Communicative Need Modulates Lexical Precision Across Semantic Domains: A Domain-Level Account of Efficient Communication

Laurestine Bradford (laurestine.bradford@mail.mcgill.ca)
Department of Linguistics, McGill University

Guillaume Thomas (guillaume.thomas@utoronto.ca)
Department of Linguistics, University of Toronto

Yang Xu (yangxu@cs.toronto.edu)
Department of Computer Science, Cognitive Science Program, University of Toronto

Abstract

Different domains exhibit different degrees of lexical precision. Existing work has suggested that communicative need may modulate the precision of word meaning in individual domains. We extend this proposal across domains by asking why languages have more precise vocabulary in some domains than others. We hypothesize that lexical precision for a domain reflects how frequently speakers need to refer to it. We test this proposal using a cross-linguistic dataset of word-concept mappings for nine diverse domains from seven languages, and word frequencies from independent corpora. We find that the more frequent domains (except for kinship) tend to be more precise in every language, supporting a domain-level account of efficient communication on the precision of the lexicon.

Keywords: the lexicon; lexical precision; semantic domains; communicative need; efficient communication

Introduction

Natural languages have vocabularies for expressing a diverse range of domains, but not all domains share the same level of lexical precision. For instance, the English vocabulary has precise terms for expressing weathers such as snow and rain, but it is ambiguous for expressing kin relations such as cousin (as a child of one's uncle or aunt, from mother's or father's side). Why are certain semantic domains in the lexicon more precise than others? We investigate this question in a computational analysis of semantic domains across languages.

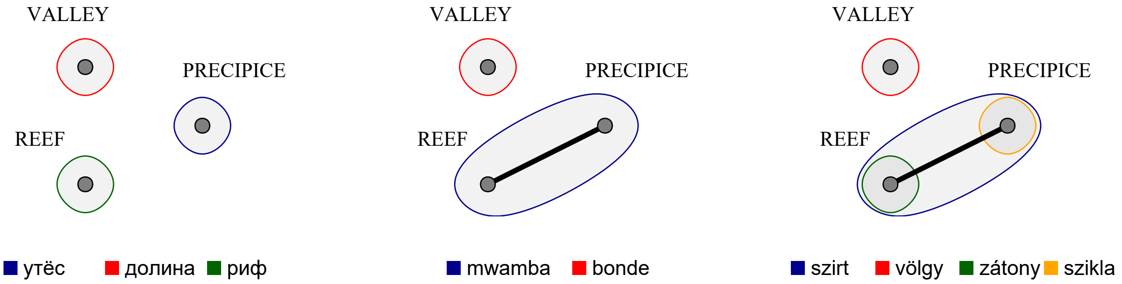
All languages exhibit lexical ambiguity. That is, words often have multiple meanings. For example, the Hungarian word *szirt* can refer to either PRECIPICE or REEF (see Figure 1c). This phenomenon is sometimes also known as colexification (François, 2008; Xu, Duong, Malt, Jiang, & Srinivasan, 2020), whereby a single word form labels multiple meanings. However, lexical ambiguity is constrained. Words are more likely to encode multiple meanings when these meanings are semantically related (Floyd, Dalawella, Goldberg, Lew-Williams, & Griffiths, 2021; Karjus, Blythe, Kirby, Wang, & Smith, 2021; Xu, Duong, et al., 2020). Here we focus on the cases where a word form encodes highly related meanings for a semantic domain, i.e., a special form of colexification known as underspecification (François, 2008).

Underspecification can impede communication. For instance, it would be more effortful to distinguish RAIN and SNOW in communication if these concepts are lexicalized under the same word form, in comparison to the case where they are labelled under distinct word forms. This issue can be exacerbated when the underlying concepts are frequently talked about in language, because these underspecified cases of colexification would likely cause constant ambiguity in communication even when context is taken into account. On the contrary, if a concept rarely needs to be mentioned in language, we might expect underspecified cases of colexification to be more tolerable, since the cost incurred due to ambiguity would presumably be low given the low need for communicating such concepts.

Indeed, recent work has suggested that communicative need—how frequently a concept is talked about or needs to be made distinct from other concepts—may modulate the probability of colexification and hence the precision of vocabulary (Hawkins, Franke, Smith, & Goodman, 2018; Karjus et al., 2021). In particular, it has been shown that concept pairs such as SNOW and ICE are more likely to have distinct word forms from languages spoken in regions of cooler climate due to greater needs for expressing and distinguishing them (Regier, Carstensen, & Kemp, 2016). Similar work has also shown how communicative need may shape lexical precision within an individual domain such as kinship (Kemp & Regier, 2012), i.e., why we do not have a single term that labels MOTHER and FATHER, reflecting the view that semantic domains across languages are structured to support efficient communication (Kemp, Xu, & Regier, 2018).

The drive for communicative efficiency appears to have shaped the lexicons of extant languages. Cross-linguistically, systems of vocabulary do not take on theoretically possible, but inefficient, configurations. This has been demonstrated in several individual domains, including but not restricted to spatial relations (Khetarpal, Neveu, Majid, Michael, & Regier, 2013), kin relations (Kemp & Regier, 2012), numerals (Xu, Liu, & Regier, 2020), and color (Conway, Ratnasingam, Jara-Ettinger, Futrell, & Gibson, 2020; Kågebäck, Carlsson, Dubhashi, & Sayeed, 2020; Zaslavsky, Kemp, Tishby, & Regier, 2020). For surveys of recent work on the role of efficiency in linguistic typology, see Gibson et al., 2019; Kemp et al., 2018. This crosslinguistic tendency toward efficient communication has also been demonstrated in language learners (Fedzechkina, Jaeger, & Newport, 2012; Kanwal, Smith, Culbertson, & Kirby, 2017), and may relate to the joint evolutionary pressures of learnability and com-

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*



(a) Russian

(b) Swahili

(c) Hungarian

Figure 1: An illustration of lexical precision based on colexification graphs. Portions of the CLICS3 colexification graphs for the Physical Geography domain in (a) Russian, (b) Swahili, and (c) Hungarian. Each colored border delineates concepts lexicalized by one word in the language.

municative expressivity (Carr, Smith, Culbertson, & Kirby, 2020; Kirby, Tamariz, Cornish, & Smith, 2015).

Our current investigation extends beyond the analysis of an individual semantic domain toward multiple (and diverse) domains. We want to understand whether similar principles of communicative efficiency might explain the cross-domain variation in lexical precision in different languages.

Efficient communication at the domain level

In order for a system of vocabulary to be simple without substantial loss of communicative precision, ambiguity should be localized in less important parts of the lexicon. For example, since someone living in Toronto would usually encounter dogs far more often than agoutis, it would be inefficient for their vocabulary to include words for different types of agouti, but not for different breeds of dog. The added complexity would not add much communicative potential. Instead, it is more efficient for them to use those cognitive resources learning breeds of dogs, and simply call all agoutis by one label. More generally, in an efficient linguistic system, those pairs of meanings that are more often relevant should be colexified less often. This is indeed the case for the concepts of ice and snow, which are colexified more in warmer climates than in cooler ones (Regier et al., 2016). Similarly, in an experiment based on a communication game, participants prefer to colexify concepts that are less often needed (Karjus et al., 2021). In an efficient vocabulary system, as communicative need goes up, colexification (or more relevantly to our study, underspecification) goes down.

The hypothesis. We propose that, within a given language, domains with higher communicative need (i.e., high-frequency) should tend to exhibit more lexical precision (i.e., low-ambiguity) than domains with less need. Our proposal is inspired by work suggesting that communicative need at the object-level and the domain-level may drive efficient structuring of semantic systems within and across domains respectively (Kemp et al., 2018). Although the idea that communicate need drives efficiency has been tested within individual domains (Kemp & Regier, 2012; Khetarpal et al., 2013; Xu, Liu, & Regier, 2020; Zaslavsky et al., 2020), whether the same idea holds across domains is an open question that we pursue here. If it does, this will suggest that communicative efficiency also explains cross-domain differences in lexical precision beyond semantic structures within a single domain.

Crosslinguistic data

We collected two primary sources of data across languages. Data for quantifying lexical precision came from the colexification database CLICS3 (Rzymski et al., 2020), and data for estimating communicative need came from word usage frequencies in different text corpora independent to CLICS3.

Database of colexification across languages

Data for lexical precision are sourced from the Database of Cross-Linguistic Colexifications, third edition (CLICS3) of which was published in 2020 (Rzymski et al., 2020). This database encodes information about the meanings of words in over 2,000 languages by linking words to a common collection of concepts taken from the Concepticon (List et al., 2020). As these concepts span a diverse range of semantic domains, this dataset is well-suited to a cross-domain comparison of lexical precision based on colexification patterns.

The database is compiled from 30 sub-datasets, with each glossing words in a number of languages. Glosses in the contributing datasets are then used in CLICS3 to associate each word in each language with the Concepticon concepts it lexifies. This association allows one to find all concepts colexified with a given concept in a given language or language family.

The CLICS3 database also provides tools for visualizing the colexification of concepts in a graph. Concepts are visualized as nodes, and nodes are connected by an edge if there is a word in some dataset that lexicalizes both concepts. An example of a portion of such a graph for Russian, Swahili, and Hungarian is given in Figure 1.

Mappings from concepts to words

The resources used to extract information about lexical precision in semantic domains were Concepticon (List et al., 2020) and CLICS3 (Rzymski et al., 2020). The former is a list of concepts intended to be cross-linguistically comparable, each of which is linked to related glosses from various dictionaries and databases. It also includes some information on inter-concept relationships like "narrower". The latter provides information on how each of these concepts is lexicalized in a number of world languages. Consequently, the semantic domains to be studied were chosen as subsets of the concepts in the Concepticon (described in the next section).

In order to find the frequencies of words from each semantic domain, it was necessary to map each of the chosen concepts to a list of words in each included language. Concepticon provides links from a subset of its concepts to two lexical databases: Babelnet (Navigli & Ponzetto, 2012) and OmegaWiki (Omegawiki, 2019). Each of these databases is organized by concept, and for each concept, lists words in a variety of languages. These databases were accessed automatically through their web interfaces to find words associated to Concepticon concepts. Note that frequencies only take into account concepts for which a word was found in Babelnet or Omegawiki.

Although CLICS3 (Rzymski et al., 2020) provides mappings between concepts and words, said words are recorded in a notation specific to CLICS3 which might or might not correspond to the actual word forms in a given language. Since this notation does not reliably align with the standard orthography, it is not useful for finding words to search for in corpora, and BabelNet and OmegaWiki were used instead.

Text corpora for estimating communicative need

To test the generality of our hypothesis across languages, we consider seven languages in Table 1 which have accessible text corpora. We consider the following constraints in selecting text corpora and thereby languages. First, we used only news corpora (or news subcorpora of larger corpora) to reduce any effect of differences in genre on the relative frequencies of semantic domains. Second, in order to have a sufficiently large sample from each domain, we used only corpora of at least ten million words, total. Third, we used only corpora with lemmatization for all tokens, to facilitate looking up all mentions of a particular lexical item more easily and reliably. Forth, we ignored any corpora with high numbers of translated texts, as translated texts would not accurately reflect need probabilities in the translation target language. For this reason, we eliminated parallel corpora and web-crawl corpora. Finally, we used only corpora which were publicly available for free. Within these constraints, we sought corpora representing languages from varied families. This left us with the seven languages and corpora described in Table 1.

In order to estimate the communicative need of each semantic domain in each language, we consider words associated to each concept in each language, then find their usage frequencies in the corresponding text corpora, as proportions of total corpus size. The frequency for each semantic domain in each language is calculated as the total frequency of all words for concepts in that domain in that language. We chose to approximate communicative need of a semantic domain this way due to its straightforward procedure and generalizability across domains and languages, but it is by no means the only or optimal way and has limitations. For instance, usage frequency of a polysemous word might represent its need in several different domains and hence overestimates the need for a single domain to a certain extent.

Choices of domains and concepts

We choose semantic domains based on the subsets of the concepts in the Concepticon (List et al., 2020), because these are the common glosses used to compile the CLICS3 dataset.

Choice of domains. We choose semantic domains based on three criteria: (1) previous mention in semantic typology literature; (2) reasonable assumption of noun-based concepts and discreteness of the underlying conceptual space; (3) reasonable assumption of a shared underlying conceptual space.

First, we searched a broad collection of semantic typology papers on diverse semantic domains including field manuals, computational studies, experimental studies, and theoretical work. Domains were only included if at least two of these papers mentioned them. We acknowledge that our search and set of domains are by no means exhaustive.

Next, we eliminated those domains which have a clearly continuous, rather than discrete, conceptual space. We expected these domains to be poorly modeled by the discrete concepts in the CLICS3 database. For example, the domain of Colour was eliminated at this stage.

Finally, we eliminated those domains where the meanings of the concepts themselves were expected to vary across languages due to cultural variation. This assessment was aided by studies such as (Rabinovich, Xu, & Stevenson, 2020; Thompson, Roberts, & Lupyan, 2020; Majid, Jordan, & Dunn, 2015) which compare the cross-linguistic semantic alignment of different domains. For example, the domain of Social Relations was eliminated at this stage. It is likely that the meaning of many social relations, such as FRIEND and GUEST, varies across cultures, and indeed Thompson et al. (2020) find a low degree of cross-linguistic semantic alignment of words in this domain. The domains that remained after this process were Animals, Body, Clothing, Emotion, Kinship, Number, Physical Geography, Plants, Speech, and Time.

Choice of concepts. We used only concepts whose ontological category in Concepticon was marked as "Person/Thing". These concepts are most likely to be lexicalized by nouns, and by restricting our attention to one grammatical class, we reduce the effect of each language's syntax on the measured frequencies of each domain. Moreover, each semantic domain underwent a further selection criterion, informed by the relevant literature, and listed below:

Animals. Types of creatures from the kingdom Animalia. No parts of animals were included. Types need not correspond to scientifically-delineated taxa.

Table 1: The seven languages used for the analysis, and the corpora from which need probabilities are calculated.

Language	Family	Corpus
Albanian	Indo-European	Albanian National Corpus (Morozova, Rusakov, & Arkhangelskiy, n.d.)
Basque	Basque	Corpus of Contemporary Basque (Sarasola, Salaburu, & Landa, 2021)
English	Indo-European	British National Corpus (Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium, 2007)
Hungarian	Uralic	Hungarian National Corpus (Oravecz, Váradi, & Sass, 2014)
Russian	Indo-European	Russian National Corpus (RNC Developers, 2016)
Swahili	Atlantic-Congo	Helsinki Corpus of Swahili 2.0 (Hurskainen & Department of World Cultures, University of Helsinki, 2016)
Turkish	Turkic	TS Columns Corpus (Sezer, 2017)

Body. From Brown (1976): "[P]art of [the human body] and ... described as 'possessed by' [the human body]." Includes only things with specifically delineated locations on the body: no liquids or other substances which move around were included. Only parts of *living* bodies were considered.

Clothing. From Roach-Higgins and Eicher (1992): "[E]nclosures that cover the body... generally omit[ting] body modifications."

Emotion. From Jackson et al. (2019): "[A] mental state that [can] be felt."

Kinship. Any kin relationship that Murdock (1970) addresses, as well as those strictly between those and the ego on a family tree structured like those in (Kemp & Regier, 2012). Excludes concepts that depend on the speaker (e.g. DAUGHTER (OF MALE EGO), defined as "A daughter, as referred to by her father.").

Number. A collection of one or more positive integers. Includes vague quantities like MANY.

Physical Geography. From Mark and Turk (2017): "[T]he natural landscape, especially landforms and water bodies... it does include toponyms and cultural and spiritual associations with landscape... and vegetation assemblages." Excludes man-made constructions.

Plants. Types of living things from the kingdom Plantae. No parts of plants were included. Types need not correspond to scientifically-delineated taxa.

Speech. From Rhodes (1986): Concepts that "refer to instances of oral communication".

Time. From Evans (2013): Concepts that "relate to experiences such as duration, simultaneity, assessment of a temporal 'point', the experience of 'now' ". Concepts meaning subsets of the timeline (e.g. JANUARY and SOMETIMES) were also included.

Using these definitions, we hand-coded each concept from the Concepticon, based on its Concepticon definition, as belonging to one or none of the above ten domains. Note that the above-defined domains do not line up perfectly with the "Semantic_Field" associated to each concept in Concepticon.

We eliminated the data from any domain-language pairs in which ten or fewer concepts had an associated wordform in CLICS3, so the entire Number domain was eliminated, Emotion and Speech domains were not used in Albanian, Swahili, or Turkish, and the Clothing domain was not used in Turkish. This left 56 domain-language pairs to be employed in the analysis. We also focused on analyzing concepts lexicalized by single word forms in a given language in CLICS3 database. We discarded word forms if they contain spaces, since words with spaces were usually multiple-word descriptions such as *paternal grandfather* which had been entered into a component dataset as a single vocabulary item.

Computational methodology

In order to test our hypothesis, we need to define a formal notion of lexical precision. We formulate lexical precision as the opposite of lexical ambiguity, and we define the lexical ambiguity of a domain by estimating the average amount of ambiguity it contains in its colexification patterns. Intuitively, a domain that shows a high degree of colexification will yield high lexical ambiguity, and the reverse holds for domains with high lexical precision. We describe two alternative methods, "expected lexical ambiguity" vs. "edge density", and suggest that the former is a more appropriate measure which we used for our analysis.

Expected lexical ambiguity

To quantify lexical ambiguity in a way that takes into account semantic breadth of words (in general) within a domain, we calculate the expected or average ambiguity of concepts within that domain.

In our terminology, the *ambiguity* of a concept is determined by identifying the narrowest word that labels that concept, and counting how many concepts, including the target concept, are labelled by that word. Formally, if a speaker intends to express a certain concept C, and they use the most precise possible word to do so, then the ambiguity of C measures the number of possible interpretations of that word by the listener. Thus for a concept C in a domain D, we have:

$$\text{Ambiguity}(C) = \min_{\text{words } w} n\left(\{C' \in D | w \text{ lexicalizes } C'\}\right)$$

Here, w ranges over all words of the language under consideration, and $n()$ measures the number of elements in a set.

The expected lexical ambiguity of a domain is then the average of $\text{Ambiguity}(C)$ over all concepts C in that domain. As an example, consider the few concepts in Figure 1 as a mini-domain, and consider the lexical ambiguity of this domain in Swahili (Figure 1b). The ambiguity of the concept VALLEY would be 1, since there is a word, *bonde*, labelling only this concept from the mini-domain. The ambiguity of REEF is 2, since the most precise word for it, *mwamba*, lexicalizes two concepts from the mini-domain. Similarly, the ambiguity of PRECIPICE is 2, resulting in the average ambiguity of this mini-domain to be 1.67. In contrast, the average ambiguity of this mini-domain in Hungarian (Figure 1c) is 1, because every concept has a maximally precise word lexicalizing it within that domain. The expected lexical ambiguity, therefore, captures the precision in vocabulary available to speakers, and we explain how this measure is better than alternative measures such as edge density described next.

Alternative measure based on edge density

We have also considered an alternative measure of lexical ambiguity by quantifying colexification in a domain using the *edge density* of the relevant subgraph of the CLICS3 colexification graph for a given language. The edge density of a graph is defined as the number of actual edges divided by the number of potential edges (pairs of distinct vertices). That is, if $n(E)$ is the number of edges of the graph, and $n(V)$ is the number of vertices, then the edge density is as follows:

$$\text{Edge Density} = n(E)\binom{n(V)}{2}^{-1} = \frac{2n(E)}{n(V)\,(n(V)-1)}.$$

This measure captures some aspects of our intuition for the "amount of ambiguity" in a domain. For example, in the subgraph corresponding to the concepts VALLEY, PRECIPICE, and REEF (Figure 1), there is clearly more ambiguity in Swahili than in Russian, since Swahili has a word colexifying PRECIPICE and REEF, which Russian does not. Correspondingly, the Russian subgraph has edge density $0/3 = 0$, while the Swahili subgraph has edge density $1/3 = 0.333...$. However, this alternative measure fails to distinguish certain cases which are different in important ways. Using edge density, we regard domains as highly ambiguous if speakers have the option of using a word that covers many concepts. This does not capture the increased precision available if speakers also have the option of a semantically narrower word. For instance, the edge density of the Hungarian graph in Figure 1c is identical to that of the Swahili graph in Figure 1b, which ignores the fact that the words *zátony* and *szikla* are available to Hungarian speakers wishing to speak more precisely about these concepts, while (according to CLICS3 data) no such single words are available to Swahili speakers. Thus, the Swahili vocabulary for these concepts is more ambiguous than the Hungarian vocabulary. However, the proposed measure of expected lexical ambiguity which we described is sensitive to this fact.

Adjustment for broader or narrower concepts

Due to the way databases were assembled into CLICS3, an additional adjustment step is necessary for computing expected lexical ambiguity for each domain. In assembling CLICS3, words from each contributing dataset were connected to concepts according to the glosses in that dataset. This means that some words were recorded as lexicalizing one broad concept, even when they also lexicalize many narrower concepts. For example, the English word *grandfather* was mapped to the concept GRANDFATHER, which ignores the fact that the concept GRANDFATHER also colexifies the two concepts PATERNAL GRANDFATHER and MATERNAL GRANDFATHER. The type of underspecification in the English word *grandfather* is exactly the type we wished to capture with our measure of ambiguity in a domain. As such, it is necessary to correct for this issue, to avoid undercounting the ambiguity of a domain.

To do so, we use the broader-narrower relations encoded in Concepticon (List et al., 2020). Specifically, we followed the general rule that if a concept from a domain has any narrower concepts from the same domain, then any words lexicalizing the broader concept are considered to colexify all of the narrower concepts, and the broader concept is ignored. For example, instead of lexicalizing GRANDFATHER, the English word *grandfather* was considered to colexify PATERNAL GRANDFATHER and MATERNAL GRANDFATHER. For cases in which the narrower concepts did not cover all logically possible cases of the broader concept, an additional "other" concept was added. So, for example, in Concepticon, the concept MONKEY is broader than the concepts SPIDER MONKEY, HOWLER MONKEY, and CEBUS MONKEY. The English word *monkey* is recorded in CLICS3 as lexicalizing MONKEY. For our analysis, we instead regarded this word as colexfying the four concepts: SPIDER MONKEY, HOWLER MONKEY, CEBUS MONKEY, and OTHER MONKEY. In cases where an "other" concept was logically impossible[1], such as in the case of splitting GRANDFATHER into PATERNAL GRANDFATHER and MATERNAL GRANDFATHER, no "other" concept was added. The average ambiguity for each domain was then computed using this adjusted colexification dataset.

Results

To evaluate our main hypothesis, we performed a regression-based analysis between (1) the expected lexical ambiguity of each domain, and (2) the communicative need of each domain (operationalized as the logarithm of the total domain usage

[1] The judgment of "logically impossible" included an assumption of binary gender and sex in humans and mammals, and of heterosexual spousal relationships.

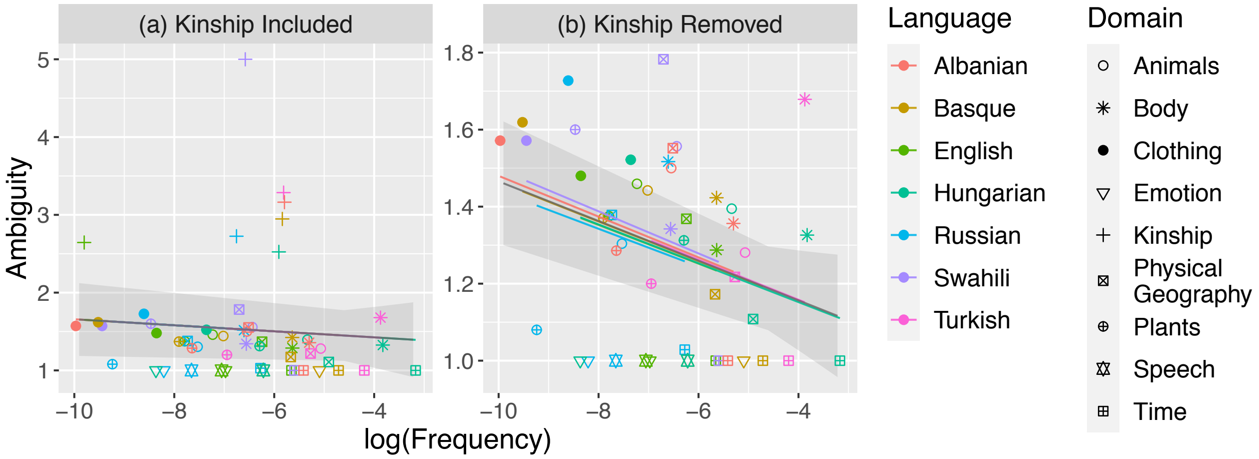


Figure 2: A summary of results for comparing the expected lexical ambiguity and communicative need (word usage frequency of a domain) for each domain in each language. Domains with high needs or frequencies are expected to have low lexical ambiguity. A line fit under a linear mixed-effects model on (a) all points, or on (b) the data with the Kinship domain removed, is shown in grey, with a 95% confidence interval. Lines fit to data of each language are plotted in color, matching those in (a).

frequency), for 56 domain-language pairs we examined. We used a linear mixed-effects model to account for these data, which includes a random effect of language on slope. Figure 2 summarizes the results.

In Figure 2a, a linear mixed-effects model was run on all the data points, expressing the ambiguity of a domain in terms of its log frequency. These initial results were not significant: although the slope of the model was estimated at -0.4 aligning with the direction that we expect from our hypothesis, the overall fit was statistically insignificant $p = 0.548$. The effect of language on slope was estimated to be exactly 0. Due to the small number of data points, the initial analysis is highly susceptible to interference from outliers. This can be seen in the difference between Figure 2a and Figure 2b, where the latter is based on a similar analysis except with the outlier datapoints removed. In order to test for interference from outliers, a robust linear mixed-effects model from the package robustlmm in R (Koller, 2016) was applied to the data. The model assigns a “robustness” from 0 to 1 to each point, with low-robustness points considered outliers. The model assigned robustness less than 0.5 to all and only points from the Kinship domain. Therefore, we removed the Kinship domain as the outlier and re-ran the mixed-effects model (Figure 2b). This time, the slope of the model was significantly negative at -0.05 with $p = 0.016$. The standard deviation of the random effect of language was estimated at 0.0051. So, the effect of the outlier Kinship domain obscured the general pattern observed in the other domains in the initial analysis: namely, more frequent domains tend to be more precise, and this effect varies in magnitude across languages.

Discussion and conclusion

We presented a computational study on the relationship between communicative need and lexical precision in semantic domains across languages. We found initial support for the domain-level efficiency hypothesis that domains with higher communicative need tend to have higher lexical precision. Our findings extend existing work that explores similar ideas in more restricted settings concerning pairs of concepts (e.g., Regier et al., 2016; Karjus et al., 2021). Our results are based on a simple frequency-based measure of communicative need and a limited set of semantic domains and languages, and are therefore subject to issues such as biased estimation and outliers. As a result, the statistical power in our analysis may be limited. However, to our knowledge this is one of the first studies on examining the role of communicative need in shaping lexical precision across a diverse set of domains.

Our results provide some support for the view that communicative need modulates the precision of vocabulary across domains, but these results can be consolidated with more domains and languages. It is possible that communicative efficiency does shape the cross-domain variation in ambiguity, but that for Kinship, other factors might influence communicative efficiency and cause elevated ambiguity. For example, Rácz, Passmore, Sheard, and Jordan (2019) suggest that social changes may influence the evolution of kinship vocabularies in ways that differ from the core vocabulary. Alternatively, it could be that kinship is a special domain from a communicative view: in cultures such as English, one does not typically refer to a relative directly by the kin term but rather by name, making kin terms a subsidiary naming system. Future work can explore this issue and broaden the analysis to understand the general role of communicative need in lexical precision across languages.

Code

Code and data for this paper are available on GitHub: https://github.com/laurestine/needandambiguity/

Acknowledgments

The authors thank Myrto Grigoroglou, Noga Zaslavsky, Charles Kemp, and Terry Regier for valuable feedback.

This work was supported by a NSERC Discovery Grant RGPIN-2018-05872, a SSHRC Insight Grant #435190272, and an Ontario ERA Award to YX.

References

Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. (2007). *The British national corpus, version 3 (BNC XML edition).* Retrieved from http://www.natcorp.ox.ac.uk/ ([Online; accessed 7-August-2021])

Brown, C. H. (1976). General principles of human anatomical partonomy and speculations on the growth of partonomic nomenclature. *American Ethnologist*, *3*(3), 400–424.

Carr, J. W., Smith, K., Culbertson, J., & Kirby, S. (2020). Simplicity and informativeness in semantic category systems. *Cognition*, *202*, 104289.

Conway, B. R., Ratnasingam, S., Jara-Ettinger, J., Futrell, R., & Gibson, E. (2020). Communication efficiency of color naming across languages provides a new framework for the evolution of color terms. *Cognition*, *195*, 104086.

Evans, V. (2013). *Language and time: A cognitive linguistics approach*. Cambridge University Press.

Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, *109*(44), 17897–17902.

Floyd, S., Dalawella, K., Goldberg, A., Lew-Williams, C., & Griffiths, T. (2021, 07). Modeling rules and similarity in colexification. In T. Fitch, C. Lamm, H. Leder, & K. Tessmar-Raible (Eds.), *Proceedings of the 43rd annual meeting of the cognitive science society.* Cognitive Science Society.

François, A. (2008, 01). Semantic maps and the typology of colexification: Intertwining polysemous networks across languages. In M. Vanhove (Ed.), *From polysemy to semantic change: Towards a typology of lexical semantic associations* (p. 163-215). Amsterdam: John Benjamins.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, *23*(5), 389-407.

Hawkins, R. X., Franke, M., Smith, K., & Goodman, N. D. (2018). Emerging abstractions: Lexical conventions are shaped by communicative context. In *Proceedings of the 40th annual meeting of the cognitive science society.*

Hurskainen, A., & Department of World Cultures, University of Helsinki. (2016). *Helsinki Corpus of Swahili 2.0 Annotated Version* [text corpus]. Kielipankki. Retrieved from http://urn.fi/urn:nbn:fi:lb-2016011301

Jackson, J. C., Watts, J., Henry, T. R., List, J.-M., Forkel, R., Mucha, P. J., ... Lindquist, K. A. (2019). Emotion semantics show both cultural variation and universal structure. *Science*, *366*(6472), 1517–1522.

Kanwal, J., Smith, K., Culbertson, J., & Kirby, S. (2017). Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, *165*, 45-52.

Karjus, A., Blythe, R. A., Kirby, S., Wang, T., & Smith, K. (2021). Conceptual similarity and communicative need shape colexification: An experimental study. *Cognitive Science*, *45*(9), e13035. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.13035 doi: https://doi.org/10.1111/cogs.13035

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054.

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*(1), 109-128.

Khetarpal, N., Neveu, G., Majid, A., Michael, L., & Regier, T. (2013, 08). Spatial terms across languages support near-optimal communication: Evidence from Peruvian Amazonia, and computational analyses. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual meeting of the cognitive science society.* Cognitive Science Society.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87-102.

Koller, M. (2016). robustlmm: An R package for robust estimation of linear mixed-effects models. *Journal of Statistical Software*, *75*(6), 1–24. doi: 10.18637/jss.v075.i06

Kågebäck, M., Carlsson, E., Dubhashi, D., & Sayeed, A. (2020, 07). A reinforcement-learning approach to efficient communication. *PLOS ONE*, *15*(7), 1-26.

List, J. M., et al. (Eds.). (2020). *Concepticon 2.4.0*. Jena: Max Planck Institute for the Science of Human History. Retrieved from https://concepticon.clld.org/

Majid, A., Jordan, F., & Dunn, M. (2015). Semantic systems in closely related languages. *Language Sciences*, *49*, 1-18. (Semantic systems in closely related languages)

Mark, D., & Turk, A. (2017, 03). Ethnophysiography.. doi: 10.1002/9781118786352.wbieg0349

Morozova, M., Rusakov, A., & Arkhangelskiy, T. (n.d.). *Albanian national corpus.* Retrieved from albanian.web-corpora.net ([Online; accessed 28-July-2021])

Murdock, G. P. (1970). Kin term patterns and their distribution. *Ethnology*, *9*(2), 165–208.

Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, *193*, 217-250.

Omegawiki. (2019). *Meta: Main page — Omegawiki, a dictionary in all languages.* Retrieved from http://www.omegawiki.org/Meta:Main_Page ([Online; accessed

28-July-2021])

Oravecz, C., Váradi, T., & Sass, B. (2014, 01). The Hungarian gigaword corpus..

Rabinovich, E., Xu, Y., & Stevenson, S. (2020). The typology of polysemy: A multilingual distributional framework. In *Proceedings of the 42nd annual meeting of the cognitive science society.*

Rácz, P., Passmore, S., Sheard, C., & Jordan, F. M. (2019). Usage frequency and lexical class determine the evolution of kinship terms in Indo-European. *Royal Society Open Science*, *6*(10), 191385.

Regier, T., Carstensen, A., & Kemp, C. (2016, 04). Languages support efficient communication about the environment: Words for snow revisited. *PLOS ONE*, *11*(4), 1-17.

Rhodes, R. (1986). The semantics of the ojibwa verbs of speaking. *International Journal of American Linguistics*, *52*(1), 1–19.

RNC Developers. (2016). *Russian national corpus (RNC)* [text corpus]. Retrieved from `https://ruscorpora.ru/new/en/index.html` ([Online; accessed 28-July-2021])

Roach-Higgins, M. E., & Eicher, J. B. (1992). Dress and identity. *Clothing and Textiles Research Journal*, *10*(4), 1-8.

Rzymski, C., Tresoldi, T., Greenhill, S. J., Wu, M.-S., Schweikhard, N. E., Koptjevskaja-Tamm, M., ... List, J.-M. (2020, 01). The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific Data*, *7*(13). Retrieved from `https://doi.org/10.1038/s41597-019-0341-x` (Database accessible at `https://clics.clld.org/`) doi: 10.1038/s41597-019-0341-x

Sarasola, I., Salaburu, P., & Landa, J. (2021, February 5). *Egungo testuen corpusa (ETC).* Retrieved from `https://www.ehu.eus/etc/` ([Online; accessed 28-July-2021])

Sezer, T. (2017). TS corpus project: An online Turkish dictionary and TS DIY corpus. *European Journal of Language and Literature*, *9*(1), 18-24.

Thompson, B., Roberts, S. G., & Lupyan, G. (2020, 10). Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour*, *4*, 1029–1038.

Xu, Y., Duong, K., Malt, B. C., Jiang, S., & Srinivasan, M. (2020). Conceptual relations predict colexification across languages. *Cognition*, *201*, 104280.

Xu, Y., Liu, E., & Regier, T. (2020, 08). Numeral systems across languages support efficient communication: From approximate numerosity to recursion. *Open Mind*, *4*, 57-70.

Zaslavsky, N., Kemp, C., Tishby, N., & Regier, T. (2020). Communicative need in colour naming. *Cognitive Neuropsychology*, *37*(5-6), 312-324.