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

FirstName1 SurName1 (firstname1.surname1@upf.edu, matriculation no. NNNNNN)
FirstName2 SurName2 (firstname2.surname2@upf.edu, matriculation no. NNNNNN)

2022-03-11

Introduction

Natural languages often use the same form to express different meanings. For instance, in Catalan the word *dit* can mean both finger and toe. Moreover, cross-linguistically, 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 Catalan but found in more than 130 diverse languages across the globe (Rzymski et al. 2020). It is not yet clear why certain meanings are more often expressed by the same form than others; but understanding this may help us uncover the cognitive underpinnings of the way in which meaning is universally organized.

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 et al. 2021). We here test whether the findings of Xu et al. can be reproduced using more cross-linguistic data, and using a different source of information for semantic similarity. This is important because we want to know whether these results generalize from about 200 languages that Xu et al. tested, to the 1261 languages that we test; and whether they are robust to changes in the source and transformations of a predictor variable.

If the results from Xu et al. are reproduced, 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 reproduce All analysis code used within this report is available as a subset of a larger study, found here: https://osf.io/hjvm5/?view_only=cde6d3ed716a4e1dbc9f271a53ae875c.

Material and methods

Xu et al. used a dataset spanning about 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, Small World of Words is the largest associativity data set available to date for English. For data on form-meaning associations across languages, 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 1261 languages (vs. 200 in Xu et al.) and 1301 meanings, totalling 283354 data points.

Following 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 Catalan or Spanish; and 0 for the same pair of meanings in English). The sole predictor is the standardized associativity of the two meanings (e.g., how associated finger and toe are according to the data experimental data from De Deyne et al. 2018).

Results

The dataset is inbalanced. There are 215070 meaning pairs that are not expressed by the same form and 68284 that are (mean of 0.2409848; total of 283354 data points). As shown in Figure 1, most meaning pairs either are –or are not– expressed by the same form in a language (see histogram on the y-axis). As a consequence, accounting for the sparse gradient in the middle is a difficult predictive task. That is, estimating the likelihood whether a pair of meanings will colexify is hard because, for most pairs, we only have evidence that they either always do or always do not.

Notwithstanding, the results from Xu et al. are reproduced: an increase of 1 standard deviation in associativity is predicted to increase the likelihood of a pair of meanings to be expressed by the same form by 1.94 on the logit scale; by a probability of 0.87; and by odds of 6.93. In other words, the more strongly associated two meanings are, the more likely they are to be expressed by the same form in a language.

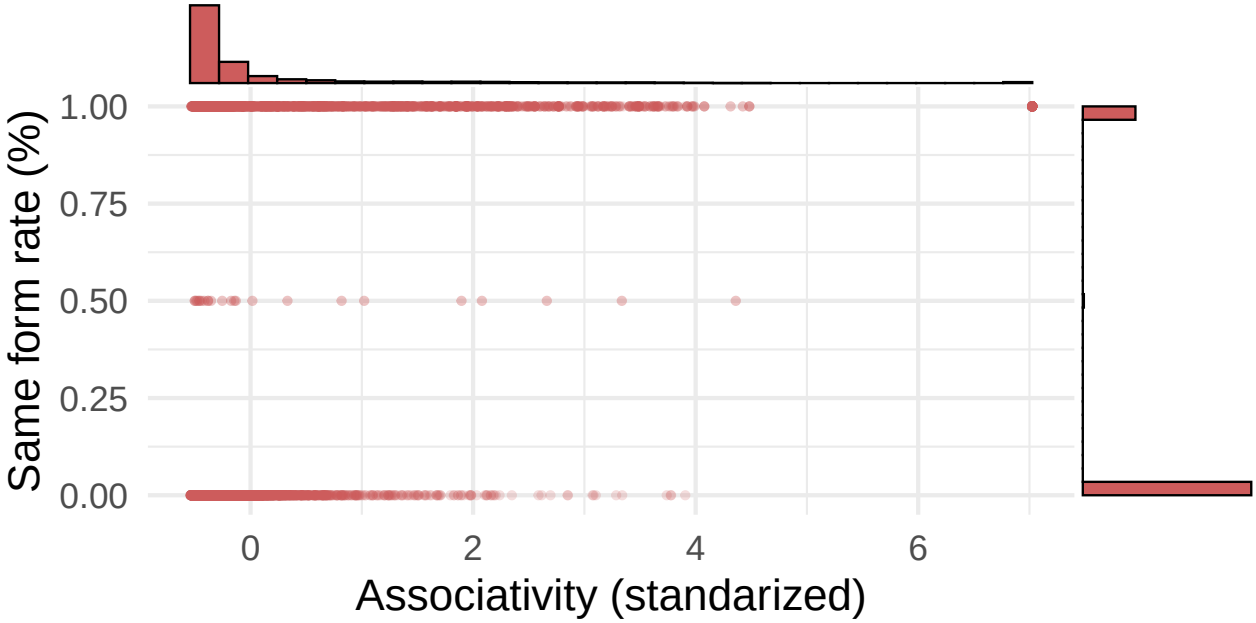

Figure 1: Associativity against rate by which two meanings are expressed by the same form across languages (only 12000 data points depicted).

Discussion / Potential issues Apart from a class imbalance (see above), we did not control for phylogenetic relatedness nor for geographic proximity. It may be that languages that are closer to each other, either geographically or linguistically, (e.g., Spanish and Catalan) may influence each other, and this fact is not modelled but could affect the results. Accordingly, while these results corroborate Xu et al.'s finding, the estimate is likely to change if these factors are appropriately modelled.

Division of labor

- FirstName1 SurName1 pre-processed and analyzed the data
- FirstName2 SurName2 collected the data; did the literature review; and wrote the documentation
- FirstName1 SurName1 and FirstName2 SurName2 both designed the analysis, analyzed the data, and wrote the report

References

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.