243	Northern Kurdish (Kurmanji) (nort2641, Indo-European, Eurasia: Haig et al., 2024)	9657
Arapaho (arap1274, Algic, North America: Cowell, 2024)	10279	Northern Alta (nort2875, Austronesian, Papunesia: Garcia-Laguia, 2024)	11137
Baïnounk Gubëeher (bain1259, Atlantic-Congo, Africa: Cobbinah, 2024)	12522	Fanbyak (orko1234, Austronesian, Papunesia: Franjieh, 2024)	18928
Beja (beja1238, Afro-Asiatic, Africa: Vanhove, 2024)	15454	Pnar (pnar1238, Austroasiatic, Eurasia: Ring, 2024)	20485
Cabécar (cabe1245, Chibchan, North America: Quesada et al., 2024)	17528	Daakie (port1286, Austronesian, Papunesia: Krifka, 2024)	11880
Cashinahua (cash1254, Pano-Tacanan, South America: Reiter, 2024)	9655	Ruuli (ruul1235, Atlantic-Congo, Africa: Witzlack-Makarevich et al., 2024)	8255
Dolgan (dolg1241, Turkic, Eurasia: Däbritz et al., 2024)	18694	Sadu (sadu1234, Sino-Tibetan, Eurasia: Xu and Bai, 2024)	11752
Evenki (even1259, Tungusic, Eurasia: Kazakevich and Klyachko, 2024)	8366	Sanzhi Dargwa (sanz1248, Nakh-Daghestanian, Eurasia: Forker and Schiborr, 2024)	5140
Goemai (goem1240, Afro-Asiatic, Africa: Hellwig, 2024)	24039	Savosavo (savo1255, Isolate, Papunesia: Wegener, 2024)	11383
Gorwaa (goro1270, Afro-Asiatic, Africa: Harvey, 2024)	19988	Nafsan (South Efate) (sout2856, Austronesian, Papunesia: Thieberger, 2024)	25204
Gurindji (guri1247, Pama-Nyungan, Australia: Meakins, 2024)	6116	Sümi (sumi1235, Sino-Tibetan, Eurasia: Teo, 2024)	11158
Hoocąk (hoch1243, Siouan, North America: Hartmann, 2024)	7431	Svan (svan1243, Kartvelian, Eurasia: Gippert, 2024)	10318
Jahai (jeha1242, Austroasiatic, Eurasia: Burenhult, 2024)	8087	Tabasaran (taba1259, Nakh-Daghestanian, Eurasia: Bogomolova et al., 2024)	5057
Jejuan (jeju1234, Koreanic, Eurasia: Kim, 2024)	9359	Teop (teop1238, Austronesian, Papunesia: Mosel, 2024)	12134
Kakabe (kaka1265, Mande, Africa: Vydrina, 2024)	46634	Texistepec Popoluca (texi1237, Mixe-Zoque, North America: Wichmann, 2024)	8468
Kamas (kama1351, Uralic, Eurasia: Gusev et al., 2024)	37861	Totoli (toto1304, Austronesian, Papunesia: Bardají i Farré, 2024)	11798
Tabaq (Karko) (kark1256, Nubian, Africa: Hellwig et al., 2024)	9318	Mojeño Trinitario (trin1278, Arawakan, South America: Rose, 2024)	17421
Komnzo (komn1238, Yam, Papunesia: Döhler, 2024)	33773	Asimjeeg Datooga (tsim1256, Nilotic, Africa: Griscom, 2024)	8782
Light Warlpiri (ligh1234, Mixed Language, Australia: O'Shannessy, 2024a)	8685	Urum (urum1249, Turkic, Eurasia: Skopeteas et al., 2024)	18797
Movima (movi1243, Isolate, South America: Haude, 2024)	10243	Vera'a (vera1241, Austronesian, Papunesia: Schnell, 2024)	17785
Dalabon (ngal1292, Gunwinyguan, Australia: Ponsonnet, 2024)	4046	Warlpiri (warl1254, Pama-Nyungan, Australia: O'Shannessy, 2024b)	7129

Table 1: The 44 languages in the DoReCo dataset. 'n tokens' is the number of target language word tokens.

target language lemmas may mismatch across seed words: English *split* might yield Spanish *ˆsepar* as a target language lemma. Without further processing, this would lead to the model's failure to recognize that *separar* translates into the reference language words *separate* and *split*.

To resolve this issue, I implement a simple heuristic to merge target language lemmas given the same or different seed words. In all cases, the basic criterion is that two target language lemmas l_i and l_j are merged iff they have a longest-common substring (1) whose length is ≥ 3 characters, and (2) that is at least half as long (in characters) as the shortest string of the two lemmas l_i and l_j. When merging *across* seed words (like *ˆsepar* given *split* and *ˆsepara* given *separate* in the example above), we further require that the whole word forms (e.g., *separamos, separaba*) that the two lemmas cover overlap, as a further way to ensure that they indeed are the same lemma. Concretely, we retrieve the set of unique whole word forms covered by l_i, i.e. all unique strings from **tokens**(l_i), and call it W_i. We do the same for l_j and call it W_j. Next, we define the two lemmas to have sufficient overlap in the word forms they cover if $|W_i \cap W_j| \geq \max(|W_i|, |W_j|) \times 0.5$, or: the intersection of their word forms is at least half the size of the largest of the two sets. All lemma pairs are considered, and an undirected graph is induced with edges between all pairs of mergeable lemmas, after which all lemmas in each connected component are merged.

2.3 Induction step

With the inferred mapping between seed words in the reference language t and merged lemmas in the target language t, we can now train a classifier to induce the merged lemma given a seed word token in the bitext between r and t. In particular, the classifier learns a mapping between contextualized vector representations $\overrightarrow{w_r}$ of each token w_r, and the merged lemmas L_t, as obtained through the previous steps.

This, then, allows for the inference of what a target language t would have used in the case of a token of some other target language t'. For every token $w_r \in B^r_{t'}$, that is: in the bitext between r and t', the contextualized vector $\overrightarrow{w_r}$ is retrieved, and **classify**$(\overrightarrow{w_r}, t)$ predicts the lemma in t for the translation of a token in t'. As such, we now know that t uses **classify**$(\overrightarrow{w_r}, t)$ for w_r, and t' uses **classify**$(\overrightarrow{w_r}, t')$ for the same token, thus making the reference language token a comparable category. Doing so for all $t \in T$ yields one row in a comparison table as obtained from a massively parallel corpus, except that most lexical labels are now inferred instead of observed. Doing so for all word tokens w_r in any bitext allows us to create the full table. I will explore the insights that can be derived from such a table in §5, but first validate the quality of this procedure.

3 Experimental set-up and materials

This paper uses the DoReCo corpus (Seifart et al., 2024), a collection of data gathered by documentary linguists for a typologically diverse sample of languages. The individual language resources form free-standing contributions that should be individually cited as part of the usage agreement. Table 1 presents the 44 languages used, along with metadata about affiliation and location and the number of (translated) words in each language.

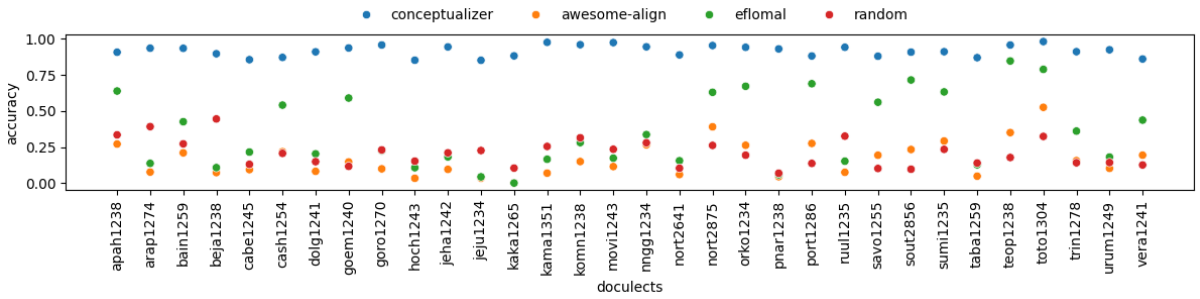


Figure 2: Mean extraction accuracy (blue) vs. random baseline (orange).

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

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

The structure of the components of the DoReCo corpora is given in Table 2, for the language Savosavo. All languages have the **[w]**ord and **[f]**ree translation layer, and a select subset of languages is interlinearly glossed with a **[m]**orphological segmentation layer and a **[g]**loss layer. Subsequently, the lexical **[l]**emma layer was derived from the **f** layer, by selecting all lemmatized words from the **f** layer whose PoS was one of Noun, Adjective, or Verb, using `spacy` for both lemmatization and PoS tagging (Honnibal and Montani, 2017).

Finally, in the induction step, BERT (Devlin et al., 2019) was used, using the `bert-base-cased` model of the `transformers` library.

4 Validation experiments

This section validates the quality of the model. As the extraction of high-quality translation equivalence relations between tokens in the target and reference language is paramount for the validity of subsequent steps, I first evaluate the Liu et al. (2023) model, which provides us with such translation equivalences, in two ways: by assessing if reference language items are aligned with the correct target language tokens (§4.1), and by assessing if the extracted 'lemmas' accurately lemmatize the target language (§4.2). Next, I consider the accuracy of the lexification induction step (§4.3).

4.1 Quality of lemma extractions

To evaluate whether the correct target language tokens are aligned with the reference language word tokens, I use the glosses, available for $32/44$ DoReCo corpora. Given that the target language tokens are associated with a morphological segmentation and a corresponding gloss in English (cf. Table 2 for an example), we can assess whether the target language token aligned with a seed language item contains the seed language item as part of its gloss. For the example in Table 2, the lemma *go* (on the **[l]**exical lemma line) might be aligned with Savosavo *bo*, which is indeed glossed as 'go' (cf. the **[g]**loss line). Only reference language words that are present in at least one gloss in the target language are considered. For instance, the verb *fish* might be aligned with *ghavilighue*, but this word does not have 'fish' in one of its glosses, but rather 'paddle'. Since no other word in the target language has 'fish' in one of its glosses, the item is not counted as correct or incorrect.

We compare the scores of the Conceptualizer mode against a weak baseline of picking a word from the target language sentence at random, and a stronger baseline of a simple extraction procedure in which the alignments over word alignments obtained through either Awesome Align (Dou and Neubig, 2021) or Eflomal (Östling and Tiedemann, 2016) combined with the 'grow-diag-final-and' heuristic were used (for both models, default settings were used). The procedure furthermore involved resolving cases where one reference language word token was mapped onto multiple target language tokens, as the evaluation procedure requires a single target language form. For such cases, only the target language token that was most frequently aligned with the reference language word type across the whole bitext was kept.

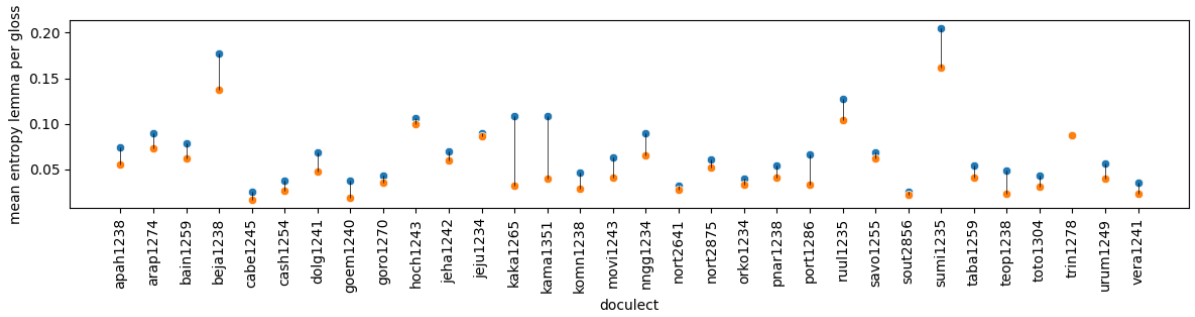


Figure 3: Entropy of lemmas given glosses. Blue: **lemma-H** without merging; orange: with merging.

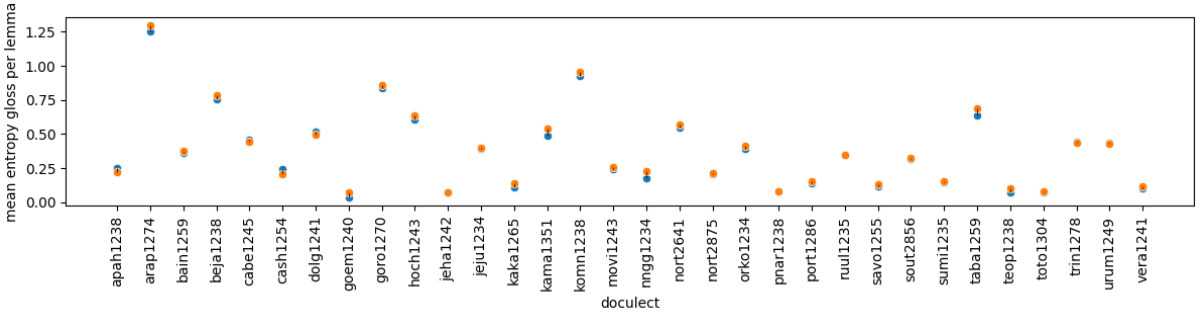


Figure 4: Entropy of glosses given lemmas. Blue: **gloss-H** without merging; orange: with merging.

Figure 2 reports the macro-averaged scores, i.e., averaged over all tokens of seed language items per language. For the Conceptualized model, the median language gets 91.1% of alignments correct, against median accuracy scores of the three baseline models of 19.7% (random), 15.5% (Awesome Align), and 25.4% (Eflomal). The variation between languages is relatively small, with an interquartile range of 88.2%-94.4%, a worst case of 81.0% (Jejuan) and a best case of 97.2% (Toto). Notably, these results substantially support the superiority of the Liu et al. (2023) Conceptualizer model over alignment-based procedures in low-resource scenarios such as the one studied here.

4.2 Effect of lemma merging

While the previous analysis supports the accuracy of the alignments between seed words and target language tokens, it does not yet validate whether the extracted lemmas, to be used in the subsequent induction step, are accurate. It may be that all target language tokens are correctly aligned, but this is done through several lemmas that all correspond to one 'true' lemma as given in the gloss. This would lead to an artificial inflation of the lexical boundaries in the language, which in turn reduces the quality of the inferred representations of crosslinguistic variation. The merger step discussed in §2.2 intends to pre-empt this situation.

It is difficult to assess the quality of the extracted lemmas directly, due to variation in how the glosses are assigned. Because of that, I approach the assessment indirectly, by considering the uncertainty in two conditional probability distributions: of extracted lemmas given annotated glosses, and, vice versa, of glosses given lemmas. I only consider gloss-lemma pairs found to be correctly aligned in the previous evaluation step.

For a target language t, let G_t be the set of all glosses that contain a seed word, i.e., the glosses used to determine the correctness of the alignment in the previous set, and L_t the set of induced lemmas (either as-is from the Liu et al. (2023) procedure, or after the merging step) found in cases of correct alignments. Primarily, I propose to measure the quality through the weighted average uncertainty of the probability of the lemmas given a gloss, or $P(L_t|g)$, for all glosses $g \in G_t$, as weighted by the frequency of occurrence of the gloss among correctly aligned cases, or $N(g)$. In an ideal case, for every gloss, there is just a single induced lemma that aligns to it. If multiple lemmas are found, aligning to the same gloss, the model might have inferred spurious lemmas. Formally, **lemma-H**$(t) =$

$$\sum_{g \in G_t} \Big(H(P(L_t|g)) \times N(g) \Big) \times \frac{1}{\sum_{g \in G_t} N(g)} \quad (1)$$

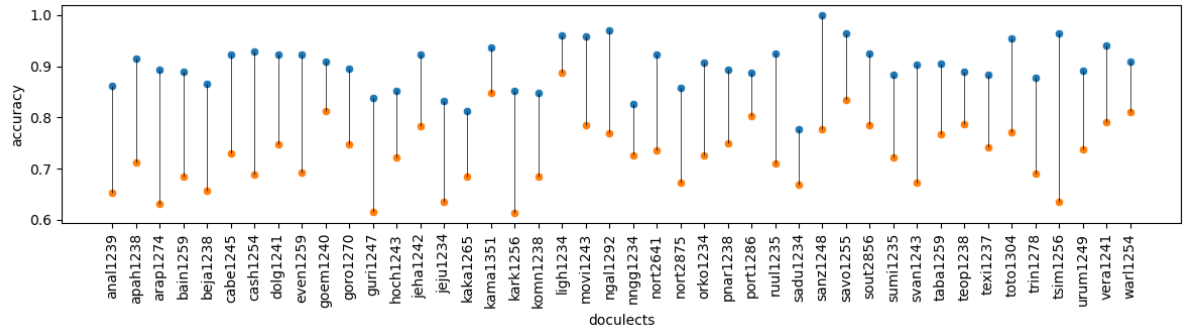


Figure 5: Accuracy on classifying the lemma for held-out data; blue: MLP-100 classifier, orange: most-frequent lemma per seed word baseline.

The inverse relation, of glosses given lemmas, similarly has an expectation of one-to-one mappings: given a lemma, we expect it to align with one unique gloss only. For this relation, however, a qualification applies: for many of the languages in the corpus, the individuable glosses contain more than just the lemma, and as such individual lemmas frequently align with multiple unique glosses, with substantial variation between the languages owing to the different approaches to writing the glosses that the documentary linguists applied. Nonetheless, the gloss entropy given the lemmas is a useful measure when assessing the effect of the merging step: if applying the merging step leads to erroneous mergers, i.e., cases where two induced lemmas are merge that should not be merged, the uncertainty over the glosses given the lemmas should go up, as the original lemmas that were erroneously merged can be expected to have rather different sets of glosses. As such, it can be expected that if the merging is accurate, the entropy over the glosses given the lemmas should *not* go up relative to the application of the model *without* the merging step. Formally, the **gloss-H** measure is defined as:

$$\sum_{l \in L_t} \Big(H(P(G_t|l)) \times N(l) \Big) \times \frac{1}{\sum_{l \in L_t} N(l)} \quad (2)$$

Figure 3 shows that across languages the **lemma-H** goes down with the addition of the merging step for each individual language, with some positive outliers being Kakabe and Kamas, where most of the uncertainty over the glosses is removed by adding the merging step (**lemma-H** values going from 0.109 to 0.032 for the former and 0.109 to 0.040 for the latter). On average, the **lemma-H** was found to decrease from 0.072 when the merging step is not applied, to 0.053 when it is applied.

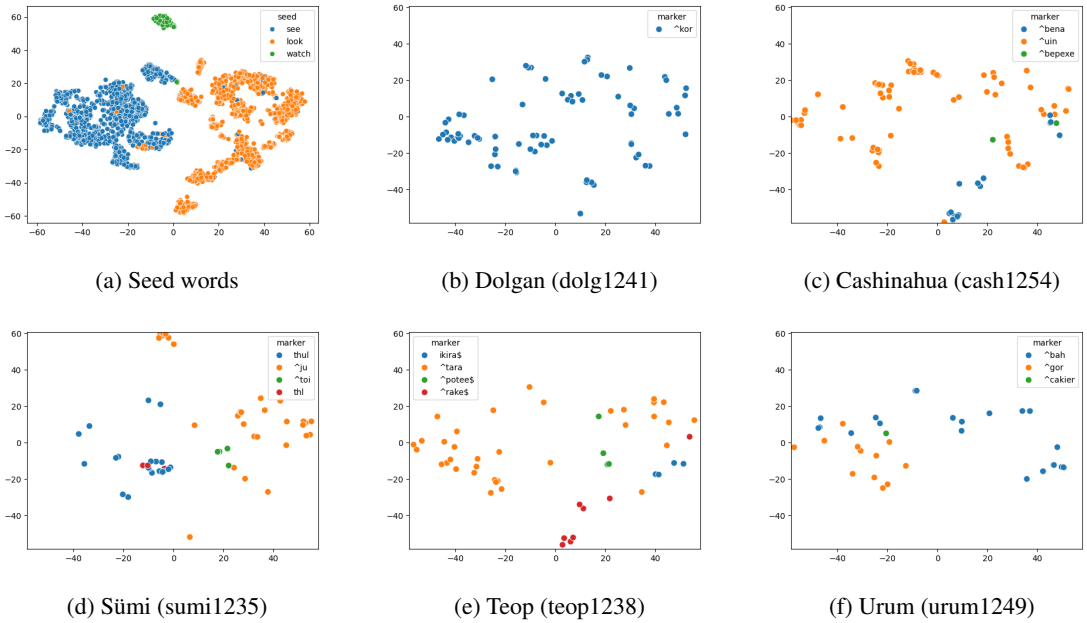
model	accuracy	ERR
baseline	0.739	-
KNN-3	0.862	0.491
SVC	0.890	0.594
MLP	0.898	0.624
MLP-100	0.900	0.631

Table 3: Induced lexification results across all languages; ERR = error rate reduction.

Conversely, the merging step does not introduce substantial new uncertainty in the $P(G_t|l)$ distributions due to erroneous lemma mergers. Compared to the magnitude of the **gloss-H** values when no merging step is applied, the **gloss-H** values when merging *is* change relatively little, as Figure 4 illustrates on a language-by-language basis. Only in 6 cases does the **gloss-H** value go up with the addition of the merging step, compared to 19 cases where it goes down, meaning that on the whole, adding the step in fact *reduces* the uncertainty over the glosses given the lemmas.

4.3 Quality of induced lexification

The two validation experiments suggest that the inferred lemmas align reasonably well with the linguistic annotations provided in the corpus. While the goal of the induction procedure is to infer the target language lemmas given contextualized usages of target words for *other* target languages, we can assess the quality of the induction procedure by assessing the classification accuracy on a held-out sample of the *same* language. For each of the 44 languages, all seed words occurring with a frequency of 10 or more were considered, and K-fold cross-validation (here: $K = 20$) over the entire lexicon of some target language t was carried out.



(a) Seed words (b) Dolgan (dolg1241) (c) Cashinahua (cash1254)

(d) Sümi (sumi1235) (e) Teop (teop1238) (f) Urum (urum1249)

Figure 6: t-SNE plots with various colour coding. For the five languages, only the observed tokens are shown.

The accuracy of this procedure was then compared against a baseline of always predicting the most common lemma given a reference language seed word, reflecting a scenario in which a model only knows the input word in the reference language. I assessed four classifiers implemented in the `sklearn` library, a k-nearest neighbours classifier with $k = 3$ (KNN-3), an Support Vector Classifier with the default settings (SVC), and two Multi-Layer Perceptron, one with no hidden layers (MLP), and the other with one hidden layer of 100 units and ReLU activation (MLP-100).

Table 3 presents the results. Seed words tend to be associated with few lemmas, one of which is typically very dominant (cf. the low entropy of the lemmas given the glosses in Figure 3, which supports this observation). As such predicting the modal lemma given a reference language seed word forms a competitive baseline. All classifiers, however, provide substantial improvement over the baseline, reducing the error by between 49% (KNN-3) and 63% (MLP-100). Figure 5 shows the results per language for the best-performing MLP-100 model, showing that the classifier surpasses the baseline and generally performs well for all languages.

5 Application

The previous section demonstrated that the model extracts generally valid target language representations (lemmas) and is reasonably well able to classify these lemmas on the basis of contextual vector representations of the seed language. The goal of this approach, however, is to provide a method for typologists to obtain massively comparable data in the absence of a massively parallel corpus. This section demonstrates how known insights can be replicated, and how novel insights can be obtained with the method.

To explore the comparability afforded by the model, here, we briefly explore the domain of visual perception verbs, translation equivalents of English *see, look,* and *watch*. A main lexical distinction between Experiencer and Activity verbs (English *see* vs. *look*) – with the former involving a more passive ('experiencing') role for the perceiver, and the latter a more active one has been postulated (Viberg, 1983), but challenged on the basis of parallel corpus data by Wälchli (2016). Using manually extracted instances from comparable corpora, San Roque et al. (2018) consider the non-literal extensions of perception verbs, noting that discourse markers (e.g., *look!* to draw attention or introduce something unexpected) are common extensions.

To explore the distribution of visual perception verbs in the DoReCo corpus, we can train the best classifier from §4.3 (MLP-100) for each language that has $N \geq 30$ instances of the three most common English visual perception verbs (*see, look,* and *watch*) in their free translations. Next, we apply

this classifier to all instances of *see, look*, and *watch* for all other languages, leading to a 3001 (instances of visual perception verbs across all 29 languages with sufficient data) by 194 (unique lemmas across the 37 languages with sufficient data) table, with the probability assigned by each MLP-100 model to the lemmas as the cell values. To visualize this table, we can apply t-SNE (Van der Maaten and Hinton, 2008) to reduce the table to two dimensions.

Figure 6 shows the t-SNE representation, with six distinct colour-codings. The top-left subfigure (6a) shows the distribution of the three English seed words, which form coherent groups of visual clusters, but with each term nonetheless covering multiple clusters. Some languages, such as Dolgan (6b) do not make any lexical distinctions in this domain – a situation predicted by Viberg (1983)), while others, such as Sümi (6d) split Activity and Experiencer meanings more or less along the lines of English. Two languages carve out a cluster near the bottom of the 2D-space – Cashinahua *bena* and Teop *rake* – these are all instances of *look for*, meaning 'search', which many languages group with the other 'look' meanings, but these two languages distinguish lexically. Finally, Urum (6f) presents an interesting case of two main terms, but with a split that differs from English or Sümi. Here, we see that *bah* covers a region containing English *look* and some of *see*, whereas *gor* covers only part of the *see* tokens. The *see* tokens covered by *bah* involve cases of modal see, like *can see*, *will see*, in several cases in the meaning 'find out', like "I will see where to go, possibly to the city". As such, Urum supports the argument of Wälchli (2016) that the Activity-Experiencer split is (a) more of a continuum, and (b) governed by properties beyond the general semantic role of the perceiver.

What the plots in Figure 6 further illustrate, is that languages differ in how often they use visual perception meanings. Urum uses visual perception verbs only 21 times per $10,000$ tokens, whereas Cashinahua shows five times that frequency at 102 tokens per $10,000$. Such usage variation is known to be meaningful in the explanation of lexification patterns, following the argument that a language's greater need to communicate about a specific concept correlates with finer-grained lexical distinctions (cf. Kemp et al., 2018). Original corpus data and methods for making such data comparable can thus be used to estimate such 'need probabilities'

6 Conclusion

This paper introduced a novel method for making original text corpora that are translated into the same reference language comparable, thus allowing for token-level typological study. The independent steps of the method were found to generally provide high-quality representations in three validation experiments, and the case study presented the potential of the method for studying lexical semantic variation across languages.

While generally successful in extracting translation equivalents and inducing lexical categorization models, room for improvement remains. While the Liu et al. (2023) approach benefits from its ability to consider substrings below the word level, it is hampered by not considering how other target language substrings translate to the seed item, something word alignment procedures from IBM-1 (Brown et al., 1993) onward do consider.

It should be stressed here that using original text does not make the method bias-free, in terms of a translationese bias from the shared reference language. Using the free translations means all lexical choice models are filtered through contextual vector representation of English. In the specific case of the data used here, this English is moreover written as a guide for the linguistically informed reader to make sense of the target language sentence; it may, by design given the genre of "free translations in language documentation", show translation effects from the target language onto the English. Calibrating the extent of this effect would require further testing the model on other comparable corpora.

Applications beyond the ones the method was designed for could be explored. Related work that considers crosslinguistic variation at a word type level and using secondary resources, like Thompson et al. (2020) and Khishigsuren et al. (2025), could be compared against the token-level mappings between a shared reference language and multiple target languages. Corpora that contain both original and translated text in comparable genres may furthermore be of use to pinpoint the precise effects of translationese in how lexical boundaries are drawn, and as such be of use for practical purposes in education and translation studies. Finally, we are reminded that languages vary on a discourse-pragmatic level, and that multilingual NLP ought to consider such variation, for instance when working with Large Language Models and Machine Translation systems pretrained on translated text.

References

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

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

Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

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

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

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

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

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

Renata Enghels, Bart Defrancq, and Marlies Jansegers. 2020. *New Approaches to Contrastive Linguistics: Empirical and Methodological Challenges*. Walter de Gruyter GmbH & Co KG.

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

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

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

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

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

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

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

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

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

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

Martin Haspelmath. 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.

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

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

Birgit Hellwig, Gertrud Schneider-Blum, and Khaleel Bakheet Khaleel Ismail. 2024. Tabaq (Karko) DoReCo dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 2.0*. Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Lyon.

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

Stig Johansson and Knut Hofland. 1994. Towards an english-norwegian parallel corpus. In G. Tottie Fries and P. Schneider, editors, *Creating and using English language corpora*, pages 25–37. Rodopi, Amsterdam.

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

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

Temuulen Khishigsuren, Terry Regier, Ekaterina Vylomova, and Charles Kemp. 2025. A computational analysis of lexical elaboration across languages. *Proceedings of the National Academy of Sciences*, 122(15):e2417304122.

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

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

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.

Natalia Levshina. 2017. Online film subtitles as a corpus: An n-gram approach. *Corpora*, 12(3):311–338.

Natalia Levshina. 2021. Corpus-based typology: applications, challenges and some solutions. *Linguistic Typology*, pages 129–160.

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.

Anthony McEnery and Zhonghua Xiao. 2007. Parallel and comparable corpora: What are they up to. *Incorporating corpora: translation and the linguist*, pages 18–31.

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

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

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

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

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

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

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.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

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

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

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

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

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

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

Lila San Roque, Kobin H Kendrick, Elisabeth Norcliffe, and Asifa Majid. 2018. Universal meaning extensions of perception verbs are grounded in interaction. *Cognitive Linguistics*, 29(3):371–406.

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

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

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

Deborah Tannen. 1980. A comparative analysis of oral narrative strategies: Athenian Greek and American English. In Wallace L. Chafe, editor, *The Pear Stories: Cognitive, cultural, and linguistic aspects of narrative production*, pages 51–87. Ablex Publishing Company Norwood^ eN. J NJ.

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

Marina Terkourafi. 2011. The pragmatic variable: Toward a procedural interpretation. *Language in Society*, 40(3):343–372.

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

Bill Thompson, Seán G Roberts, and Gary Lupyan. 2020. Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour*, 4(10):1029–1038.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

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

Annemarie Verkerk. 2014. Where Alice fell into: Motion events from a parallel corpus. In Benedikt Szmrecsanyi and Bernhard Wälchli, editors, *Aggregating*

dialectology, typology, and register analysis: Linguistic variation in text and speech, pages 324–354.

Åke Viberg. 1983. The verbs of perception: A typological study.

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

Bernhard Wälchli. 2016. Non-specific, specific and obscured perception verbs in Baltic languages. *Baltic Linguistics*, 7:53–135.

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

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

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

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