“Un-reject-able-ish”: Learning and generalization of novel compositional meanings

Xiaochen Zheng
Donders Institute
Radboud University
x.zheng@donders.ru.nl

Mona Garvert
MPI for Human Cognitive and
Brain Sciences
mona.garvert@gmail.com

Jonne Roelofs
MA linguistics
Radboud University
j.roelofs@student.ru.nl

Hanneke den Ouden
Donders Institute
Radboud University
h.denouden@donders.ru.nl

Roshan Cools
Donders Institute
Radboudumc
r.cools@donders.ru.nl

Abstract

The ability to generalize previously learned information to novel situations is fundamental for adaptive behavior. When seeing the word “un-reject-able-ish” for the first time, one can quickly infer its meaning by generalizing the knowledge of its constituent parts and integrating them based on certain abstract structural rules (e.g., the sequential order of the word parts). How do we generate novel, compositional meaning? What are the neuro-computational mechanisms that underlie structural inference in not only meaning generalization but also across different cognitive domains? This efficient but also flexible inferential process may leverage neural mechanisms commonly studied in the nonlinguistic domains of action planning, relational memory and model-based reinforcement learning, including medial prefrontal-hippocampal circuitry.

To address these questions, we developed a novel experimental paradigm for quantifying novel structural inference for the generation of word meaning. We taught participants compositional words from an artificial language and tested them with novel words using a semantic priming task. Results from two behavioral experiments showed that participants can learn and generalize structural (sequential order) rules for inferring novel, compositional meanings on the fly. An ongoing neuroimaging study in which we combine this paradigm with fMRI adaptation will unravel the neural mechanisms of meaning composition, allow us to test the prediction that correct compositional inference can be predicted from neural activity in a medial prefrontal-hippocampal network, measured during the generation of the novel word meaning.

Keywords: structural inference, abstract rule learning, compositionality, prefrontal-hippocampal network

Acknowledgements

This research is supported by the Language in Interaction consortium (Gravitation Grant 024.001.006 funded by the Dutch Research Council). We thank Andrea E. Martin for stimulating discussions.

1 Background

The ability to generalize previously learned information to novel situations is fundamental for adaptive behavior in an uncertain world. We are good at combining linguistic "building blocks" and inferring the meaning of these combinations on the fly. When we see the word "un-rejectable-ish" for the first time, we can quickly infer its meaning by generalizing our knowledge of the constituent morphemes and integrating them. Relational structure plays an essential role in linguistic composition. For instance, the two sentences "The cat chased the mouse." and "The mouse chased the cat." have identical linguistic building blocks but different meanings. The human ability to compose complex representations from basic building blocks depends critically on abstract, generalizable knowledge and certainly goes beyond language (Sablé-Meyer et al., 2021; Schwartenbeck et al., 2021; Roumi et al., 2021), facilitating learning and problem solving in many cognitive domains (Behrens et al., 2018; Liu et al., 2019). Recent imaging work has shown an essential role of the prefrontal-hippocampal circuitry in the construction of complex visual configurations using simple building blocks (Schwartenbeck et al., 2021).

To explore the cognitive process of structural inference for meaning generalization and its neural codes, we have developed a novel experimental paradigm that utilizes the linguistic phenomenon of "affixation", i.e., forming different words by adding morphemes at the beginning (prefix) or the end (suffix) of words. We taught participants artificial, compositional words with different rules of affixation, and later tested them on novel, compositional words that follow the same structural (i.e., sequential order) rules.

2 Experiment 1

We collected online behavioral data from 36 participants. In the learning phase (Figure 1), we taught them artificial, compositional pseudo-words consisting of a known stem (e.g., "good" in "good-kla") and an unknown affix (e.g., "kla"). Crucially, the affix alters the word meaning depending on its position (e.g., "-kla" as a suffix means "the opposite", whereas "kla-" as a prefix means "young version", see "KEY" in Figure 1). These position-dependent affix meanings lead to unique compositional meanings in different sequential combinations with the stems (e.g., "good-kla" means "bad", whereas "kla-human" means "child"). In the testing phase, we presented participants with a new set of compositional pseudo-words (i.e., not presenting in the learning phase). The pseudo-words were manipulated using the same structural rules in the learning phase. The affixes attached to the stems can either be congruent in sequential order (e.g., "rich-kla", where "-kla" means "opposite"), or incongruent in order (e.g., "kla-rich", where "-kla" but not "kla-" means "opposite"), or in a totally mismatched meaning regardless of order (e.g., "rich-ran/ran-rich", where neither "-ran" or "ran-" means "opposite"). We measured participants' reaction times when they made semantic decisions on the target synonym word, following the prime pseudo-words in the three conditions (Figure 1). To ensure that participants attended to the prime words, we included on 10% of trials a probe question in which the meaning of the preceding two words need to be compared. As a result of semantic priming, we expected participants to respond faster on a target word when the preceding compositional word carries the same meaning.

Results (Figure 2A) showed that participants were faster in making semantic decisions when the target word was primed by a congruent ($\beta = 0.04$, SE = 0.01, $z = 3.97$, $p < .001$) or incongruent ($\beta = 0.03$, SE = 0.01, $z = 2.84$, $p = .01$) pseudo-word, compared with a mismatched word. This finding provides evidence that people are able to compute novel compositional meaning online. However, there was no difference in the priming effect between congruent vs. incongruent conditions ($\beta = 0.01$, SE = 0.01, $z = 0.99$, $p = .58$). This lack of an effect of congruence suggests that sequential order did not matter. Nevertheless, when explicitly asked whether the prime matches the target, half of the participants indicated that the congruent compositional primes matched the targets in meaning, whereas the incongruent primes did not (Figure 2B), suggesting they consider a different meaning of the same affix form based on its sequential order. The discrepancy between the online priming task and the offline posttest could have two reasons: (1) Participants did not learn well enough to make the inference process automatic; (2) Participants did not have enough time to process the novel prime words online.

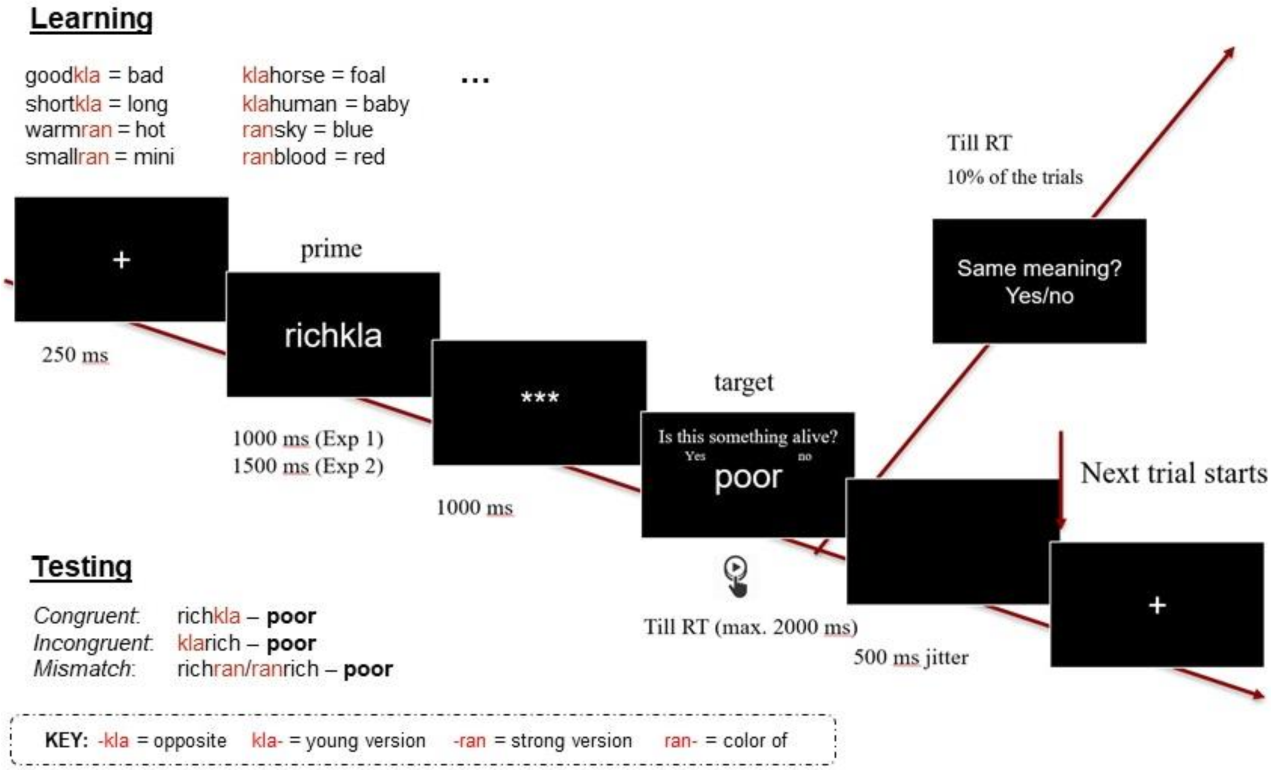

Figure 1. Experimental paradigm. The actual stimuli was in participants' native language.

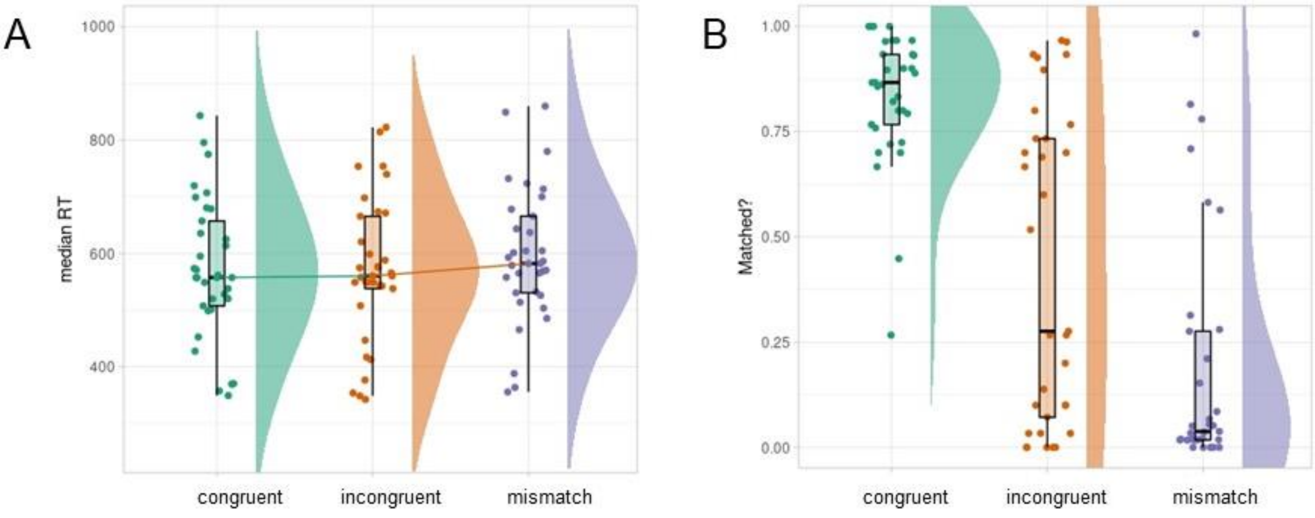

Figure 2. Main findings from Experiment 1. (A) Raincloud plots of median reaction times of the semantic priming task across three experimental conditions. The outer shapes represent the distribution of the data over participants, the thick horizontal line inside the box indicates the group median, and the bottom and top of the box indicate the group-level first and third quartiles of each condition. Each dot represents one participant. (B). Raincloud plots of participants' responses on whether the prime words match the target words in meaning.

3 Experiment 2

The second experiment was adapted from experiment 1 in the following ways: (1) prolonged presentation of the prime word to ensure sufficient processing time; (2) included another block of learning (in total four). We also doubled the number of participants (N = 72) to ensure a well-powered between-subject comparison based on participants' learning strategies, i.e., whether they took into account the sequential order rule.

Results showed again a priming effect of the congruent (β = 0.05, SE = 0.01, z = 6.91, p < .001) or incongruent condition (β = 0.03, SE = 0.01, z = 3.89, p < .001) compared with the mismatch condition, replicating Experiment 1. Moreover, the semantic priming effect was larger following a congruent than an incongruent prime word (β = 0.02, SE = 0.01, z = 2.67, p = .02, Figure 3A), suggesting that the identical affix forms in the compositional words nonetheless lead to different compositional meaning given their different sequential order (i.e., pre vs. post). This supports the idea that sequential order plays an essential role in meaning composition. Similar to Experiment 1, more than half of the participants explicitly reported the incongruent primes to *not* match the meaning of the targets. We explored the two types of learning strategies by splitting the participants into two groups based on the posttest (Figure 3B): those who did not consider sequential order ("BLENDer", N = 24), and those who did ("BUILDer", N = 47). The BLENDers showed no priming difference between the congruent and incongruent conditions (β = 0.002, SE = 0.01, z = 0.20, p = .978), whereas the BUILDers showed a larger priming effect of the congruent than incongruent ones (β = 0.03, SE = 0.01, z = 3.04, p = .007), as they could not compute the meaning of the latter (Figure 3C).

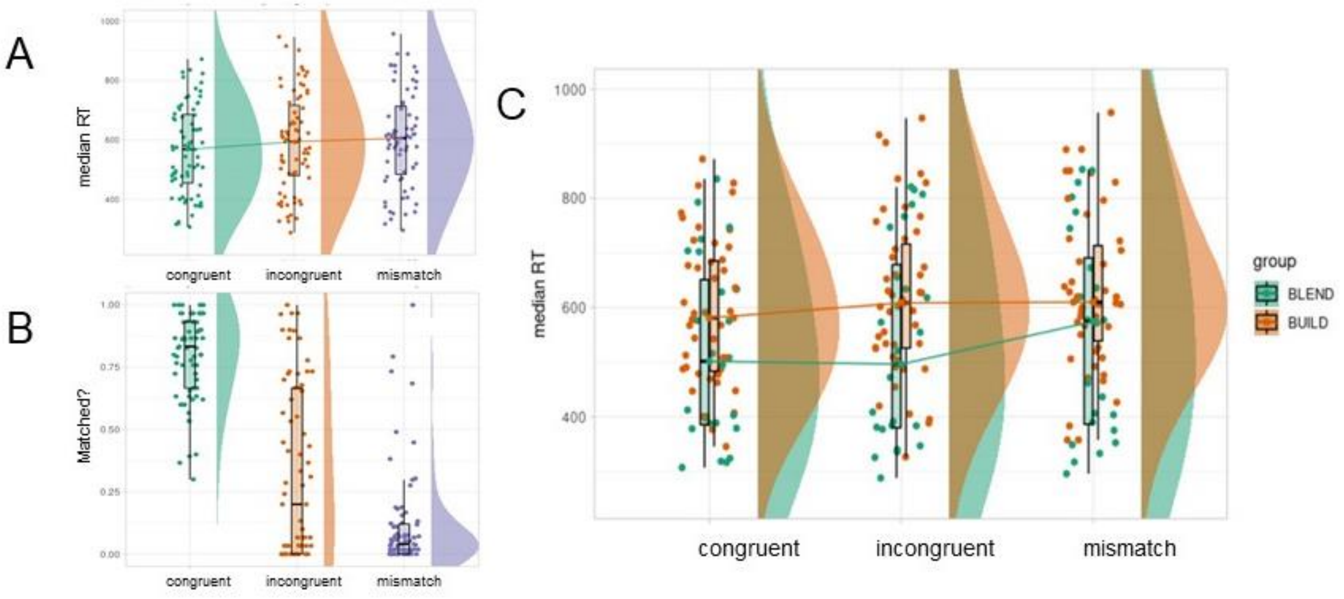

Figure 3. Main findings from Experiment 2. (A) Raincloud plots of median reaction times of the semantic priming task across three experimental conditions. (B) Raincloud plots of participants' responses on whether the prime words match the target words in meaning. (C). Raincloud plots of median reaction times of the semantic priming task across three

experimental conditions, split between two participant groups.

4 Conclusion and Ongoing work

We have developed a paradigm that allows us to quantify the ability to infer and represent novel compositional word meaning, based on the learning and generalization of sequential order rules. To assess the neural mechanisms of compositional inference and generalization, we have adapted the paradigm for use in an fMRI adaptation study. We will test our prediction that correct compositional inference can be predicted from neural activity in a medial prefrontal-hippocampal network, given the hippocampal function of representing learned and inferred structural relationships for inference and generalization (Barron et al., 2020; Behrens et al., 2018; Bellmund et al., 2018; Garvert, Dolan, & Behrens, 2017) and medial prefrontal cortex's role in abstracting and generalizing across structures and constructs novel experience (Baram et al., 2021; Barron, Dolan, & Behrens, 2013; Garvert et al., in prep).

References

Baram, A. B., Muller, T. H., Nili, H., Garvert, M. M., & Behrens, T. E. J. (2021). Entorhinal and ventromedial prefrontal cortices abstract and generalize the structure of reinforcement learning problems. *Neuron*, 109, 713-723.e7. https://doi.org/10.1016/j.neuron.2020.11.024

Barron, H. C., Dolan, R. J., & Behrens, T. E. J. (2013). Online evaluation of novel choices by simultaneous representation of multiple memories. *Nature Neuroscience*, 16, 1492–1498. https://doi.org/10.1038/nn.3515

Barron, H. C., Reeve, H. M., Koolschijn, R. S., Perestenko, P. V., Anna Shpektor, Nili, H., Rothaermel, R., Campo-Urriza, N., O'Reilly, J. X., Bannerman, D. M., Behrens, T. E. J., & Dupret, D. (2020). Neuronal computation underlying inferential reasoning in humans and mice. *Cell*, 1–16. https://doi.org/10.1016/j.cell.2020.08.035

Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron*, 100, 490–509. https://doi.org/10.1016/j.neuron.2018.10.002

Bellmund, J. L. S., Gärdenfors, P., Moser, E. I., & Doeller, C. F. (2018). Navigating cognition: Spatial codes for human thinking. *Science*, 362. https://doi.org/10.1126/science.aat6766

Garvert, M. M., Dolan, R. J., & Behrens, T. E. J. (2017). A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *eLife*, 6, 1–20. https://doi.org/10.7554/eLife.17086

Liu, Y., Dolan, R. J., Kurth-Nelson, Z., & Behrens, T. E. J. (2019). Human replay spontaneously reorganizes experience. *Cell*, 178, 640-652.e14. https://doi.org/10.1016/j.cell.2019.06.012

Roumi, F. A., Marti, S., Wang, L., Amalric, M., & Dehaene, S. (2021). Mental compression of spatial sequences in human working memory using numerical and geometrical primitives. *Neuron*, 109, 2627-2639. https://doi.org/10.1016/j.neuron.2021.06.009

Sablé-Meyer, M., Ellis, K., Tenenbaum, J., & Dehaene, S. (2021). A language of thought for the mental representation of geometric shapes. *PsyArXiv*. https://doi.org/10.31234/osf.io/28mg4

Schwartenbeck, P., Baram, A., Liu, Y., Mark, S., Muller, T., Dolan, R., Botvinick, M., Kurth-Nelson, Z., & Behrens, T. (2021). Generative replay for compositional visual understanding in the prefrontal-hippocampal circuit. *BioRxiv*, 2021.06.06.447249. https://doi.org/10.1101/2021.06.06.447249