VERBAL LABELS INCREASE CONCEPTUAL ALIGNMENT

ELLISE SUFFILL[1], JEROEN VAN PARIDON[1], and GARY LUPYAN*[1]

*Corresponding Author: lupyan@wisc.edu
[1]Dept of Psychology. University of Wisconsin-Madison, USA

Successful linguistic communication requires conversants to mean at least approximately the same thing by the same words. But how is this alignment achieved? One possibility is that participants have pre-existing concepts to which verbal labels are mapped. Alignment is then a matter of ensuring that in members of the same speech community, the same word points to the same concept. But how do the underlying conceptual representations become aligned in the first place? One source of alignment is shared sensory experiences mediated by similar perceptual systems. But is this enough? We test the possibility that language itself serves to align conceptual representations. Participants were asked to sort novel shapes and we measured the similarity between people's sorts. By separately manipulating previous perceptual experience with the shapes, and exposure to (entirely redundant) category labels, we tested (1) the role of shared perceptual experience and (2) the effect of labels on representational alignment. The results showed that shared experience with labels increased representational alignment more than shared perceptual experience alone. We consider the implications of this finding for the cognitive functions of language and for how language may be used to enable coordination in the face of non-shared perceptual experiences.

Introduction

The idea that it is possible to communicate through direct transfer of mental states between minds has been a popular trope in science fiction for over a century. It has now has become the subject of empirical investigation, e.g., Rao et al. ask "Can information that is available in the brain be transferred directly in the form of the neural code, bypassing language altogether?" (2014). A common thread in science fiction treatments of telepathy and its modern revival is that natural language obscures communication because it is ambiguous, imprecise, and slow. For example, the computer scientist Yan LeCun, the chief AI scientist at Facebook, recently asserted that "Language is an imperfect, incomplete, and low-bandwidth serialization protocol for the internal data structures we call thoughts" (LeCun, 2021). Bypassing language is hence seen as a way of improving, or at least speeding up, communication.

The work we describe here is motivated by challenging a core assumption behind the telepathy trope: that our mental states are naturally aligned such that one person's thought is syntactically and semantically homologous to another person's. In the absence of this prior such alignment, transferring neural patterns that constitute mental states between people—even if technologically possible—would not lead to successful communication. What is at stake is important to the study of language evolution because understanding the evolution of a trait is greatly helped by understanding the range of its functions (Griffiths, 1993) and work on the functions of language are curiously under-represented in the study of language evolution (cf. Dessalles, 2007).

Conceptual alignment and language: what is the connection?

Everyday communication seems to require conceptual alignment (e.g., Pickering & Garrod, 2021). When one person says "Pass the salt please" and another person passes them the salt, the two have achieved some amount of alignment: one person's request was successfully represented by the other. But where does this alignment comes from and what role, if any, natural language plays in establishing (rather than obscuring) it.

The idea that linguistic communication is possible only because our thoughts are already sufficiently aligned is a basic premise of philosophical positions such as Fodor's language of thought (Fodor, 1975). It is also a common starting point in theories of language learning that view children as mapping words onto pre-existing (and largely shared) concepts (Bloom, 2002; Pinker, 1994; Snedeker & Gleitman, 2004).

But another possibility is that alignment is achieved—in part—*through* language itself (e.g., Casasanto & Lupyan, 2014; Dingemanse, 2017; Gomila, 2011; Lupyan & Bergen, 2016). On this view, learning and using the syntax and semantics of a natural language helps people to structure and convey their thoughts in ways that are (more or less) understandable to others. Rather than just a device for conveying our thoughts, language provides an interface between minds (e.g., Clark, 1998; Gentner & Goldin-Meadow, 2003; Lupyan, 2012). If true, then telepathy could in principle be possible by transferring neural representations of words from one person to another, but such a scheme would not be the language-bypassing telepathy we've been promised, but rather an over-engineered form of texting.

The idea that language may play a causal role in promoting conceptual alignment is supported by several lines of evidence, some circumstantial, others more direct. First, there is the simple fact of substantial cross-linguistic diversity

in all aspects of language (Evans & Levinson, 2009). If our conceptual representations were naturally aligned—either due to our shared biology, shared environment, shared goals, or all three—one might expect lexicons to show more similarity than they do. And although it is clear that the lexical systems of natural languages occupy a small space of all possible systems (e.g., Zaslavsky et al., 2018), it is striking that finding universal basic units of linguistic meaning has been so difficult. Even in the domain of perception, where one might find vocabularies to be most constrained by shared biology, one finds tremendous diversity of naming schemes (Majid, 2020; Majid et al., 2018). Diversity within a language is smaller (Forder & Lupyan, 2017) although it can depend on the measure one uses (Kuehni, 2004). Second, experimental evidence suggests that verbal labels can increase conceptual alignment across people, in both communicative (e.g., Markman & Makin, 1998) and non-communicative contexts (Suffill et al., 2016, 2019).

Current study

Here we tested a strong version of the prediction that language promotes conceptual alignment. We exposed people to novel shapes grouped into two distinct categories (Figure 1). We then computed alignment among participants assigned to each of group using a sorting task. Our design allowed us to compare how conceptual alignment is affected by shared labels compared to alignment achieved through shaped perceptual experiences. Using materials with a clear pre-existing category structure allowed us to test whether labels help to align categories even when the existence of the categories is made plain by perceptual discontinuities (Fig. 1). This makes the current experiment substantially different from work examining the ways that labels can help mark distinctions in perceptually equidistant continua such as colors (Davidoff, 2001) and shapes (Plunkett et al., 2008) as well as from past work showing that labels facilitate the learning of new categories (Lupyan et al., 2007).

Methods

Participants

We recruited 129 (85 female, ages 18-22) psychology students from University of Wisconsin-Madison. Participants were randomly assigned to a *Baseline* (N = 45), *No Labels* (N = 43) or *With Labels* (N = 41) condition.

Materials

We constructed two visual family-resemblance categories designed to be easy to distinguish and difficult to name. We began by generating two prototype shapes by creating a random collection of points and connecting them with a spline (Fig 1A, 1B). We then generated 18 exemplars per category by perturbing the points and fitting a new spline, creating low, medium, or high distortions (e.g., Fig. 1 bottom).

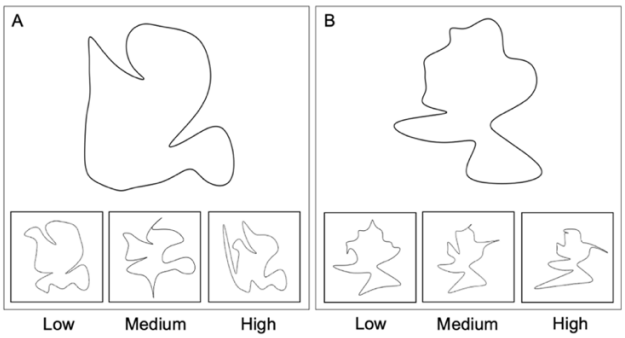

Figure 1. Category A (left) and B (right) prototypes with examples of "low", "medium" and "high" category exemplars.

Procedure.

Pre-exposure. Participants assigned to the *With Labels* or *No Labels* conditions began with a match-to-sample task designed to familiarize participants with the stimuli and have them repeatedly contrast within-category and between-category stimuli. On each trial, participants saw one of the shapes (the *standard*) for 1 sec followed by a 1 sec blank screen. Two shapes then appeared side by side. One of these was identical to the standard (*target*) and the other was a *foil*—a shape from the contrasting category. Participants had to choose which of the two shapes exactly matched the standard, i.e., on each trial they had to choose the target and not the foil. In the *With Labels* condition, the standard was presented with its corresponding nonsense label; half the *With Labels* participants heard category *A* shapes labeled as "a talp"; half heard them labeled as "a gek". Notice that the label is completely unnecessary for making a correct response and is therefore informationally redundant. Errors were signaled with a short buzzing sound. Participants assigned to the *Baseline* condition did not complete this phase and proceeded directly to the free sort.

Free sort. We quantified alignment by measuring how participants in the three conditions arranged the shapes in a free sort—a common method for

assessing people's conceptual representations (Goldstone, 1994; Kriegeskorte et al., 2008; Malt et al., 1999). Participants were shown 20 shapes (10 A shapes and 10 B shapes) arranged around the perimeter of the screen. These included 3 previously seen exemplars, 6 novel exemplars, and the previously unseen prototype. Participants were asked to cluster the shapes together in way that made sense to them, creating as many or as few clusters as they needed.

Analytic approach

We computed alignment between people's item arrangements as follows: For each participant, we take the pairwise distances between all item pairs (20*19/2 = 190). We then compute the rank correlations between that participant's pairwise item distances and the pairwise item distances of the other participants in the same condition. The Fisher's z-transformed mean of these correlations represents the participant's average alignment to other participants. These are the values shown in Fig 2A. To statistically compare the groups in an unbiased way we counted each participant pairing as a single observation, but attributed the variance associated with this observation to both participants in the pairing using *lmerMultiMember* (van Paridon et al., 2022), an R package that allows for specifying multiple membership random effects. In addition, we computed for each participant a measure of *categoricality*, the median Euclidean distance between exemplars from different categories (e.g., A_1 and B_2) minus the median distance between exemplars from the same category (e.g., A_1 and A_2).

Results

Pre-exposure. Accuracy on the delayed match-to-sample task was nearly identical for the *No Labels* (M = 0.98) and *With Labels* groups (M= 0.98). Given the task's simplicity, this was expected, and confirms that the categories were trivially easy to distinguish, regardless of labels.

Categoricality. Participants in all groups grouped within-category items closer together than between-category items: $\text{Categoricality}_{\text{baseline}}$=151 pixels, $\text{Categoricality}_{\text{no-labels}}$=171 pixels, $\text{Categoricality}_{\text{with-labels}}$=264 pixels. All three values were significantly greater than 0, t's> 5, p<.0001, confirming that even

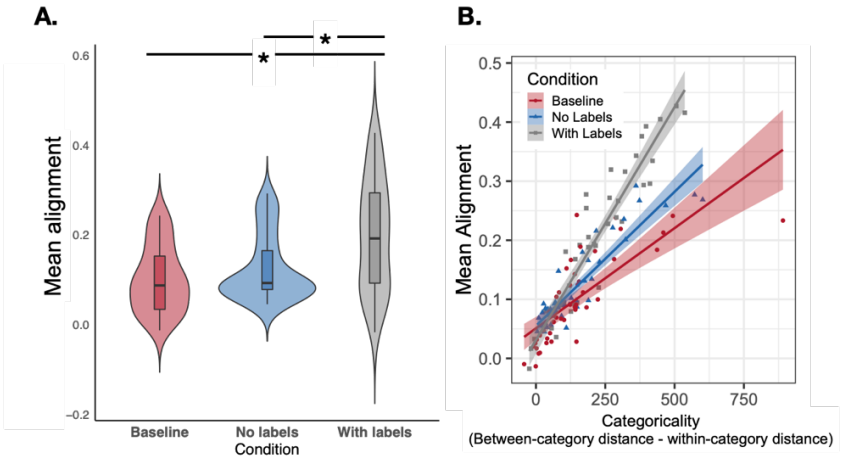

Figure 2. A. Mean alignment for the three tested conditions. B. The relationship between alignment and categoricality for each condition.

Baseline participants—who had no prior experience with seeing or contrasting the shapes—were still sensitive to the designed category structure. Categoricality was not significantly different between the *Baseline* and *No Labels* groups, $t<1$. In contrast, participants exposed to the nonsense labels produced more categorical sorts than those who had identical experience seeing and contrasting the shapes, but without being exposed to their names ($b=93$, $t=2.4$, $p=.02$).

Alignment. Average alignment for each group is shown in Fig. 2A. Those in the *Baseline* condition were as similar to one another in their shape arrangements as those who encountered the shapes several hundred times, but without the accompanying labels ($b=.02$, $t=1.1$, $p=.24$). In contrast, participants who were exposed to category labels, sorted shapes more similarly to one another than those in the *Baseline* condition ($b=.09$, $t=4.9$, $p<.0001$) and those in the *No Labels* condition ($b=.07$, $t=3.7$, $p=.0004$). The increase in alignment caused by labels was significantly greater than that caused by shared perceptual experiences ($t=2.27$, $p=.03$).

Relationship between categoricality and alignment. Why did labels increase alignment? One possibility is that alignment was mediated by categoricality: labels increased categoricality—leading to an increase in between category distance and a decrease in within-category, and these more tightly-clustered sorts were more aligned. As shown in Fig. 2B, there was indeed a strong relationship between categoricality and alignment (overall $r=.83$, $p<.0001$). And yet, categoricality ($b=.08$, $t=21.9$, $p<.0001$) and condition ($b=.02$, $t=6.4$, $p<.0001$) accounted for unique variance in predicting alignment; including categoricality

made the effect of condition on alignment even stronger with the two predictors accounting for 83% of the variance. Fig. 2B also makes clear that there was a condition-by-categoricality interaction (t=7.71, p<.0001). For the same level of categoricality, exposure to labels yielded greater alignment. In short, labels increased both alignment and categoricality, but there was no evidence of mediation of alignment by categoricality.

General Discussion

Language allows one to activate thoughts, old and new, in other people. The promise of telepathy—a direct exchange of mental states that bypasses natural language—is predicated on the assumption that thoughts are entirely independent of language; language is merely a medium by which the thoughts are transmitted. This assumption, however, may be wrong. The study we describe here provides a very limited, but nevertheless strong test of the hypothesis that even very stripped-down forms of language—redundant and seemingly uninformative verbal category labels—can increase conceptual alignment and do so to a greater extent than shared perceptual experiences alone. Our finding that verbal labels increase conceptual alignment is just one result using specific stimuli and task. Our hope is that future investigations can map out the generality of this result and the mechanisms by which labels achieve this effect.

The technology to transfer mental states may one day exist. Will it enable telepathy? The present results offer an early hint that however "imperfect and incomplete" language may be, attempts to bypass it may lead to a semantic disconnect and communicative failure. It may be possible to devise a system for re-aligning our thoughts into a mutually understandable form. Natural language is *just such a system* and the role natural language plays in aligning our thoughts may be another piece in the puzzle of its evolution.

Acknowledgements

This work was supported by NSF-PAC Awards 2020969 to G.L.

References

Bloom, P. (2002). *How Children Learn the Meanings of Words*. MIT Press.

Casasanto, D., & Lupyan, G. (2014). All Concepts are Ad Hoc Concepts. In E. Margolis & S. Laurence (Eds.), *Concepts: New Directions* (pp. 543–566). MIT Press.

Clark, A. (1998). Magic words: How language augments human computation. In P. Carruthers & J. Boucher (Eds.), *Language and Thought: Interdisciplinary themes* (pp. 162–183). Cambridge University Press.

Davidoff, J. (2001). Language and perceptual categorization. *Trends in Cognitive Sciences*, *5*(9), 382–387.

Dessalles, J.-L. (2007). *Why We Talk: The Evolutionary Origins of Language* (J. Grieve, Trans.). Oxford University Press, USA.
Dingemanse, M. (2017). On Brain-to-Brain Interfaces, Distributed Agency and Language. In N. J. Enfield & P. Kockelman (Eds.), *Distributed Agency* (pp. 59–66). Oxford University Press.
Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, *32*(05), 429.
Fodor, J. A. (1975). *The language of thought* (Vol. 5). Harvard university press.
Forder, L., & Lupyan, G. (2017). Facilitation of color discrimination by verbal and visual cues. *Meeting of the Vision Sciences Society*.
Gentner, D., & Goldin-Meadow, S. (2003). *Language in mind: Advances in the study of language and thought*. MIT Press.
Goldstone, R. (1994). An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers*, *26*(4), 381–386.
Gomila, T. (2011). *Verbal Minds: Language and the Architecture of Cognition* (1 edition). Elsevier.
Griffiths, P. E. (1993). Functional Analysis and Proper Functions. *The British Journal for the Philosophy of Science*, *44*(3), 409–422.
Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational Similarity Analysis—Connecting the Branches of Systems Neuroscience. *Frontiers in Systems Neuroscience*, *2*.
Kuehni, R. G. (2004). Variability in Unique Hue Selection: A Surprising Phenomenon. *Color Research & Application*, *29*(2), 158–162. https://doi.org/10.1002/col.10237
LeCun, Y. (2021, March 6). Language is an imperfect, incomplete, and low-bandwidth serialization protocol for the internal data structures we call thoughts. [Tweet]. *@ylecun*. https://twitter.com/ylecun/status/1368208931265388554
Lupyan, G. (2012). Linguistically modulated perception and cognition: The label-feedback hypothesis. *Frontiers in Cognition*, *3*(54). https://doi.org/10.3389/fpsyg.2012.00054
Lupyan, G., & Bergen, B. (2016). How Language Programs the Mind. *Topics in Cognitive Science*, *8*(2), 408–424. https://doi.org/10.1111/tops.12155
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.
Majid, A. (2020). Human Olfaction at the Intersection of Language, Culture, and Biology. *Trends in Cognitive Sciences*.
Majid, A., Roberts, S. G., Cilissen, L., Emmorey, K., Nicodemus, B., O'Grady, L., Woll, B., LeLan, … Levinson, S. C. (2018). Differential coding of perception in the world's languages. *Proceedings of the National Academy of Sciences*, *115*(45), 11369–11376.
Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language*, *40*(2), 230–262.
Markman, A. B., & Makin, V. S. (1998). Referential communication and category acquisition. *Journal of Experimental Psychology: General*, *127*(4), 331.
Pickering, M. J., & Garrod, S. (2021). *Understanding dialogue: Language use and social interaction*. Cambridge University Press.
Pinker, S. (1994). *The language instinct. William Morrow & Co*.
Plunkett, K., Hu, J.-F., & Cohen, L. B. (2008). Labels can override perceptual categories in early infancy. *Cognition*, *106*(2), 665–681.
Rao, R. P. N., Stocco, A., Bryan, M., Sarma, D., Youngquist, T. M., Wu, J., & Prat, C. S. (2014). A Direct Brain-to-Brain Interface in Humans. *PLoS ONE*, *9*(11).
Snedeker, J., & Gleitman, L. (2004). Why it is hard to label our concepts. *Weaving a Lexicon*, *257294*.
Suffill, E., Branigan, H., & Pickering, M. (2016). *When the Words Don't Matter: Arbitrary labels improve categorical alignment through the anchoring of categories*. Proceedings of the 38th Meeting of the Cognitive Science Society, 1715–1720.
Suffill, E., Branigan, H., & Pickering, M. (2019). Novel labels increase category coherence, but only when people have the goal to coordinate. *Cognitive Science*, *43*(11), e12796.
van Paridon, J., Bolker, B., & Alday, P. (2022). *lmerMultiMember: Multiple membership random effects for mixed effects models in lme4* [R]. https://github.com/jvparidon/lmerMultiMember
Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, *115*(31), 7937–7942.