HOW MUCH DO WE AGREE ON WHAT WORDS MEAN?

GARY LUPYAN[*1], STEVEN PIANTADOSI[2], LOUIS MARTÍ[2], CELESTE KIDD[2], ELLA LIU[1], JEROEN VAN PARIDON[1] and YANCHAO BI[3]

[*]Corresponding Author: lupyan@wisc.edu

[1]Dept of Psychology, University of Wisconsin-Madison, USA [2]Dept of Psychology, UC Berkeley, USA [3]IDG/McGovern Institute for Brain Research, Beijing Normal University, China

At the core of every language is a vocabulary—a set of building blocks from which language users can construct arbitrarily complex meanings. Vocabularies are, of course, learned. Members of a speech community must learn the form of each word—its phonology—but also, critically, its meaning. Where do word meanings come from? And how similar are the meanings of the same word in different people within a speech community? These questions are of central importance to the study of language evolution for at least three reasons. First, understanding the emergence of vocabularies requires that we understand the extent to which words demarcate cognitively privileged categories (i.e., words label our concepts, e.g., Snedeker & Gleitman, 2004), as opposed to demarcating categories that are shaped by communicative needs and history—categories that might not be learned in the absence of labels (words help create concepts, e.g., Lupyan et al., 2007; Thompson et al., 2020). Second, for languages to function as effective communicative systems, it is generally thought that people must closely agree on what words mean (e.g., Hutchins & Hazlehurst, 2006). If word meaning variability within a speech community is pervasive, it would raise the question of how much agreement is really necessary for effective communication. Third, if word meaning variability is indeed pervasive, how can we use language to coordinate as effectively as we seem to? When we talk, misunderstandings seem to be more the exception than the norm. Here, we bring together several sources of recent evidence collected in our respective labs and observed in published word norms that suggest that within a speech community, differences in meanings for everyday words may be more extensive than previously realized (cf. Clark, 1998).

There is no gold standard for quantifying and comparing word meanings, and so we sought to estimate variability in a few different ways: (1) Sorting, e.g., sort these words according to how similar they are to one another. (2) Similarity judgments, e.g., is a penguin more similar to a whale or a seal? (3) Elicitation and endorsement of word meanings, e.g., what does "energy" mean? Which of these meanings of "energy" best approximates your meaning? (4) Comparison of people's judgments of semantic dimensions such as concreteness. (5) Beyond

explicit behaviors, we can also compare (using fMRI) neural representations elicited by the same word in different people.

Each measure revealed clear similarities, e.g., people clustered body parts together and separately from animals, clustered positive emotions together and separately from negative emotions. In general, agreement for concrete words was greater than for abstract words. But substantial variability existed even for concrete words. Fig. 1A shows an example of how two people sorted a group of common animals. Sorting-based correlations of such concrete words rarely exceed r=.5 (Wang & Bi, 2021). We also saw substantial variability when using similarity probes, e.g., when asked whether a seal is more similar to a penguin or a whale, 44% of people chose "whale". Interestingly, people were largely unaware that judgments like these produce divergent responses, believing a large majority will respond as they themselves did. Those who responded with "whale" thought that 75% would do the same (Martí et al., 2021). Fig. 1B shows endorsement patterns of various senses of "energy" (generated by a separate set of participants), revealing three distinct profiles. Fig. 1C shows concreteness ratings (Brysbaert et al., 2014). While many words show expected unimodal distributions, many others show clear bimodality hinting at systematic differences in how different people construe these words.

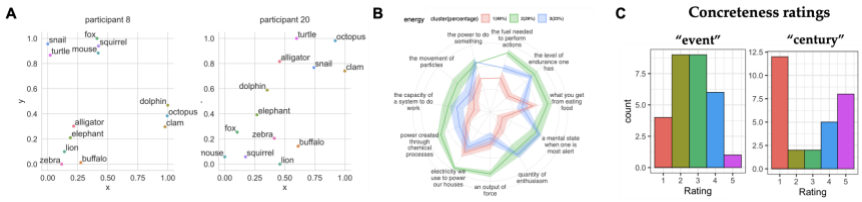

Figure 1. (A) Two solutions on an animal-sorting task. (B) Results from a meaning elicitation and endorsement task for the word "energy". (C) Distributions of concreteness ratings for two example words (rating of 1 corresponds to maximally abstract).

Do results such as these indicate true variability? If so, where do these differences come from? We consider several possibilities including different learning biases, different sensorimotor experiences, and different linguistic experiences. How do people communicate in the presence of these differences? We will discuss three possibilities: (1) These differences have no consequences for everyday communication, only showing up in specific contexts. (2) Misalignments are quickly repaired (Healey et al., 2018) or compensated through pragmatic inference. (3) Consequential errors in communication are more common than generally acknowledged.

References

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Clark, H. H. (1998). Communal Lexicons. In K. Malmkjær & J. Williams (Eds.), *Context in Language Learning and Language Understanding* (p. 63). Cambridge University Press.

Healey, P. G. T., Mills, G. J., Eshghi, A., & Howes, C. (2018). Running Repairs: Coordinating Meaning in Dialogue. *Topics in Cognitive Science*, *10*(2), 367–388. https://doi.org/10.1111/tops.12336

Hutchins, E., & Hazlehurst, B. (2006). *How to invent a lexicon: The development of shared symbols in interaction*. https://doi.org/10.4324/9780203993699-17

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: Labels facilitate learning of novel categories. *Psychological Science*, *18*(12), 1077–1082.

Martí, L., Wu, S., Piantadosi, S. T., & Kidd, C. (2021). *Latent diversity in human concepts*. PsyArXiv. https://doi.org/10.31234/osf.io/fb8s2

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

Thompson, B., Roberts, S. G., & Lupyan, G. (2020). Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour*, *4*, 1–10. https://doi.org/10.1038/s41562-020-0924-8

Wang, X., & Bi, Y. (2021). Idiosyncratic Tower of Babel: Individual Differences in Word-Meaning Representation Increase as Word Abstractness Increases. *Psychological Science*, 09567976211003877. https://doi.org/10.1177/09567976211003877