Quantifying Semantic Alignment Across Languages

Bill Thompson[1] (biltho@mpi.nl)
Sean Roberts[2] (Sean.Roberts@bristol.ac.uk)
Gary Lupyan[3] (lupyan@wisc.edu)

[1]Max Planck Institute for Psycholinguistics, Netherlands; [2]University of Bristol. [3]University of Wisconsin-Madison

Abstract

Do all languages convey semantic knowledge in the same way? If language simply mirrors the structure of the world, the answer should be a qualified "yes". If, however, languages impose structure as much as reflecting it, then even ostensibly the "same" word in different languages may mean quite different things. We provide a first pass at a large-scale quantification of cross-linguistic semantic alignment of approximately 1000 meanings in 55 languages. We find that the translation equivalents in some domains (e.g., Time, Quantity, and Kinship) exhibit high alignment across languages while the structure of other domains (e.g., Politics, Food, Emotions, and Animals) exhibits substantial cross-linguistic variability. Our measure of semantic alignment correlates with known phylogenetic distances between languages: more phylogenetically distant languages have less semantic alignment. We also find semantic alignment to correlate with *cultural* distances between societies speaking the languages, suggesting a rich co-adaptation of language and culture even in domains of experience that appear most constrained by the natural world.

Keywords: word meanings; distributional semantics; word2vec; language; culture; relativity

Introduction

English speakers call them "chairs", Spanish-speakers, "sillas", and Turkish speakers "sandalye". Despite their varying phonology, these words would seem to denote the very same objects in the world—namely chairs. But is the meaning of words even as seemingly straightforward as "chair" the same across languages? How can we know?

In this work we present one of the first large-scale quantitative examinations of semantic structure across languages (see Youn et al., 2016 for an alternate approach). We examine the extent to which supposed translation equivalents such as "chair"-"silla" have the same meanings, as assessed by analyses of distributional semantics. We use these results to quantify cross-linguistic alignment in various semantic domains, and examine how this measure of similarity relates to cultural and historical distance.

To the extent that languages *name* and *describe* the world thereby reflecting "joints of nature" that exist independently of human observers, we might expect to find for a word in any language a corresponding word in any other. For example, we might expect languages to agree on the meanings of "five", "rat", "near", and "triangle" as long as speakers of these languages have comparable exposure to the relevant data. Even the most ardent universalist would not expect a language spoken in a place without rats to have a word corresponding to "rat". On such a universalist position[1], semantic divergence between languages would be expected to be limited to cases where languages have come to name different artifacts and institutions. A language spoken by a culture without cars would not be expected to have a word for "carburetor"—a type of semantic misalignment. On the other hand, words for common animals, plants, natural objects, spatial relations, and common objects would all be expected to align. And so on a universalist position the primary reason why the semantic systems of different languages would diverge is when one language names an entity that is not named by the other language.

To the extent that language does not simply map onto existing joints of nature, but plays an important role in *creating* them, different languages may take different paths in the cultural fitness landscape (Lupyan & Dale, 2016). Consider that the category of human creations is far broader than it may at first appear. It includes color categories (the world and our physiology constrains color, but does not give us definite color boundaries) (Anderson, Biggam, Hough, & Kay, 2014; Wierzbicka, 2006), spatial terms (the world does not contain well-marked categories of "in", "out", and "across") (Bowerman & Levinson, 2001; Majid, Bowerman, Kita, Haun, & Levinson, 2004), and number systems (there is nothing natural about a decimal number system) (Calude & Verkerk, 2016; Harald Hammarström, 2010). In these and many other domains, there are numerous ways that languages *could* carve up the world. This is true even in domains where one might expect the least variability such as words for human body parts. Although people speaking different languages have objectively similar bodies, there are different solutions to partitioning the body into linguistic categories (Majid, 2015). As a result, translation equivalents of words as seemingly simple as "hand" often do not actually mean the same thing in different languages (Wierzbicka, 2013).

How can we tell if two words mean the same thing?

On first glance, one might assume that the meaning of a word in one language (L_1) and another (L_2) is the same if the two words denote an identical set of entities. If on hearing "chair" and "silla," English and Spanish speakers, respectively, pick out the same objects, we might say the

[1] We outline this position as a hypothetical limiting case rather than as a theory associated with a particular group of researchers.

words mean the same thing in English and Spanish. If it were that easy, however, there would be little need to study semantics. We review some of these difficulties below.

The first problem with this simplified definition of meaning equivalence is that most words refer to abstract and relational entities (Lupyan & Winter, 2017). How exactly would one obtain the set of entities picked out by words like "fun"? The second problem is that an equivalent word in one context may not be an equivalent in another context. For example, in English we "wash" our clothes and wash our face, but "brush" our teeth. Italian uses the same verb "lavare" for all three contexts. So does "lavare" mean "to wash" or doesn't it? A related problem is that psychologically informed word meanings are not limited to *denotative* referents, but include *connotations*. For example "impressive" translates to "impressionante" in Italian, but the former word has a positive connotation while the latter has a largely negative connotation. These connotations are psychologically real for both L1 speakers (Onnis et al., 2008) and L2 learners (e.g., Partington, 1998).

The final problem is polysemy. Even very concrete words often have multiple senses. The English word "chair" can (and in the world of this paper's readers, often does) denote people occupying managerial positions. This meaning is not shared by the Spanish translation equivalent, "silla". To the extent that "chair" even partially activates these multiple senses in the minds of English speakers, the "chair"-"silla" alignment is reduced. The issue of differential polyseymy is magnified when we look to more abstract words.

With these caveats (familiar to anyone who has attempted translation) we may define overall semantic equivalence as the aggregate similarity in the effect that the words w_1 and w_2 have on speakers of L_1 and L_2, respectively. The best way to actually quantify this measure is through rigorous and laborious consultation with native speakers (Majid, 2015). This approach is difficult to scale, however. Here, we take as a starting point, the idea that word meanings are revealed by their contexts: "you shall know a word by the company it keeps" (Firth, 1957). Recent advances in text digitization and machine-learning have made it possible to construct models of distributional semantics of unparalleled size (e.g., Mikolov et al., 2013). By being exposed to large amounts of text, these models are able to capture semantic relationships to a surprising degree of subtlety (Baroni, Dinu, & Kruszewski, 2014; Hollis & Westbury, 2016; Nematzadeh, Meylan, & Griffiths, 2017) though varying considerably for different kinds of similarity (Hill et al., 2016; Chen, Peterson, & Griffiths, 2017)

To assess semantic alignment, we take models trained on different languages and align them by using translation equivalents. This provides a fairly conservative test of semantic equivalence in that we restrict our analysis only to words which are attested to *have* translation equivalents (so we are excluding words like "carburetor", culture-specific plant and animal names, etc.). We then compute semantic alignment based on distributional patterns of these translation equivalents.

To assess the extent to which the results support linguistic universality versus diversity, we examine how semantic alignment differs by semantic domain. To reiterate: no position would predict high alignment across all domains. A more universal position gains support if the only variable domains are those that name human constructs. Relativity gains support if we find lack of semantic alignment in domains that name allegedly objective joints of nature.

Similarity and Diversity of Word Meanings

Methods

Embedding Models As our primary data we use word-embedding models trained on Wikipedia in different languages (Bojanowski et al., 2016). These models were trained using the Skipgram technique (Mikolov et al., 2013), which positions words in a semantic vector space based primarily on collocation patterns. From these models we construct semantic *networks* by computing the cosine distance between embeddings for all pairs of relevant concepts. We are of course aware that Wikipedia datasets in some languages (e.g., Spanish and Portuguese) are more similar *in content* to one another than between other languages (e.g., English and Russian). We conduct extensive modeling of these similarities (to be presented elsewhere) to ensure that the results we report below cannot be explained by the specific content contained in Wikipedia.

Translation Sets We made use of the NORTHEURALEX (NEL) dataset (Dellert & Jäger, 2017) which provides word forms, part-of-speech information and translation equivalents for 1,016 concepts in 107 languages, covering 20 language families.

Semantic Domains For semantic domains we used the chapters of Intercontinental Dictionary Series (IDS) project (Key & Comrie, 2015). These domains include Kinship, Time, Quantity, Religion and Belief, and Food & Drink. From these chapters, we were able to tag semantic domain for roughly half of the NEL concepts (~600). This subset was large enough to impute a semantic domain for the remaining NEL concepts, using multi-class regression on the embeddings, with around 70% accuracy. We compare these rankings to Wordnet classifications of each word (details presented elsewhere).

Combined Data The intersection of these datasets contains the languages present in both the embedding models and the NEL data. The concepts in the data are limited to those which are given parallel wordforms by NEL *and* vectors by the embeddings models. After combining data, our primary dataset consists of 46,089 wordforms across 55 languages (1485 unique language pairs). This allows us to make 1,012,330 unique comparisons of a concept's network structure between language pairs.

Computing the Semantic Alignment of a Concept Between Languages Intuitively, our procedure is as follows. Take a concept, and look around it in semantic space to identify its near neighbors. Do the same for this concept in another language. Count up the number of neighbors common to both languages. Align the common-neighbor networks in both languages, and measure their agreement. More formally: for every unique pair of languages (L_1 and L_2), we computed, for every individual concept (C) that had a vector embedding available in both L_1 and L_2, the following statistic (which we call r_c). Compute, in L_1, the semantic similarity between C and all other terms in the NORHEURALEX set of concepts (for which embeddings are available in L1). Using these distances, find the N closest neighbors of C in L_1 (words with the smallest cosine distance to C). Repeat this procedure to find the N closest neighbors to C in L_2. Identify the concepts that appear in both neighbor lists, and call this set the *neighbor intersection*. Compute how strongly the similarity scores between C and the neighbor intersection in L_1 correlate with the similarity scores between C and the neighbor intersection in L_2 using Pearson's *r* (similar results are obtained using Spearman's *rho*). Take the correlation coefficient to be a measure of the structural similarity of C in L_1 and L_2. A high coefficient ($i.e. r_c(L_1, L_2) \to 1$) indicates that—at least within the network of words available to our analyses—C keeps a similar pattern of company in L_1 and L_2, and so (on this definition of semantic equivalence) the word means close to the same thing in L_1 and L_2.

As an example, Figure 1A shows neighbor sets for "Friday" / *"vendredi"* in English and French (setting N = 40 for all analyses presented here; ongoing work is investigating). This meaning behaves very similarity in these two languages: its closest neighbors in both languages tend to be in the neighbor intersection (i.e. if "Friday" has a close neighbor in English, then the translation of that neighbor is likely a close neighbor of "vendredi" in French). Neighbors of "Friday" / "vendredi" that are language specific (i.e. neighbors in only one of the two languages) tend to be relatively distant semantic neighbors (low cosine similarity). Therefore, the semantic alignment of the meaning conveyed by "Friday"/"vendredi" is quite high: $r_{Friday}(En, Fr) = 0.94$. Figure 1B shows neighbor sets in French and English for the meaning correspond to "worker". The pattern of shared close neighbors is much reduced: around half of the neighbors of this meaning are language-specific. "Worker" / "ouvrier" tends not to have closely concentrated neighbors in either language per se (note scale differences between A and B in Fig1). In this respect, our metric identifies a similarity (the correlation would be lower if the concept had close neighbors in one language but not the other). In the same way, although neighbors aren't generally close, shared neighbors tend not to show distance disparities between the two languages. These properties lead "worker" to gain an intermediate alignment , $r_{Worker}(En, Fr) = 0.5$.

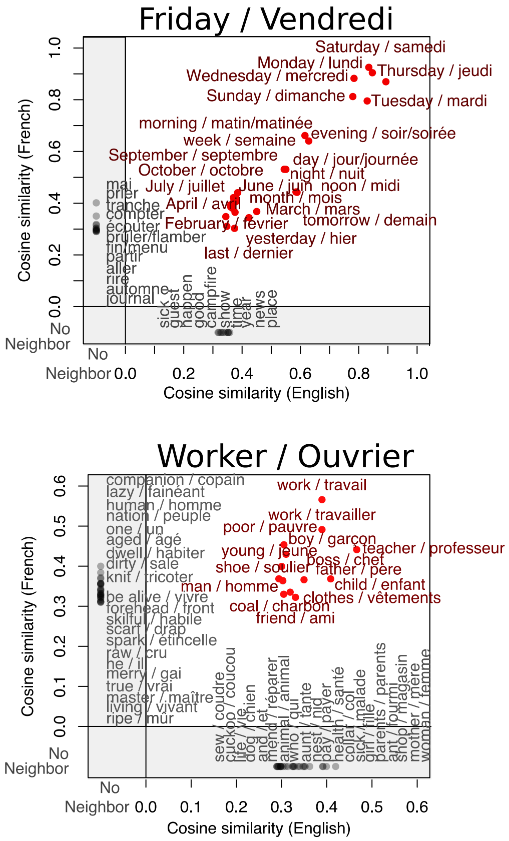

Figure 1: Example semantic neighbor sets in English and French for (A) *Friday / Vendredi* which shows high alignment, and (B) for *Worker / Ouvrier* which show low alignment. Values lower than 0 indicate that the form was not a neighbor of the target word in the given language.

Results

We computed this structural alignment statistic for all available concepts and language pairings. We explored a number of data filters and subsets (e.g. filtering by Wikipedia size and quality, or by minimum number of language pairs per concept, etc.), but found none to challenge the general pattern of results we report. As such, we simply subset the data to only those comparisons whose *neighbor intersection* included more than five concepts, and to only those semantic domains which comprise 20 or more unique concepts. Here we focus on two key results: divisions of the data by semantic domain and word class.

Cross-linguistic Structural Diversity by Domain Figure 2 shows a ranking of semantic domains by average semantic alignment across languages. To compute this ranking, we took the average value of *r* over all concepts tagged within a domain, over all pairs of languages in which a comparison could be made. Shared vocabulary relating to Quantity (e.g. *first, second, last, third, sixty, eighty, a thousand, half*), Time (e.g. *December, January, Wednesday, tomorrow, winter, wait, begin*), and Kinship (*father, old, sister, son, mother, child, husband, uncle, brother, grandfather, woman, you*) exhibit the most structural alignment across languages in our sample. Food and Drink (e.g. *dish, cup, egg, boil, ripe, prepare, onion, hunger, raw*) and Social & Political Relations (e.g. *village, town, friend, master, people, invite, king, meet, help, hinder, power*) feature at the opposite end of the ranking, exhibiting variety. Figure 3 demonstrates this difference between cross-linguistically regular versus idiosyncratic domains, showing matched semantic networks among concepts belonging to the domains Time and Food, in 3 different languages.

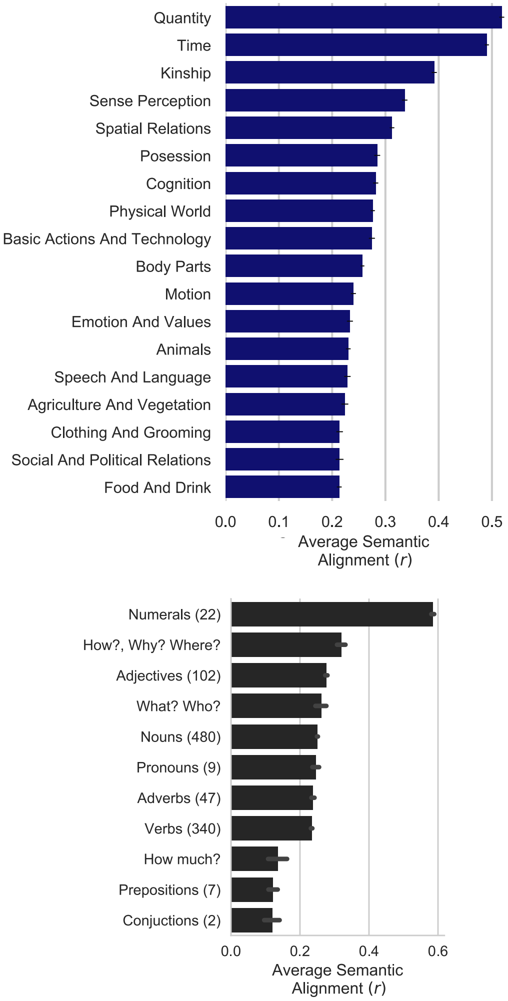

Figure 2: Overall cross-linguistic semantic alignments of IDS semantic domains (Top), and parts-of-speech with some words of interest singled out (Bottom).

Cross-linguistic Structural Diversity by Word Class We also examined semantic alignment by word-class (Fig. 2 bottom). Semantic alignment of Numerals is around twice that of next closest word class (note that *Quantity* in Fig. 2 (top) additionally includes quantifiers like "whole" and "half"). Two insights stand out. First, Numerals are known independently to have exceptionally slow rates of diachronic change in general. Second, the ranking shows a striking agreement with an independent ranking of word classes by rates of *phonological* change (Meade, Pagel, & Atkinson, 2007).

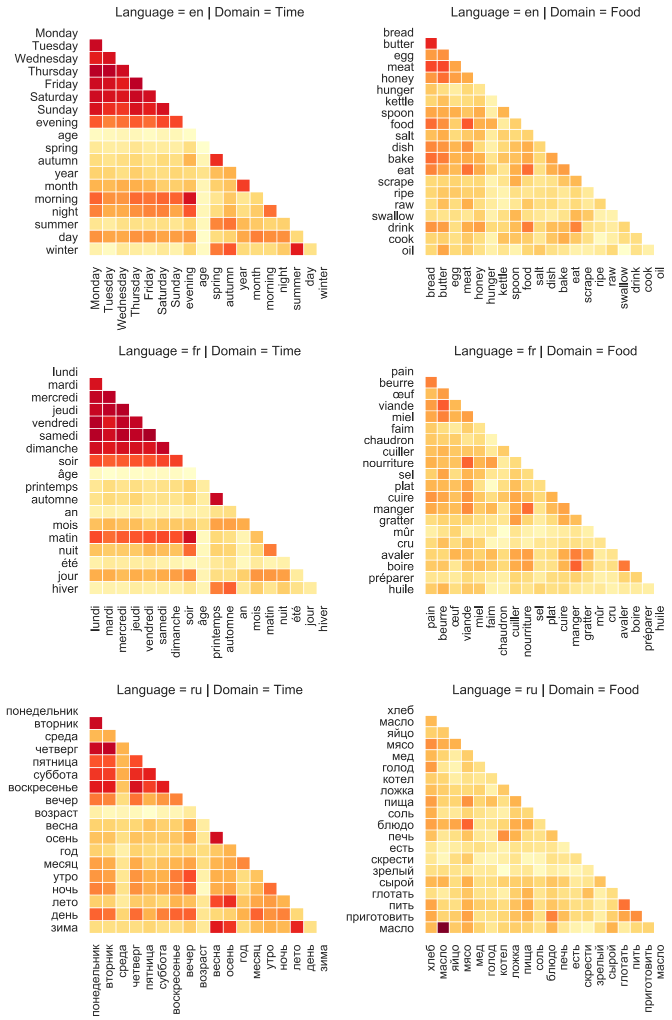

Figure 3: Matched Semantic networks for Time (left) and Food (right) related concepts, in three languages.

Semantic Alignment Predicts Language Phylogeny

Given the link to historical change, we can test whether semantic similarity correlates with historical relatedness.

Methods

Semantic distances For every pair of languages (1485 unique pairs), we calculated the mean pairwise semantic similarity, $\rho_*(L1, L2)$, averaging over concepts and domains, to approximate what we will call the 'linguistic distance' between languages, based on semantics. Figure 4 shows the 50 language pairs judged by our model to be most similar, and their mean similarities.

Phylogenetic differences For 19 Indo-European languages in our data (171 pairs), established historical distances are available from a phylogenetic tree based on linguistic forms (independent of semantics, Bouckaert et al., 2010). Patristic distances between languages in the tree are used as a measure of historical distance between societies.

Results

Mantel test correlations suggest that semantic alignment between language pairs correlate with their historical distance (r = -0.39, one-tailed p = 0.003). More historically distant languages are less semantically aligned.

Semantic Alignment Predicts Cultural Distances Between Languages

Different societies may conceptualize the world in different ways, or make finer distinctions in domains that matter to them. Languages should adapt to these differences (Lupyan & Dale, 2016), which predicts that semantic alignment should decrease with greater cultural distance.

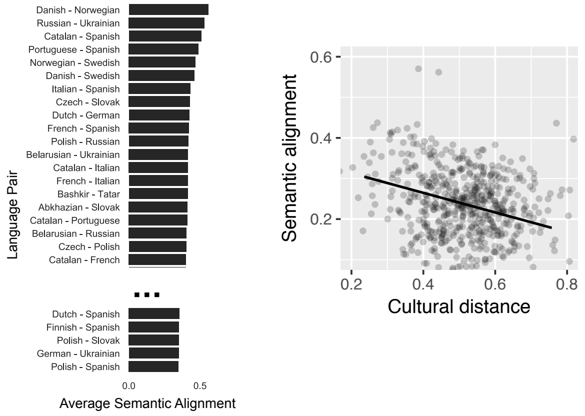

Figure 4: **Left:** Language pairs by semantic alignment; **Right:** The relationship between semantic alignment (r) and cultural distance for 561 language pairs. Regression line derived from a mixed effects model controlling for shared ancestry.

Methods

We obtained 92 cultural traits (e.g. norms for marital residence, rules for political succession) for 34 societies from the Ethnographic Atlas as linked to languages in D-PLACE (Kirby et al., 2016). Missing values were imputed by multiple imputation using classification and regression trees (van Buuren & Groothuis-Oudshoorn, 2010). During testing, this method imputed the correct value for unseen data 74% of the time, compared to a random sampling baseline of 19%. Cultural distances were calculated as the average Gower distances between traits in 100 imputed sets. We compared cultural distances between societies to linguistic similarities between societies, controlling for shared history in two ways: 1: mixed effects modelling with Language-family pair (according to Glottolog, H. Hammarström, Bank, Forkel, & Haspelmath, 2018) included as a random effect. This enabled the model to capture the likelihood that, for example, two languages from the Indo-European language family will be more similar to each other than two languages from different language families. The same was done with geographic area according to the AUTOTYP database (Bickel et al., 2017). The models included random intercepts and slopes for the effect of cultural distance. The second test controls for history using the phylogenetic tree of Indo-European with a partial Mantel test.

Results

Linguistic and cultural distances were significantly correlated under both controls for common history. Controlling for language family and geographic area (test one) we found a significant relationship (β= -0.34, χ^2=10.2, p=0.001, Fig. 4). Likewise, linguistic similarities and cultural distances were moderately correlated in test two (Mantel r = -0.40[-0.54,-0.3], one-tailed p=0.02), even when partialing out the effect of historical divergence (Mantel r= -0.31[-0.45,-0.21], one-tailed p=0.04). These results suggest that the semantic differences between languages are to some extent reflecting cultural differences. The effect was stronger for concepts related to kinship, and weaker for those related to agriculture and vegetation.

General Discussion

A vocabulary of a language is an organizational scheme. If this organizational scheme is largely determined by the objective joints of nature and shared joints of our minds, we would expect vocabularies of different languages to largely align. If, instead vocabularies not only reflect some pre-existing structures in the world, but also *impose* structure, we might expect different vocabularies to impose detectably different organizational schemes. In this work we present one of the first large-scale quantitative investigations of this question by examining the extent to which word meanings—defined here using distributional semantics—align across languages. We found that words pertaining to Quantity and Time have the greatest semantic alignment. This suggests that these words have a natural structure, which may result from objective joints in the world and/or common cognitive organizing principles. This does not mean that these semantic domains are not human constructions. Numeric and calendar systems are human inventions. What the high alignment for these domains shows is that for languages using decimal systems, 7 days of the week, etc., the matching words are closely aligned, a proxy for meaning the same thing.

The domains showing least semantic alignment pertain to human institutions (as expected), but interestingly, words relating to Animals (e.g., "fish") common actions (e.g., "wash") and the physical world (e.g., "stone", "sea") show only intermediate levels of alignment: these domains appear more variable than expected on a universalist thesis (Youn et al., 2016) (although a quantifiable baseline is currently missing). These words do not align in the way they should if their full meanings simply picked out natural categories in the world. Our findings support the possibility that languages and cultures co-adapt to forge a human-constructed representation of the world that can vary across populations (Majid, 2015) and is not predicted by the view that lexical semantics are strongly constrained by objective joints of nature. While the current data are highly preliminary, our approach is capable of making strong predictions about the semantic variation we should find among native speakers of the world's languages. We recognize that our conclusions derive from semantics based on distributional models that, while correlating with human judgments, only roughly approximate psychologically real semantic representations. Testing model predictions experimentally is a key priority going forward.

References

Anderson, W., Biggam, C. P., Hough, C., & Kay, C. (Eds.). (2014). *Colour Studies: A broad spectrum*. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/z.191

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 238–247). Baltimore, MD. Retrieved from http://anthology.aclweb.org/P/P14/P14-1023.pdf

Bickel, B., Nichols, J., Zakharko, T., Witzlack-Makarevich, A., & et al. (2017). *The AUTOTYP typological databases. Version 0.1.0*.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching Word Vectors with Subword Information. *ArXiv:1607.04606 [Cs]*. Retrieved from http://arxiv.org/abs/1607.04606

Bowerman, M., & Levinson, S. C. (2001). *Language acquisition and conceptual development*. Cambridge University Press.

Calude, A. S., & Verkerk, A. (2016). The typology and diachrony of higher numerals in Indo-European: a phylogenetic comparative study. *Journal of Language Evolution*, *1*(2), 91–108.

Chen, D., Peterson, J. C., & Griffiths, T. L. (2017). Evaluating vector-space models of analogy. *ArXiv:1705.04416 [Cs]*. Retrieved from http://arxiv.org/abs/1705.04416

Dellert, J., & Jäger, G. (Eds.). (2017). NorthEuraLex - Lexicostatistical Database of Northern Eurasia. University of Tubingen. Retrieved from http://northeuralex.org/

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955.

Hammarström, H., Bank, S., Forkel, R., & Haspelmath, M. (2018). *Glottolog 3.2*. Jena: Max Planck Institute for the Science of Human History.

Hammarström, Harald. (2010). Rarities in numeral systems. *Rethinking Universals: How Rarities Affect Linguistic Theory*, *45*, 11–53.

Hill, F., Reichart, R., & Korhonen, A. (2016). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*. Retrieved from http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00237

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. https://doi.org/10.3758/s13423-016-1053-2

Key, M. R., & Comrie, B. (Eds.). (2015). *The Intercontinental Dictionary Series*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from http://ids.clld.org/

Kirby, K. R., Gray, R. D., Greenhill, S. J., Jordan, F. M., Gomes-Ng, S., Bibiko, H.-J., … others. (2016). D-PLACE: A global database of cultural, linguistic and environmental diversity. *PloS One*, *11*(7), e0158391.

Lupyan, G., & Dale, R. (2016). Why are there different languages? The role of adaptation in linguistic diversity. *Trends in Cognitive Sciences*, *20*(9), 649–660. http://dx.doi.org/10.1016/j.tics.2016.07.005

Lupyan, G., & Winter, B. (2017). Language is more abstract than you think, or, why aren't languages more iconic? *PsyArXiv*. https://doi.org/10.17605/OSF.IO/YZ3UN

Majid, A. (2015). Comparing lexicons cross-linguistically. In J. R. Taylor (Ed.), *The Oxford handbook of the word* (pp. 364–379). New York, NY: Oxford University Press.

Majid, A., Bowerman, M., Kita, S., Haun, D. B. M., & Levinson, S. C. (2004). Can language restructure cognition? The case for space. *Trends in Cognitive Sciences*, *8*(3).

Meade, A., Pagel, M., & Atkinson, Q. D. (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature*, *449*(7163), 717. https://doi.org/10.1038/nature06176

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *ArXiv Preprint ArXiv:1301.3781*. Retrieved from https://arxiv.org/abs/1301.3781

Nematzadeh, A., Meylan, S. C., & Griffiths, T. L. (2017). Evaluating Vector-Space Models of Word Representation, or, The Unreasonable Effectiveness of Counting Words Near Other Words. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*. London.

Onnis, L., Farmer, T., Baroni, M., Christiansen, M., & Spivey, M. (2008). Generalizable distributional regularities aid fluent language processing: The case of semantic valence tendencies. *Italian Journal of Linguistics*, *20*, 125–152.

Partington, A. (1998). *Patterns and meanings: Using corpora for English language research and teaching* (Vol. 2). John Benjamins Publishing.

van Buuren, S., & Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 1–68.

Wierzbicka, A. (2006). The semantics of colour: A new paradigm. In C. P. Biggam & C. Kay (Eds.), *Progress in Colour Studies* (pp. 1–24). Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/z.pics1.05wie

Wierzbicka, A. (2013). *Imprisoned in English: The Hazards of English as a Default Language* (1 edition). Oxford: Oxford University Press.

Youn, H., Sutton, L., Smith, E., Moore, C., Wilkins, J. F., Maddieson, I., … Bhattacharya, T. (2016). On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences*, *113*(7), 1766–1771.