CO-EVOLUTION OF CULTURE AND MEANING REVEALED THROUGH LARGE-SCALE SEMANTIC ALIGNMENT

BILL THOMPSON[*1], SEÁN G. ROBERTS[2], and GARY LUPYAN[3]

[*]Corresponding Author: billdthompson@berkeley.edu
[1]Social Science Matrix, University of California, Berkeley, USA
[2]ENCAP, Cardiff University, UK; Anthropology and Archaeology, University of Bristol, UK
[3]Department of Psychology, University of Wisconsin, Madison, USA

Do natural languages evolve to reflect the objective structure of the world (Gleitman & Fisher, 2005; Snedeker & Gleitman, 2004) or do they impose their own structure, with each language adapting to local communicative needs (Evans & Levinson, 2009; Davidson, 1973)? If languages mirror the objective structure of the world, words referring to natural kinds, common artifacts, and universal human actions should mean the same things in different languages. In contrast, if different languages impose their own structure, carving joints *into* nature, word meanings may exhibit substantial variability between languages, making cross-linguistic semantic alignment more difficult. This would not be surprising for specialised artefacts, regional animals or distinctions that relate to specific local conditions (e.g. distinctions between "ice" and "snow" are more likely in colder climates, Regier, Carstensen, & Kemp, 2016). However, the extent of the alignment between more common meanings (common animals and artifacts, natural features, quantifiers, body parts, and common verbs) is an open question. For example, do the English words 'five', 'near', and 'arm' mean the same thing as the Spanish words 'cinco', 'cerca" and 'brazo', respectively?

Quantifying semantic structure is difficult because word meanings are not directly observable (Cuyckens, Dirven, & Taylor, 2009). Here, we present a large-scale analysis of word meanings by taking advantage of recent advances in distributional semantics using machine-learning on natural language text. We obtained translation equivalents for 1,016 concepts in 74 languages using the NorthEuraLex dataset (Dellert & Jäger, 2017). We began by deriving within-language word-to-word similarities using the fast-text skipgram algorithm trained on language-specific versions of Wikipedia (Bojanowski, Grave, Joulin, & Mikolov, 2017). We also replicated on word embeddings derived from the OpenSubtitles database (Lison & Tiedemann, 2016) and a combination of Wikipedia and the Common Crawl dataset (Grave, Bojanowski, Gupta, Joulin, & Mikolov, 2018)). To compute semantic alignments for meaning c for language pair L_i and L_j, we first found the closest k semantic neighbours of c of L_i along with their context similarity score.

For example, the closest neighbours to the English word 'beautiful' are 'colourful' (.55), 'love' (.53) and 'delicate' (.51). We then found the translations of these neighbours in L_j and their corresponding proximity to the translation of c. The directional semantic alignment $L_i \rightarrow L_j$ is the correlation between c's similarity to these neighbours in both languages. For example, the French translations of these neighbours are more distant from 'beau' ('multicolore'=.22, 'aimer'=.32 and 'fin'=.2), while other words are closer ('frère', 'père') so the alignment is low. This was repeated in the opposite direction: the k closest semantic neighbours to c in L_j were identified and matched to their translations in L_i; the same correlation was calculated for $L_i \rightarrow L_j$. Final semantic alignment is the average of these two correlations. We validated the measure by correlating it with human translatabiliy judgements (e.g., Tokowicz et al., 2002; Allen & Conklin, 2014).

The most alignable meanings across languages stand out not as being especially concrete or reflecting natural joints, but as domains that have high internal coherence such as number words and kinship terms. In comparison, words for common artifacts, actions, and natural kinds have much lower alignments indicating that these words have different semantic neighborhoods in different languages.

If languages reflect cultural factors, then languages should be more aligned if they are spoken by people with similar cultures. We confirmed that cultural similarity (the proportion of cultural traits in common based on 92 non-linguistic cultural traits for 39 societies, Kirby et al., 2016) predicted semantic alignment between languages, even when controlling for historical relatedness and geographic proximity (b= 0.2, $\chi^2(1)$= 16.56, p <.001). Cultural similarity related to subsistence type was correlated with semantic alignment in domains including 'food and drink' (r = .3), 'animals' (r = .29), 'agriculture and vegetation' (r = .25), 'clothing and grooming' (r = .25), 'social and political relations' (r = .15), and 'spatial relations' (r = .1, all adjusted p-values $<$ .05). These reflect well-known relations between subsistence types and culture (Murdock & Provost, 1973; Sellen & Smay, 2001; Peoples & Marlowe, 2012; Botero et al., 2014; Gavin et al., 2018; Majid et al., 2018). This indicates that cultural and historical processes influence the evolution of natural language semantics. Consistent with the idea that languages that emerge in larger communities have more systematic structure (Raviv, Meyer, & Lev-Ari, 2019), we find that semantic alignment positively associated with population size. Controlling for shared history, languages spoken by larger groups tend to align better with one another (b = .002, t = 7.1, $p < .001$).

Our results show that even frequent concrete meanings show substantial cross-linguistic differences – differences which are predictable from shared culture and history. Despite some of the shortcomings of corpus-derived semantics (which makes our analysis more conservative), we believe the present work provides a major step forward for understanding the evolutionary factors that shape the emergence and evolution of linguistic meaning, and particularly the impact of shared culture (Thompson et al., 2016).

References

Allen, D., & Conklin, K. (2014). Cross-linguistic similarity norms for Japanese–English translation equivalents. *Behavior Research Methods*, *46*(2), 540-563.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Botero, C. A., Gardner, B., Kirby, K. R., Bulbulia, J., Gavin, M. C., & Gray, R. D. (2014). The ecology of religious beliefs. *Proceedings of the National Academy of Sciences*, *111*(47), 16784–16789.

Cuyckens, H., Dirven, R., & Taylor, J. R. (2009). *Cognitive approaches to lexical semantics* (Vol. 23). Walter de Gruyter.

Davidson, D. (1973). On the Very Idea of a Conceptual Scheme. *Proceedings and Addresses of the American Philosophical Association*, *47*, 5-20. (ArticleType: research-article / Full publication date: 1973 - 1974 / Copyright © 1973 American Philosophical Association)

Dellert, J., & Jäger, G. (2017). *NorthEuraLex (version 0.9).*

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and brain sciences*, *32*(5), 429–448.

Gavin, M. C., Kavanagh, P. H., Haynie, H. J., Bowern, C., Ember, C. R., Gray, R. D., Jordan, F. M., Kirby, K. R., Kushnick, G., Low, B. S., Vilela, B., & Botero, C. A. (2018). The global geography of human subsistence. *Royal Society Open Science*, *5*(9), 171897.

Gleitman, L., & Fisher, C. (2005). Universal aspects of word learning. In J. McGilvray (Ed.), *The Cambridge Companion to Chomsky* (p. 123-142). New York, NY: Cambridge University Press.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the international conference on language resources and evaluation (lrec 2018).*

Kirby, K. R., Gray, R. D., Greenhill, S. J., Jordan, F. M., Gomes-Ng, S., Bibiko, H.-J., Blasi, D. E., Botero, C. A., Bowern, C., Ember, C. R., et al.. (2016). D-place: A global database of cultural, linguistic and environmental diversity. *PloS one*, *11*(7), e0158391.

Lison, P., & Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the international conference on language resources and evaluation (lrec 2016).*

Majid, A., Roberts, S. G., Cilissen, L., Emmorey, K., Nicodemus, B., OGrady, L., Woll, B., LeLan, B., De Sousa, H., Cansler, B. L., Shayan, S., Vos, C. de, Senft, G., Enfield, N. J., Razak, R. A., Fedden, S., Tufvesson, S., Dingemanse, M., Ozturk, O., Brown, P., Hill, C., Guen, O. L., Hirtzel,

V., Gijn, R. van, Sicoli, M. A., , & Levinson, S. C. (2018). Differential coding of perception in the worlds languages. *Proceedings of the National Academy of Sciences*, *115*(45), 11369–11376.

Murdock, G. P., & Provost, C. (1973). Factors in the division of labor by sex: A cross-cultural analysis. *Ethnology*, *12*(2), 203–225.

Peoples, H. C., & Marlowe, F. W. (2012). Subsistence and the evolution of religion. *Human Nature*, *23*(3), 253–269.

Raviv, L., Meyer, A., & Lev-Ari, S. (2019). Larger communities create more systematic languages. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1907), 20191262.

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

Sellen, D. W., & Smay, D. B. (2001). Relationships between subsistence and age at weaning in "preindustrial" societies. *Human Nature*, *12*, 47–87. (Accessed on: 2019-01-18)

Snedeker, J., & Gleitman, L. (2004). Why is it hard to label our concepts? In D. G. Hall & S. Waxman (Eds.), *Weaving a Lexicon* (illustrated edition ed., p. 257-294). Cambridge, MA.: The MIT Press.

Thompson, B., Kirby, S., & Smith, K. (2016). Culture shapes the evolution of cognition. *Proceedings of the National Academy of Sciences*, *113*(16), 4530–4535.

Tokowicz, N., Kroll, J. F., De Groot, A. M., & Van Hell, J. G. (2002). Number-of-translation norms for dutchenglish translation pairs: A new tool for examining language production. *Behavior Research Methods, Instruments, & Computers*, *34*(3), 435–451.