The Typology of Polysemy: A Multilingual Distributional Framework

Ella Rabinovich[†] **Yang Xu**[†*] **Suzanne Stevenson**[†]

[†]Department of Computer Science, [*]Cognitive Science Program
University of Toronto
{ella,yangxu,suzanne}@cs.toronto.edu

Abstract

Lexical semantic typology has identified important cross-linguistic generalizations about the variation and commonalities in polysemy patterns—how languages package up meanings into words. Recent computational research has enabled investigation of lexical semantics at a much larger scale, but little work has explored lexical typology across semantic domains, nor the factors that influence cross-linguistic similarities. We present a novel computational framework that quantifies *semantic affinity*, the cross-linguistic similarity of lexical semantics for a concept. Our approach defines a common multilingual semantic space that enables a direct comparison of the lexical expression of concepts across languages. We validate our framework against empirical findings on lexical semantic typology at both the concept and domain levels. Our results reveal an intricate interaction between semantic domains and extra-linguistic factors, beyond language phylogeny, that co-shape the typology of polysemy across languages.

Keywords: semantic typology, cross-linguistic similarity, word meaning, distributional semantics, multilingual word embeddings

A central issue in cognitive science is the nature of the mental mapping between language and the world. One oft-studied question is how and why languages vary in the way they use words to partition semantic space (e.g., Berlin & Kay, 1969; Levinson & Meira, 2003). Polysemy—the use of a single word form to express multiple related senses—is a fundamental property of language that exemplifies this variation. Figure 1 shows how word forms across languages may differ in the sets of senses they cover. Despite this variation, there is also much cross-linguistic commonality in word meanings, as seen in the overlap of sense expression in Figure 1. How much do languages vary in their lexical semantics, and what contributes to the observed cross-linguistic patterns of polysemy? Here we present a principled and large-scale computational approach to these questions.

Work in typology—studies of the constrained variation exhibited by languages—has identified important cross-linguistic generalizations regarding polysemy patterns across many semantic domains. For example, some research has focused on the primitives that underlie cross-linguistic lexical categorization (e.g., Berlin & Kay, 1969; Levinson & Meira, 2003), while other work has studied the degree of universality of polysemy patterns (e.g., Majid, Jordan, & Dunn, 2015; Youn et al., 2016). However, such studies have been restricted in scope due to reliance on manual methods. To find robust answers to the above questions on semantic typology, automatic methods are required to enable larger scale study.

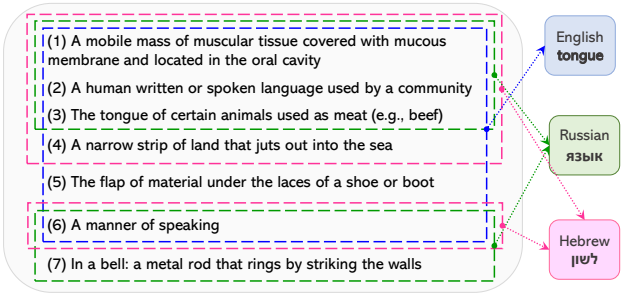

Figure 1: A partial list of meanings given in Babelnet for the English word "tongue", as well as for the corresponding Russian and Hebrew word forms.

Computational research has proposed various methodologies to explore lexical semantic structure at a more comprehensive scale. Previous work exploiting distributional representations has studied how language-pair semantic similarity is influenced by various factors, including phylogeny (e.g., (Thompson, Roberts, & Lupyan, 2018; Beinborn & Choenni, 2019)), geography (Eger, Hoenen, & Mehler, 2016), culture (Thompson et al., 2018), and conceptual structure (Xu, Duong, Malt, Jiang, & Srinivasan, in press). To the best of our knowledge, the only study considering lexical variation at the level of semantic domain is that by Thompson et al. (2018). That study used monolingual word embeddings for quantifying cross-linguistic semantic alignment in an inherently multilingual setting. Importantly, previous work has typically focused on descriptive analyses that are not evaluated against the empirical generalizations reported in the literature.

We propose a novel framework[1] for quantifying lexical semantic variation that addresses these limitations in two respects. First, we develop a measure of *semantic affinity* that assesses the degree of semantic similarity of the corresponding word forms for a concept across many languages. We take an alternative approach to existing work in which we construct a *common multilingual semantic space* that enables a direct comparison of the lexical expression of concepts across multiple languages. Second, we evaluate our approach against known findings in the typological literature to assess the validity of our measure.

[1]All code and data are available at https://github.com/ellarabi/semantic-affinity

Background on Lexical Semantic Variation

We focus on some key findings regarding lexical variation and factors that influence it, at the level of individual concepts, domains, and languages. We summarize these empirical findings in Table 1, which we will assess our framework against.

Individual Concepts. Semantic change is an important source of polysemy, and factors that influence that process may also influence the degree of semantic affinity of concepts. A number of studies have shown that the rate of semantic change is correlated with the psycholinguistc factors of frequency and degree of polysemy (estimated by number of senses), and minimally correlated with word length (a proxy of frequency) when frequency and polysemy are both controlled for (e.g., Hamilton, Leskovec, & Jurafsky, 2016). Pagel, Atkinson, and Meade (2007) also found that numbers (e.g., "two") are slowest to change among the grammatical categories, which follow a specified order (Table 1).

Semantic Domains. Recently, much research on lexical semantic typology has studied cross-linguistic universals and principles of variation in patterns of polysemy (e.g., Berlin & Kay, 1969; Levinson & Meira, 2003; Majid et al., 2015; Youn et al., 2016). Two studies in particular enable us to assess the relative level of semantic affinity across different semantic domains. First, Majid et al. (2015), using naming tasks to elicit lexical data, manually determine an ordering of the degree of semantic variation among four conceptual domains across 12 Germanic languages; see Table 1. Second, Youn et al. (2016), using manual translation across 81 languages, found that a set of 23 basic concepts pertaining to the physical environment exhibits "universal tendencies" in lexical semantics—i.e., has a high degree of cross-linguistic similarity. In particular, they show that this similarity is unaccounted for by phylogeny, geography, or climate (with one exception, which we return to in our results). Interestingly, Regier, Carstensen, and Kemp (2016) did find an effect of environmental factors on the cross-linguistic lexicalization of "snow" and "ice" (but this subdomain is too small to assess).

Language-Level Influences. Studies quantifying similarity between pairs of languages have exploited distributional properties extracted from monolingual (e.g., Beinborn & Choenni, 2019; Thompson et al., 2018) or bilingual (Eger et al., 2016) semantic spaces. The findings by and large highlight the correlation between languages' semantic and *phylogenetic* similarity (Beinborn & Choenni, 2019). Correlations of geographical (Eger et al., 2016) and cultural (Thompson et al., 2018) factors with cross-linguistic semantic similarity have been shown. However, an analysis of their influence across various semantic domains, and evaluation against empirical observations, have been lacking.

Our Approach. Our work is closely related to that by Thompson et al. (2018), who presented a large-scale study of cross-linguistic semantic alignment at the level of the domain. That study used monolingual semantic spaces to quantify cross-linguistic semantic alignment, where the similarity between words representing a concept in two languages was estimated indirectly through the proximity of these words to their (partial) neighbourhood in individual spaces. Our work explores an alternative approach based on *semantic affinity*, which differs in that we: (1) quantify cross-linguistic semantic similarity in a direct and unmediated way, by constructing a common multilingual semantic space shared across languages of interest; (2) evaluate this framework against empirical findings in the literature, a critical aspect that was not explored in the previous work; and (3) leverage this framework to perform analysis of factors—both linguistic and extra-linguistic—that influence semantic affinity of a concept, at the levels of a single concept and a domain.

Datasets

Translation Sets. Measuring cross-linguistic semantic affinity of a concept requires a set of words representing that concept in various languages. We used NorthEuralex (Dellert et al., 2017), a large lexicostatistical database providing accurate (manual) translations of over 1000 basic concepts in 107 languages from 20 distinct language families of Northern Eurasia, including over 30 Indo-European languages. Each concept, represented by a corresponding German term, is annotated for part-of-speech (POS), and links to a set of word forms representing this concept in other languages. This yields a set of common concepts, spanning multiple domains, and including accurate translations of the same concept into words across multiple languages. Since these words naturally have various additional meanings across the languages, reflecting various patterns of polysemy, this introduces a natural testbed for our analysis. Despite limitations due to known quality control issues,[2] this is one of the most comprehensive multilingual datasets, suitable for this study.

Cross-Linguistic Polysemy Data. BabelNet (Navigli & Ponzetto, n.d.) is a very large multilingual semantic network in which each node represents a language-independent meaning, to which words across the represented languages can link. For example, as illustrated in Figure 1, the node for the meaning "A human written or spoken language used by a community" will be linked from the English word "tongue", as well as the corresponding words in Russian and Hebrew. Crucially, as seen in the figure, our target words that represent the same NorthEuralex concept in different languages may cluster different (sub)sets of meanings – sharing a common set of meanings, but deviating in language-specific ones. For each of our concepts, we document the total number of *distinct* meanings associated with it cross-linguistically, as accessed through the words representing this concept in the set of languages used in this work. We restricted the list of concepts to those supported in Babelnet by at least 30 of

[2] See note at `http://northeuralex.org/`.

Level of analysis	Summary of empirical findings from the literature
Individual concepts	Rate of semantic change correlates with frequency (-), polysemy (+), word length (≈0) Lexical evolution rate: number<pro.<adv.<noun<verb<adj.<conj.<prep.
Semantic domains	Semantic variation: Color, Body Parts<Containers<Spatial Relations Universal tendencies in lexicalizing basic concepts of the physical environment
Language-level influences	Language phylogeny correlates with semantic similarity across languages Environmental factors (geography/climate) influence lexical semantic typology

Table 1: Condensed summary of recent findings on lexical semantic typology to which we compare our approach.

the 35 languages considered in this study (i.e., at least 30 of languages have a corresponding word-form entity in the database). This results in 697 concepts across many domains.

Computational Framework

Our goal is to measure the degree of semantic similarity of the corresponding words for a concept across many languages. We adopt a distributional semantics approach given the success of such models in capturing subtleties of word meaning (e.g., Hollis & Westbury, 2016; Pereira, Gershman, Ritter, & Botvinick, 2016). We construct *multilingual* common semantic spaces that enable the projection of words from multiple languages into a shared space (e.g., Conneau, Lample, Ranzato, Denoyer, & Jégou, 2017). Specifically, words in different languages that have roughly the same meaning are brought close to each other within a single vector space. For a given concept, we operationalize its semantic affinity across languages by the degree of similarity of the corresponding words' representations in the common semantic space. This notion of affinity can be extended to a semantic domain (a collection of concepts) and to languages (across all concepts).

Building a Multilingual Semantic Space. We use the Facebook MUSE framework (Conneau et al., 2017), shown to obtain good results on many tasks (e.g., Artetxe & Schwenk, 2019; Beinborn & Choenni, 2019), for construction of a multilingual semantic space. The model uses a set of automatically extracted bilingual dictionaries between pairs of languages to project monolingual word representations in two languages onto a common space. It does so while optimizing the mutual proximity of representations of an automatically extracted set of translation equivalents (words referring to the same entity; e.g., English "apple", French "pomme"). Using English as a pivot language, the procedure can then be scaled to any number of languages L, assuming an English–L bilingual dictionary, and ultimately resulting in a common massively multilingual semantic space. Further details on this procedure can be found in Appendix A.

For building our multilingual space, we use the set of 35 geographically-diverse languages supported by NorthEuralex, Babelnet and MUSE bilingual dictionaries, and the corresponding fastText monolingual embeddings (Grave, Bojanowski, Gupta, Joulin, & Mikolov, 2018). In training and validation, we excluded the entire set of words representing our target concepts from the list of translation equivalents whose proximity is optimized by MUSE in creating the common embedding space. Figure 2 illustrates that different concepts can have differing degrees of cross-linguistic similarity in the resulting common semantic space.

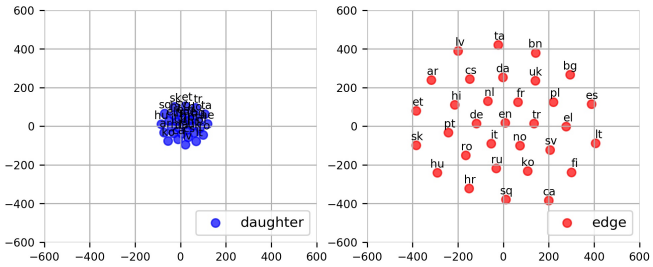

Figure 2: t-SNE projections of multilingual embeddings, corresponding to English terms "daughter" and "edge".

Quantifying Semantic Affinity. The semantic affinity of a concept w.r.t. a set of languages amounts to the mutual proximity of embeddings representing the concept across various languages; that is, the "tighter" the cluster of embeddings, the more (cross-linguistically) similar the concept is (cf. Figure 2). Formally, given a concept c and a set of N word forms representing c across a set of languages $\mathcal{L}=\{L_1, L_2, ..., L_N\}$, we denote its corresponding 300-dimensional vector representations by $V^c = \{v_1^c, v_2^c, ..., v_N^c\}$. Using cosine similarity as our similarity metric, we compute the centroid of V^c by calculating the average of its constituents. This procedure results in a vector in the direction of the cluster centroid:

$$\text{Cent}(V^c) = \frac{1}{N}\sum_i v_i^c, \quad i\in[1..N], \|v_i^c\|=1 \qquad (1)$$

We then estimate cross-linguistic semantic affinity of c with respect to $\mathcal{L}$ by computing its *cluster density*, specifically, we average over individual words' cosine similarities to the (virtual) cluster centroid in Equation 1 ($i \in [1..N]$):

$$\text{SemAff}(V^c) = \frac{1}{N}\sum_i \cos(v_i^c, \text{Cent}(V^c)) \qquad (2)$$

Intuitively, semantic affinity mirrors the extent of meaning similarity of a concept as expressed across a set of languages. For example, as expected from Figure 2, we find higher semantic affinity for the concept corresponding to "daughter" (0.766) than for that corresponding to "edge" (0.572).

Results on Concepts and Domains

We first evaluate how well our measure of cross-linguistic semantic affinity matches empirical findings at the level of individual concepts and semantic domains (Table 1).

Semantic Affinity of Concepts

We hypothesize that factors that play a role in lexical semantic change (within a language) may also influence the degree of semantic affinity across languages. We thus suggest the following variables as predictors of cross-linguistic affinity:

Mean Word Rank. We derive a ranked list of the top-N words in each language using frequencies recorded in wordfreq[3] For a given concept c, we then average the ranks of its corresponding words across the languages.[4]

Degree of Polysemy. We computed the total number of unique senses of the words associated with a concept across our languages (see the Datasets section for details).

Mean Word Length. We computed the average length of word forms corresponding to a concept across our languages.

We perform multiple regression analysis using the semantic affinity of concepts (SemAff, Equation 2) as the dependent variable, and the predictors above as independent variables; see Table 2. All variables together explain nearly 40% (adj. r^2=0.381) of the variance. Our results are in line with previous findings on the relation of these psycholinguistic variables to semantic change (cf. Table 1), as we expected since lexical evolution is an important source of polysemy. Mean word rank is negatively correlated with semantic affinity, implying that less frequently used concepts have lower cross-linguistic semantic affinity. As well, concepts with a higher degree of polysemy exhibit higher cross-linguistic semantic diversity. Finally, mean word length shows the weakest correlation with affinity among the variables.

predictor	coef.($\beta\times10$)	std err(β)	t-stat
coeff	6.615	0.001	445.747
mean word rank	-0.242	0.002	-13.294
degree of polysemy	-0.200	0.002	-16.037
mean word length	0.129	0.001	8.640

Table 2: Multiple regression analysis. The response variable is concept semantic affinity, a real value in the 0–1 range. $p < .001$ in all cases.

We further computed cross-linguistic semantic affinity for various POS categories; that is, by averaging SemAff over concepts that share the same POS in the NorthEuralex dataset, requiring a minimum of five concepts per tag. Table 3 (left) reports the results. Here too we find that our relative rankings replicate the relative stability over time of these categories as found by Pagel et al. (2007), with a single exception of a swap in ordering between verbs and nouns (cf. Table 1).

[3] https://pypi.org/project/wordfreq/

[4] Word frequency (which strictly correlates with rank in a language) is incomparable across different languages.

To provide a fine-grained qualitative view of our framework, we visualize semantic affinities of 10 common concepts in the domain of kinship. Figure 3 reveals an interesting symmetry in this domain. Specifically, semantic affinity is higher for kin terms that are more closely related to ego than those farther away, and this trend is symmetric between male and female kin types. Concretely, "aunt" and "uncle" (and "grandmother" and "grandfather") show the lowest semantic affinity across languages, in comparison to the closer kin relations such as children, siblings, and parents of ego. This observation is consistent with independent empirical findings suggesting that remote kin terms, e.g., "aunt" and "uncle", are most often extended to unrelated persons (Ballweg, 1969).

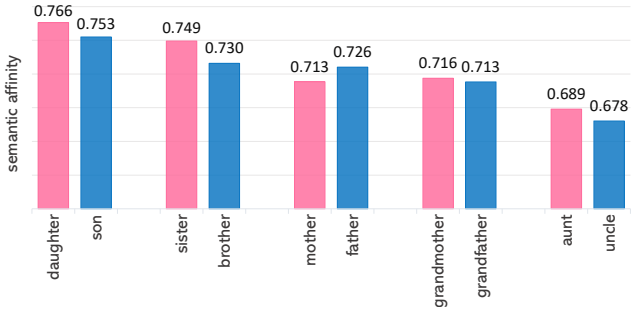

Figure 3: Semantic affinity of kinship female and male terms located by their relatedness to ego (left-to-right).

Semantic Affinity across Domains

We derive cross-linguistic semantic affinity at the level of a domain by averaging the SemAff of its individual concepts. We use semantic domains similar[5] to those used by Majid et al. (2015) and Pagel et al. (2007), as well as the set of 23 concepts analysed by Youn et al. (2016), denoting this of words set as the "Youn et al. Set" hereafter.

Table 3 (right) reports the SemAff of each domain along with its number of concepts. The results generally support fundamental observations in the literature: specifically, that words used for Numerals, Colors, and Body Parts have greater semantic affinity across languages than Spatial Relations (relational words) and Containers (artifacts). Our approach thereby provides additional empirical evidence for theoretically-motivated hypotheses on the nature of lexical semantic structure across semantic domains. However, our findings do not strictly mirror the results reported by Majid et al. (2015), e.g., spatial relations and containers are ranked in the opposite order; that, possibly due to the slightly different set of concepts in both categories, and the much larger set of languages used in our experiment. The relatively high affinity of the Youn et al. Set reflects the empirical finding by Youn et al. (2016) regarding the *universal semantic structure* of the underlying set of words—a finding that our results suggest does not systematically generalize to additional domains.

[5] Restricted by the set of words used in this study, and by limited availability of data used in previous work.

domain	count	SemAff	SD	POS	count	SemAff	SD
NUMERAL	21	0.701	0.034	Numerals	21	0.701	0.034
ADVERB	37	0.672	0.050	Youn et al. Set	22	0.683	0.031
VERB	204	0.668	0.041	Colors	9	0.675	0.028
NOUN	474	0.656	0.052	Body Parts	42	0.643	0.033
ADJECTIVE	102	0.645	0.038	Spatial Relations	8	0.621	0.043
PREPOSITION	5	0.631	0.046	Containers	9	0.611	0.030

Table 3: Cross-linguistic semantic affinity and standard deviation by part-of-speech and domain.

Results on Language-level Influences

Above we considered semantic affinity of concepts and domains; we can also calculate a measure of semantic affinity between languages (across concepts from a range of domains). As noted earlier, there is much evidence that such broad semantic affinity between languages is correlated with phylogenetic similarity, but the evidence is sparser and less clear regarding the influence of other factors, such as geography and climate. Here, we extend both these strands of work, by considering the influence of phylogeny, geography, and/or climate on a large scale sample of concepts and domains across languages. We expect genealogical similarities between languages to be predictive of their semantic affinity. Moreover, we hypothesize that geography and climate exhibit predictive power on semantic similarity above and beyond genealogy, thereby highlighting the effect of environmental factors on shaping lexical semantic systems.

Semantic Distance Between Languages

We can measure semantic affinity between two languages w.r.t. a single concept as the cosine similarity between the projection of the two words representing that concept onto our common semantic space. We then define semantic affinity between two languages across a set of concepts as the average such similarity across the individual concepts. Finally, to align with terminology in the literature on phylogenetic *distance* (as opposed to similarity), we convert this semantic affinity measure to a semantic distance by subtracting it from 1. Then, given a set of concepts $\mathcal{C}$, semantic distance (SDist) between two languages L_i and L_j w.r.t. $\mathcal{C}$ is defined as:

$$\text{SDist}(L_i, L_j) = 1 - \frac{1}{|\mathcal{C}|}\sum \cos(v_i^c, v_j^c), \quad c \in \mathcal{C} \qquad (3)$$

We limit the following analysis to 22 IE languages in our set,[6] which have well-established historical data.

Phylogenetic and Environmental Factors

We use a well-accepted tree (Gray & Atkinson, 2003) for computing phylogenetic distances between pairs of languages. We define phylogenetic distance between two languages as their (unweighted) path length in the tree. We further model two environmental factors: geographical and climate distances between languages.We model language-pairwise geographical distance as the Euclidean distance between their corresponding (longitude, latitude) coordinates in the NorthEuralex database. We model language-pairwise climate distance as the Euclidean distance between their climate vectors. These vectors are formed from *temperature* and *precipitation* data in a climate database (Kottek, Grieser, Beck, Rudolf, & Rubel, 2006). For each language, we extract this data from the region whose (longitude, latitude) coordinates are closest to those of the language in NorthEuralex.

Language-level Results and Discussion

Pairwise correlations show that the three predictor variables—phylogeny (PHY), geography (GEO), and climate (CLM)—are correlated with each other as expected: a high correlation between PHY and GEO (Pearson's r = 0.559; genetically related languages are often close in space), and between GEO and CLM (0.807; regions close in space generally have similar climates). Importantly, all three predictors also exhibit a significant association with semantic distance (SDist): 0.402 for PHY, 0.518 for GEO and 0.516 for CLI. Notably, a higher correlation of SDist is found with geographical and climate predictors than with phylogenetic distance, suggesting that these have a considerable effect on lexical semantic structure.

We next perform a multiple regression analysis to estimate the relative contribution of each of the three factors to predicting language-pair semantic distance. The three independent variables together explain 30% of SDist variance (adj. r^2=0.301); Table 4 reports the details. Although all predictors share similar coefficients, the highest coefficient is assigned to climate distance, implying its substantial predictive power on semantic diversity of concepts in our data. The contribution of geographical distance appears only marginally significant, likely due to its interaction with climate.

predictor	coeff. ($\beta\times10$)	std err (β)	t-stat	pval
coeff	5.589	0.003	202.396	$<$0.001
PHY	0.086	0.003	2.587	0.005
GEO	0.087	0.006	1.452	0.063
CLM	0.141	0.006	2.505	0.006

Table 4: Multiple regression analysis predicting SDist.

Evidence beyond Language Phylogeny. We hypothesize that the effects of environmental factors (GEO and CLM) are

[6] We excluded English and Spanish because their widespread native use prevents isolating their geographic and climate data.

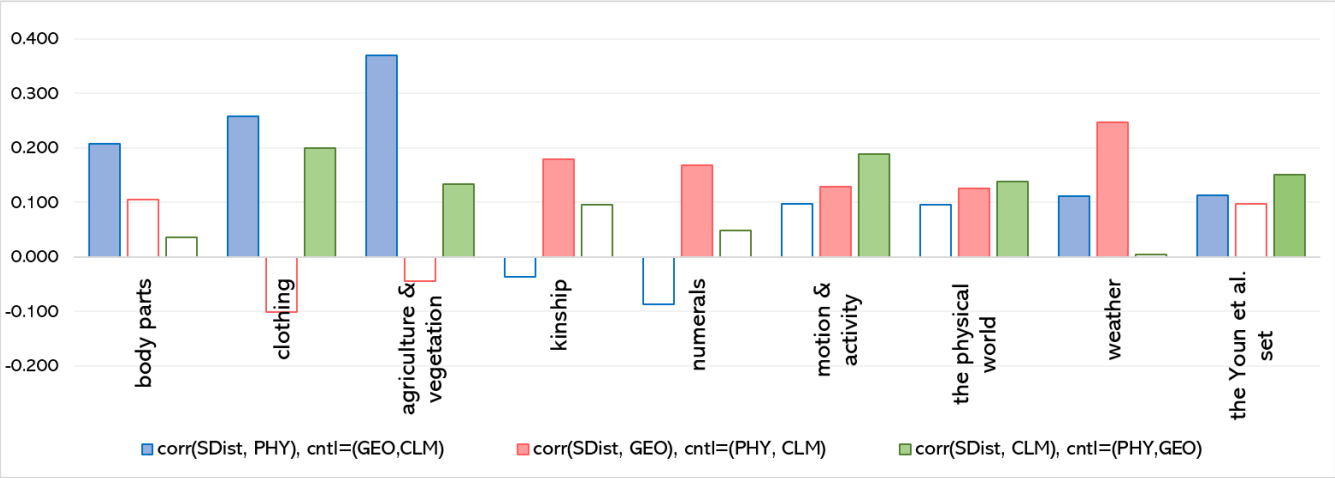

Figure 4: Partial correlation tests across focused domains. Each bar shows the correlation of language-pair SDist with one of the factors (PHY, GEO, and CLM) while controlling for the other two. Empty bars show non-sig. correlation ($p>0.05$).

not uniformly distributed across domains: we expect them to be most evident in concepts that are more subject to interpretation associated with environmental conditions. We construct focused sets of concepts that represent a variety of semantic domains presumed to be affected by environmental factors to a varying extent. For example, the Clothing domain includes words like "shirt", "cap" and "boot"; Motion&Activity includes words like "ski", "boat", "sway"; and the Weather domain includes words like "cloud", "frost" and "thunder". We perform partial correlation tests of the predictive power of each of the predictors—PHY, GEO and CLM—on pairwise language distance within each domain (while controlling for the other two). Figure 4 presents the results.

Our methodology reveals an intricate interaction between the semantic domains and the influencing factors. We found that each factor—not just phylogeny—plays a non-trivial role in explaining the cross-language semantic affinities, and our results are in accord with some independent findings from the literature. In particular, phylogeny is the strongest and the only significant predictor in the domain of Body Parts. This finding is consistent with evidence that semantic shifts in body-part terms provide important clues to proto-language reconstruction (Matisoff, 1985). In contrast, we observed that phylogeny alone is not sufficient to account for affinity of Clothing, where climate would naturally co-shape the semantic typology. Additionally, geography is a salient factor in the domains of Kinship and Numerals, which relates to findings that suggest kinship networks vary along geographical areas (Murphy, 2008), and that numeral systems in a language family are shaped by areal diffusion (Epps, 2005). Finally, the significant effect of climate on the Youn et al. Set—a domain that is reported to exhibit cross-linguistic "universals"—mirrors their own finding that the split of languages by *humid* and *arid* areas was an exceptional case to their universal semantics hypothesis (cf. Youn et al., 2016).

Conclusions

Lexical semantic typology reflects both variation and commonalities in the patterns of polysemy across languages. We proposed a principled and large-scale approach to the study of cross-linguistic lexical semantic structure at the levels of individual concept and semantic domain. We evaluated our framework against existing findings in previous studies, demonstrating results that conform to established fundamentals pertaining to semantic variation across languages. Through the analysis of a subset of Indo-European languages, our framework discovered that extra-linguistic factors of geography and climate carry over explanatory value regarding semantic variation between languages—above and beyond genealogical relations. Our work suggests that the environment may play an important role in explaining the cross-linguistic variation in polysemy.

Despite these advances, our approach is currently limited by the reliance on a manually curated dataset to provide the translation equivalents across languages of the concepts we investigate. In the future, we plan to apply our framework on translation equivalents extracted automatically (e.g., via word alignment in bilingual corpora), thereby extending it to additional concepts and languages.

Acknowledgements

ER and SS are funded through NSERC grant RGPIN-2017-06506 to SS. YX is funded through an NSERC Discovery Grant, a SSHRC Insight Grant, and a Connaught New Researcher Award.

References

Artetxe, M., & Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *TACL*, *7*, 597–610.

Ballweg, J. A. (1969). Extensions of meaning and use for kinship terms. *American Anthropologist*, *71*(1), 84–87.

Beinborn, L., & Choenni, R. (2019). Semantic drift in multilingual representations. *arXiv preprint arXiv:1904.10820*.

Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution*. Berkeley: UC Press.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & Jégou, H. (2017). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Datta, B. N. (2010). *Numerical linear algebra and applications* (Vol. 116). Siam.

Dellert, J., Daneyko, T., Münch, A., Ladygina, A., Buch, A., Clarius, N., ... others (2017). Northeuralex: A deep-coverage lexical database of northern eurasia.

Eger, S., Hoenen, A., & Mehler, A. (2016). Language classification from bilingual word embedding graphs. In *COLING* (pp. 3507–3518).

Epps, P. (2005). Areal diffusion and the development of evidentiality: Evidence from Hup. *Studies in Language.*, *29*(3), 617–650.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. In *LREC*.

Gray, R. D., & Atkinson, Q. D. (2003). Language-tree divergence times support the anatolian theory of indo-european origin. *Nature*, *426*(6965), 435.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *ACL* (pp. 1489–1501).

Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic bulletin & review*, *23*(6), 1744–1756.

Kottek, M., Grieser, J., Beck, C., Rudolf, B., & Rubel, F. (2006). World map of the köppen-geiger climate classification updated. *Meteorologische Zeitschrift*, *15*(3), 259–263.

Levinson, S. C., & Meira, S. (2003). 'Natural concepts' in the spatial topologial domain–adpositional meanings in crosslinguistic perspective: An exercise in semantic typology. *Language*, *79*, 485–516.

Majid, A., Jordan, F., & Dunn, M. (2015). Semantic systems in closely related languages. *Elsevier*.

Matisoff, J. A. (1985). Out on a limb: Arm, hand, and wing in Sino-Tibetan. *Linguistics of the Sino-Tibetan area: The state of the art*, 421–450.

Murphy, M. (2008). Variations in kinship networks across geographic and social space. *Population and Development Review*, *34*(1), 19–49.

Navigli, R., & Ponzetto, S. P. (n.d.). Babelnet: Building a very large multilingual semantic network. In *ACL*.

Pagel, M., Atkinson, Q. D., & Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout indo-european history. *Nature*, *449*(7163), 717.

Pereira, F., Gershman, S., Ritter, S., & Botvinick, M. (2016). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology*, *33*(3-4), 175–190.

Regier, T., Carstensen, A., & Kemp, C. (2016). Languages support efficient communication about the environment: Words for snow revisited. *PloS one*, *11*(4), e0151138.

Thompson, B., Roberts, S., & Lupyan, G. (2018). Quantifying semantic similarity across languages. In *CogSci.*

Xu, Y., Duong, K., Malt, B., Jiang, S., & Srinivasan, M. (in press). Conceptual relations predict colexification across languages. *Cognition*.

Youn, H., Sutton, L., Smith, E., Moore, C., Wilkins, J. F., Maddieson, I., ... Bhattacharya, T. (2016). On the universal structure of human lexical semantics. *PNAS*, *113*(7).

Appendix A

In this appendix we lay out some intuition regarding the construction of a multilingual semantic space. The procedure involves fixing a pivot language (e.g., English), and performing multiple steps of alignment of two semantic spaces (e.g., English and French), thereby generating a *bilingual* space. The process can be further scaled up to an arbitrary number of languages, pairwise aligned with the pivot, and, therefore, with each other. The essence of the construction of a bilingual semantic space lies in aligning two monolingual spaces. The input to the alignment process includes an (automatically or manually constructed) dictionary of n words in two languages $\{x_i, y_i\}, i \in \{1, .., n\}$, and two matrices—X and Y—containing d-dimensional representations (embeddings) of the n words in the two languages: source (represented by the matrix X) and target (represented by Y). The alignment procedure then learns a linear mapping (matrix W^*) between the source and the target semantic space such that:

$$W^* = \underset{W \in M_d(\mathbb{R})}{\arg\min} ||WX - Y||_F, \tag{4}$$

where d is the dimension of the embeddings, $M_d(\mathbb{R})$ is the space of $d \times d$ matrices of real numbers, and X and Y are two matrices of size $d \times n$ containing the embeddings of the words in the aligned vocabulary (Conneau et al., 2017). The 'F' notation on the right-hand side of Equation 4 denotes extracting the matrix norm (a single number) by applying the Frobenius norm (Datta, 2010), defined as:

$$||A||_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2} \tag{5}$$

for an arbitrary matrix A with $m \times n$ dimensions.

Polysemous extensions are preserved in the bilingual dictionaries by mapping a single word form with multiple meanings in a certain language into distinct words in another language, i.e., $m \times n$ mapping. As such, the French word 'mandat' is mapped into two English translation equivalents: 'mandate', 'warrant' in the automatically extracted French-English dictionary. All binlingual dictionaries used in this work are those provided by Conneau et al. (2017).