Discursive distinctiveness explains lexical differences between languages

Barend Beekhuizen[1]

[1]Department of Language Studies, University of Toronto Mississauga, Canada
Department of Linguistics, University of Toronto, Canada
barend.beekhuizen@utoronto.ca

Languages differ in their lexical semantic inventories. Recently, such differences have been explored through the lens of differences in the frequencies with which certain concepts are employed in discourse. In this paper, I depart from such work by suggesting that it is not merely the frequency of usage, but also the diversity of ways in which concepts are used that explains whether languages group together two concepts with a single lexical item. I provide a theoretically grounded account of why we should expect this to be the case, and develop a methodology for operationalizing such ideas in a multilingual corpus, finding that variation in the discursive practices of using words indeed predicts whether languages co-express concepts or split them.

1. Introduction

Languages differ in their inventories of lexically encoded meanings. While English co-expresses brothers of both parents as *uncle*, Croatian distinguishes *stric* 'father's brother' from *ujak* 'mother's brother'. Such crosslinguistic variation is the outcome of the cultural evolutionary processes through which only some word meanings are replicated in a community of users. Kemp, Xu, and Regier (2018) explore the usage frequency ('need probability') of concepts as a communicative pressures on the processes of replication: the more often a concept is brought up in discourse, the less likely it is to be co-expressed ('colexified', cf. François, 2008) with similar concepts. This insight has been fruitfully applied to various domains: colour (Twomey, Roberts, Brainard, & Plotkin, 2021), precipitation (Regier, Carstensen, & Kemp, 2016), and kinship (Anand & Regier, 2023).

Here, I propose that rather than the 'need' to express a concept, it is itextb-fusage diversity, the diversity of *the ways in which* the concept is employed in discourse that forms a source of selective pressure on word meanings. After presenting the theoretical motivation, I provide support for this position using computational methods, cross-linguistic corpus data, and a lexicon-wide sample of concepts. This result contributes to a more complete account of the pressures shaping the lexicon.

2. Background

The proposed connection between the discursive use of lexical semantic concepts and the inventories of word meanings is motivated by various starting points.

First, lexical selection in usage events is influenced by the inferences afforded by the expressed concepts (Anscombre & Ducrot, 1983; Rommetveit, 1974): we pick lexical items because they steer towards certain conclusions. These inferences do not universally derive from the concept itself (knowing a concept does not entail knowing how it should be used; Goodwin, 1994), but instead depends on semi-conventional 'practices' of speaking (Hanks, 2018) – sets of behavioural patterns governing how a word ought to be used. Such practices of speaking, then, being cultural phenomena, are tethered to a language community, and as such may differ between language communities, as noted by (Hymes, 1961). This motivates the assumption of this paper that the 'rules of use' of lexical items expressing the same or very similar concepts may differ across languages.

Acknowledging a role for language-specific practices of lexical selection is only half of the story. The second half consists of linking those in-the-moment lexical choices to population-level conventions. One proposal to do so comes from Enfield (2014), who takes the in-the-moment decisions, dubbed the 'enchronic' dimension of language, to be one of the 'natural causes' of why language structures are the way they are. Enfield develops a useful conception of how such in-the-moment decisions 'percolate up' to population level conventions in a later paper (Enfield, 2023) in which he argues that part of understanding how concepts are used in discourse is understanding what interpretive effects they have in the past given rise to. Croft (2000) similary takes this 'pool' of experienced usage events to be the source of the selective replication of certain variants over others.

The assumed cultural-evolutionary process for my case is similar. When, in a community, the conventional ways of using two similar concepts are also similar to each other, there is little need to lexically distinguish them, and so new lexical items expressing only one concept are unlikely to emerge and spread. Conversely, when the conventional ways of using the two concepts are different from each other, the concept-level similarity (which might lead to colexification) competes with the dissimilarity on the level of the practice of usage, and we can expect a greater likelihood for e.g., novel lexical items specializing for the expression of one of the concepts to emerge. This paper aims to demonstrate the consequences of these hypothesized pathways for the crosslinguistic patterning of colexification.

3. Method

Studying variation in the discursive usage of word meanings requires a substantially novel set of corpus methods in order to make the crosslinguistic comparison between usage events possible. My method draws on the translation into a shared language (English) to do so. A succinct description is given here, with more in-

formation and code made available as part of a planned journal paper.

Corpus: I use the DoReCo corpus (Seifart, Paschen, & Stave, 2022), a typologically diverse sample of fieldwork-based documentation of 51 spoken languages. For comparability, only narrative data was used, resulting in corpora of 500 to 76,000 word tokens per language, with 4 languages excluded for having no narrative data. Around half of the languages have glosses provided for them (e.g., Ex. (1)-(2)), whereas for the remaining languages only the free translation is available (e.g., Ex. (3-4)).

(1) nam na toku nom tea gono ta peha taba tahii
1PL.EX.PRON TAM2 not.know IPFV COMPL1 get NSPEC2.SG one2 thing sea
'we - we don't know (how) to get anything from the sea.'

(2) a abana paa nata vaevuru tea vagana
ART2.SG men TAM3 know already COMPL1 go.fishing
the men already knew to fish
Teop; Austronesian, Papunesia; (Mosel, 2022)

(3) tayley katiji kastellano
'I already knew Spanish.'

(4) nish taylejtij
'We do not know.'
Yurakaré; Isolate, South-America; (Gipper & Ballivián Torrico, 2022)

Extraction of translation equivalents: I use word tokens in the free translations to compare how 'the same' concept is expressed across languages. Using SpaCy spacy2, I selected all free translations tokens with 'lexical' parts of speech (nouns, adjectives and verbs) and lemmatized them. Next, the most likely orthographic segments and corresponding tokens for each lemma were extracted from the source set using the best-matching string procedure of (Liu et al., 2023). For instance, in Ex. (1) above, for the three lexical items *know*, *get* and *sea*, the Teop strings *toku*, *gono*, and *tahii* were identified as translation equivalents. For the morphologically more complex language Yurakaré (Exx. (3-4)), the English lexical item *know* was linked to the substring *yle* of *tayley*.[1]

Token-level comparability: Massively parallel corpora (e.g., Bible translations) alluw us to compare patterns of colexification through translations of the same source language utterance into all the target languages, but they don't let us study how concepts are used differently in discourse across languages, as the translations all draw on the same pattern of verbalization in the source language. This motivated the present use of a non-massively parallel corpus that nonetheless has translations *into* a shared target language. To make tokens comparable across languages, I apply computational linguistics techniques for representing the usage of a word through contextualized distributional semantic representations (CDSRs)

[1]The extraction method was found to be highly reliable: evaluating the procedure by considering, for the languages with glosses available, whether the orthographic segment extracted given a free translation matches a target language token glossed with the free translation, we found that the extraction procedure performed at 89% precision and 88% recall (cf. 19% precision/recall if guessing randomly).

in the form of high-dimensional vectors. When tokens of a word are used in similar contexts, their CDSRs will be more similar to each other than when used in different contexts. We expect the CDSRs for *know* in Ex. (1) and Ex. (4) to be similar, as well as those for *knew* in Ex. (2) and Ex. (3), given that each pair represents a similar context. CDSRs for all tokens were retrieved using BERT (`bert-base-cased`; Devlin, Chang, Lee, & Toutanova, 2018).

We can then use the CDSRs to train a supervised classifer to predict the lexical choice in a particular language. Here, I am using the linear Support Vector Machine classifier of `sklearn` (Pedregosa et al., 2011).[2]. A trained classifier allows us to ask, for a token of an English lemma, how that token would be translated in any other language. In other words, we can determine, given the CDSR for *know* in Ex. (4), that the Teop-trained classifier would pick *toku* (as in Ex. (1)), rather than *nata* (as in Ex. (2)), given that the former's contexts are more similar. Doing so for every token and every language, we arrive at a $146,821$-by-47 token-by-language table, where for every token (row) we have the inferred lexical item for each of the target languages (column) in the cells of the table.

Defining lexical fields: To analyze variation in colexification patterns, we need sufficiently large groups of tokens that display crosslinguistic variation. I use the imputed extension of all $9,534$ extracted terms as the starting point, as they reflect groups of tokens colexified by at least one language. I then pairwise merged (by taking the union) term extensions with a Jaccard similarity of $\geq .90$ in order to avoid redundancy (which would affect the regression analyses in Sec. 5), leading to $8,210$ groups of tokens or 'fields'.

Given that the data is a (dummy coded) binary valued table, I ran logistic PCA (Collins, Dasgupta, & Schapire, 2001) to study the patterning of the variation between languages (using the `logisticPCA` library in `R`). Only the first principal component was used as further components might be redundant with the first component of extensions of other terms.

4. A look at the PCA spaces:

As an exploration, I consider a group of tokens colexified by Asimjeeg Datooga (Nilotic, Africa: Griscom, 2022) *nal*, which nearly all translate to English *know*. To understand what the variation along the first component (PC1) of a logistic PCA means, I considered the free translations for the tokens with the lowest and highest value on PC1. The former are overwhelmingly cases of present-tense negated know (e.g., *We don't know because it's a stranger's plan*), whereas the latter consist mainly of instances of past-tense know (e.g., *It was that (which) they knew*). Teop (Mosel, 2022), in Fig. 1a appears to dislexify these two functions,

[2]Classifiers were evaluated using 100-fold cross-validation. The model obtained 87% accuracy in predicting the target language lexical item – outperforming an informed baseline (guessing the most frequent translation equivalent given the English free translation lemma) obtaining 74% accuracy.

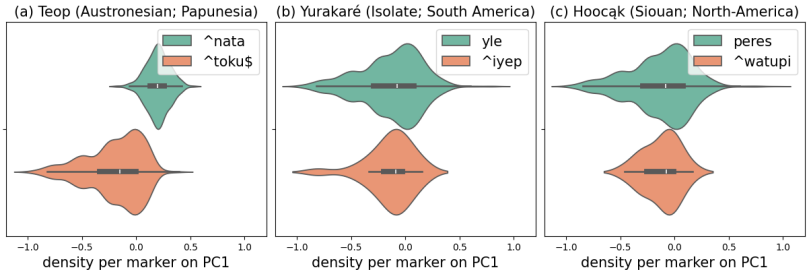

Figure 1. Examples of languages on PC1 of the Asimjeeg Datooga *nal* field.

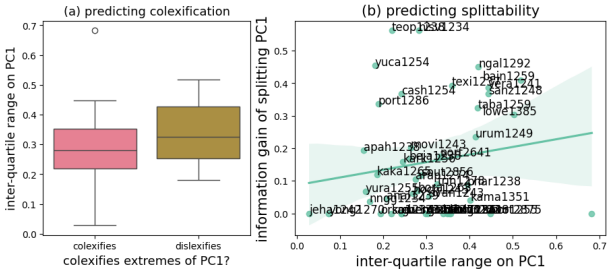

Figure 2. Demonstrating the negative correlation of usage diversity and colexification.

with the two markers seen in the example occupying left and right positions on PC1. Other languages may have multiple terms, but ones that don't line up with the distinction on PC1. Yurakaré (Gipper & Ballivián Torrico, 2022; Fig. 1b), for example, has a second term *iyep* that translates to Spanish *conocer* 'know someone' while *yle* translates to *saber* 'know something', and Hoocąk ((Hartmann, 2022); Fig. 1c), which has, per the provided glosses, a 'know-how' (*watupi*) and a 'know-that' (*peres*) verb. Note that for both languages, the two terms are not linearly separable on PC1.

Notably, the greatest density of Yurakaré and Hoocąk tokens is around the middle of PC1 whereas the tokens of Teop appear to be more spread out. This observation is in line with the central thesis of this paper, that languages that display greater usage diversity colexify less. In particular, I argue that greater variation in the usage tokens on a semantic scale (such as PC1) will go hand in hand with a lower propensity of colexifying the two ends of the scale. Here, I operationalize the usage diversity through the use of the inter-quartile range (IQR) of tokens of a language on the PC1 of a logistic PCA over a group of tokens. As the dependent measure we can consider (1) whether languages would categorize the two extreme points of PC1 with the same term or not (using an SVC trained on the observed markers for each language), and (2) how 'splittable' the language is along the axis. The latter measure makes colexification a continuum by considering the highest information gain (IG) of splitting anywhere along PC1 for a particular lan-

guage: languages that 'lump' will have a zero IG, whereas languages with a 50/50 split between tokens, perfectly splittable along PC1, will have a high IG, and languages with uneven frequency distributions and less-than-perfect splits will fall in between these extremes. Fig. 2a and 2b demonstrate the covariance of two dependent measures covary with usage diversity (IQR) for the *nal* field, showing colexifying languages have a lower IQR than dislexifying ones, and the IG measure correlates positively with the IQR.

5. A lexicon-wide study

Does this correspondence hold in the lexicon at large? I contrast usage diversity (through the IQR) with need probability, defined as the log-transformed word-per-million count of the tokens of a language in the group of tokens considered. Several groups of tokens were omitted for displaying too little variation. This leaves us with 4,679 groups of tokens and 33,843 observations (values for individual languages per field). For our two dependent variables (colexification and splittability) we fit a logistic resp. a linear regression over the two independent variables, z-transforming them for comparability.

For **colexification**, a higher need probability predicts less **colexification** ($\beta = -.08, p < .001$) and a higher usage diversity also predicts less colexification ($\beta = -.72, p < .001$). Both effects are in the expected direction. Moreover, comparing the β values informs us that usage diversity is the more impactful predictor, suggesting that it is not the mere need probability, but the make-up of the discursive need to use a concept that explains differences in colexification. Similarly, we find an effect of usage variation on the **splittability** (information gain) measure in the expected direction ($\beta = .13, p < .001$) but a (smaller) effect for need probability ($\beta = -.02, p < .001$), in the opposite direction, predicting more splittability the less frequent a group of tokens is instantiated for a language.

6. Discussion

This paper studies crosslinguistic variation using naturalistic data for which a substantial methodology had to be developed. The pay-off is that we can study the factors explaining divergence in lexical inventories at scale and using discursive factors that would otherwise not be accessible. The central finding is that crosslinguistic differences in the ways word meanings are used in discourse covary with the types of lexical inventories. This is an initial finding that encourages further consideration of usage events as loci of selectional pressures on the lexicon.

Substantial questions remain, such as the direction of causality between discursive practices and lexical inventories. One could argue that a language having two lexical items nudges their discursive applications to be more distinct. This would not be entirely unexpected, so it is a possibility that we are dealing with a loop, where more discursive distinctiveness drives the need for separate lexicalization, in turn increasing the likelihood of more distinct discourse practices.

Acknowledgements

This research was made possible by an NSERC *Discovery* Grant to the author (RGPIN-2019-06917). I would like to thank audiences at the 2022 International Cognitive Linguistics Conference and the 2022 Societas Linguisticae Europaea for feedback on earlier versions of this work. I am very grateful for the substantial methodological comments presented by the EVOLANG reviewers. I would finally like to express my gratitude to the developers and contributors to the DoReCo corpus, without whom this research would not have been possible.

References

Anand, G., & Regier, T. (2023). Kinship terminologies reflect culture-specific communicative need: Evidence from hindi and english. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45).

Anscombre, J.-C., & Ducrot, O. (1983). *L'argumentation dans la langue.* Editions Mardaga.

Collins, M., Dasgupta, S., & Schapire, R. E. (2001). A generalization of principal components analysis to the exponential family. *Advances in neural information processing systems*, *14*.

Croft, W. (2000). *Explaining language change: An evolutionary approach.* Pearson Education.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Enfield, N. (2023). Linguistic concepts are self-generating choice architectures. *Philosophical Transactions of the Royal Society B*, *378*(1870), 20210352.

Enfield, N. J. (2014). *Natural causes of language: Frames, biases, and cultural transmission.* Language Science Press.

François, A. (2008). Semantic maps an the typology of colexifications: Intertwining polysemous networks across languages. In M. Vanhove (Ed.), *From polysemy to semantic change: Towards a typology of lexical semantic associations* (pp. 163–216). Amsterdam: John Benjamins.

Gipper, S., & Ballivián Torrico, J. (2022). Yurakaré DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language documentation reference corpus (doreco) 1.2.* Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2).

Goodwin, C. (1994). Professional vision. *American anthropologist*, *96*(3), 606–633.

Griscom, R. (2022). Asimjeeg Datooga DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language documentation reference corpus (doreco) 1.2.* Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & labo-

ratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2).
Hanks, W. F. (2018). *Language & communicative practices.* Routledge.
Hartmann, I. (2022). Hoocąk DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language documentation reference corpus (doreco) 1.2.* Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2).
Hymes, D. H. (1961). On typology of cognitive styles in language. In *Anthropological linguistics: Uses of typology in language or culture or both: A symposium presented at the 1960 meetings of the american anthropological association* (Vol. 3, pp. 22–54). JSTOR.
Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*(1).
Liu, Y., Ye, H., Weissweiler, L., Wicke, P., Pei, R., Zangenfeind, R., & Schütze, H. (2023). A crosslingual investigation of conceptualization in 1335 languages. *arXiv preprint arXiv:2305.08475*.
Mosel, U. (2022). Teop DoReCo dataset. In F. Seifart, L. Paschen, & M. Stave (Eds.), *Language documentation reference corpus (doreco) 1.2.* Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2).
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.
Regier, T., Carstensen, A., & Kemp, C. (2016). Languages support efficient communication about the environment: Words for snow revisited. *PloS one*, *11*(4), e0151138.
Rommetveit, R. (1974). *On message structure: A framework for the study of language and communication.* John Wiley & Sons.
Seifart, F., Paschen, L., & Stave, M. (Eds.). (2022). *Language documentation reference corpus (DoReCo) 1.2.* Berlin & Lyon: Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2).
Twomey, C. R., Roberts, G., Brainard, D. H., & Plotkin, J. B. (2021). What we talk about when we talk about colors. *Proceedings of the National Academy of Sciences*, *118*(39), e2109237118.