The relationship between associativity and form-meaning associations: A large scale replication

Thomas Brochhagen (thomas.brochhagen@upf.edu, matriculation no. NNNNNN)
FirstName SurName (firstname.surname@upf.edu, matriculation no. NNNNNN)

2022-02-23

Introduction

Natural languages often use the same form to express different meanings. For instance, in Spanish the word *dedo* means both finger and toe. Moreover, certain meanings are expressed by the same form more often than others (Jackson et al. 2019; Xu et al. 2020). The finger-toe ambiguity, for instance, is not unique to Spanish but found in more than 130 diverse languages (Rzymski et al. 2020). However, it is not yet clear why certain meanings are more often expressed by the same form than others. Past research on this question suggests that meanings that are more similar are more likely to be expressed by the same form across languages (Xu et al. 2020,@karjus+etal:2021). We here test whether the findings of Xu et al. [@-xu+etal:2021] can be reproduced using more cross-linguistic data and with a different source of information for semantic similarity. This is important because we want to know whether these results generalize from 200 languages, in Xu et al., to over 1200 languages; and whether they are robust to small changes in the setup.

If the results from Xu et al. replicate, the likelihood of two meanings to be expressed by the same form should increase with how closely related they are. A priori, we have no reason to expect them not to replicate. All analysis scripts used in this report are available in link.

Material and methods

(Xu et al. 2020) used a dataset spanning 200 languages and used associativity as a proxy for semantic relatedness. Associativity measures how close two meanings are based on human free association data: Subjects are prompted with a word (e.g., *dog*) and are asked to provide three associates (e.g., *cat*; *bone* and *cuddly*). We here use the best performing English associativity scores from Small World of Words (De Deyne et al. 2018). As far as we know, this is the best and largest associativity data set. For linguistic data, we base our analysis on CLICS (Rzymski et al. 2020). As noted above, this data set is much larger than that used by Xu et al.; it covers 1200 languages (vs. 200 in Xu et al.) and more than 1400 meanings, totalling 203056 data points.

As Xu et al., we use a logistic regression to characterize this data. The response variable is whether a pair of meanings is expressed by the same form in a language (e.g., 1 for finger and toe in Spanish and 0 for the same pair in English). The only predictor is the associativity of the two meanings (e.g., how associated finger and toe are according to the data from De Deyne et al. 2018).

De Deyne, Simon, Danielle J. Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2018. "The 'Small World of Words' English Word Association Norms for over 12,000 Cue Words." *Behavior Research Methods* 51 (3). Springer Science; Business Media LLC: 987–1006. https://doi.org/10.3758/s13428-018-1115-7.

Jackson, Joshua Conrad, Joseph Watts, Teague R. Henry, Johann-Mattis List, Robert Forkel, Peter J. Mucha, Simon J. Greenhill, Russell D. Gray, and Kristen A. Lindquist. 2019. "Emotion Semantics Show Both

Cultural Variation and Universal Structure." *Science* 366 (6472). American Association for the Advancement of Science (AAAS): 1517–22. https://doi.org/10.1126/science.aaw8160.

Karjus, Andres, Richard A. Blythe, Simon Kirby, Tianyu Wang, and Kenny Smith. 2021. "Conceptual Similarity and Communicative Need Shape Colexification: An Experimental Study." *Cognitive Science* 45 (9). Wiley. https://doi.org/10.1111/cogs.13035.

Rzymski, Christoph, Tiago Tresoldi, Simon J Greenhill, Mei-Shin Wu, Nathanael E Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, et al. 2020. "The Database of Cross-Linguistic Colexifications, Reproducible Analysis of Cross-Linguistic Polysemies." *Scientific Data* 7 (1). Nature Publishing Group: 1–12.

Xu, Yang, Khang Duong, Barbara C Malt, Serena Jiang, and Mahesh Srinivasan. 2020. "Conceptual Relations Predict Colexification Across Languages." *Cognition* 201. Elsevier.