Neural Encoding of Continuous Word Meaning Likelihood During Semantic Disambiguation

Andrey Zyryanov (andrey.zyryanov@uni-tuebingen.de)
Centre for Integrative Neuroscience, University of Tübingen
Otfried-Müller-Straße 25, 72076 Tübingen, Germany

Graduate Training Centre of Neuroscience, University of Tübingen
Otfried-Müller-Straße 27, 72076 Tübingen, Germany

Victoria Pierz (victoria.pierz@student.uni-tuebingen.de)
Centre for Integrative Neuroscience, University of Tübingen
Otfried-Müller-Straße 25, 72076 Tübingen, Germany

Yulia Oganian (yulia.oganian@uni-tuebingen.de)
Centre for Integrative Neuroscience, University of Tübingen
Otfried-Müller-Straße 25, 72076 Tübingen, Germany

Abstract

Language comprehension hinges upon our ability to resolve semantic ambiguities, yet the neural representations underlying disambiguation remain unclear. To fill this gap, we recorded MEG while participants listened to German sentences containing an ambiguous target word, *Blatt* (meaning *paper* or *leaf*) or *Tor* (meaning *gate* or *goal*). In a behavioral pre-study, participants read these sentences and rated which target meaning was most likely. While group-averaged ratings showed that meaning likelihood varied continuously across sentences, single-trial ratings were categorically biased towards either meaning. To test whether the neural representation of meaning likelihood is categorical or continuous, we decoded the target word from neural activity and examined the effect of meaning likelihood on cross-meaning generalization. Around 800 ms *before* target onset, cross-meaning generalization was most accurate for neutral sentence contexts where target meanings were equally likely. Crucially, this improvement in generalization accuracy was parsimoniously modelled by a linear function of meaning likelihood. Thus, although explicit semantic judgments are distributed categorically, neural processing of ambiguous words reflects continuous meaning likelihood.

Keywords: semantics; language processing; decoding; neural representations

Context enables us to understand ambiguous words like 'fly' on-the-fly, but how? Psycholinguistic theories argue that context modulates activation strength within a set of categorical meaning representations (Duffy et al., 2001; Rodd, 2020). Neuroimaging evidence, however, questions the existence of categorical representations in the first place. Instead, it shows that word representations are linearly approximated by those of large language models (LLMs) and, therefore, are context-specific and continuous (Caucheteux & King, 2022; Goldstein et al., 2022). A simple linking hypothesis is that categorical meaning representations are coactivated proportionally to their contextual likelihood, yielding a summed activation pattern that linearly fits LLM representations. This hypothesis predicts that the neural representation of an ambiguous word in context changes continuously as a function of contextual meaning likelihood. Here, we test this prediction in a set of German sentences that constrained the meaning of semantically ambiguous words to different degrees.

Results

We constructed German sentences that contained one of two ambiguous target words, *Blatt* (meaning *paper* or *leaf; N* = 150) or *Tor* (meaning *gate* or *goal; N* = 150). In an online behavioral pre-study, participants (45 – 52 per sentence) read the sentences word-by-word and rated which target meaning was most likely on a continuous scale. Group-averaged ratings (Fig. 1a) covered the full range of meaning likelihood, including strongly biased and neutral sentences.

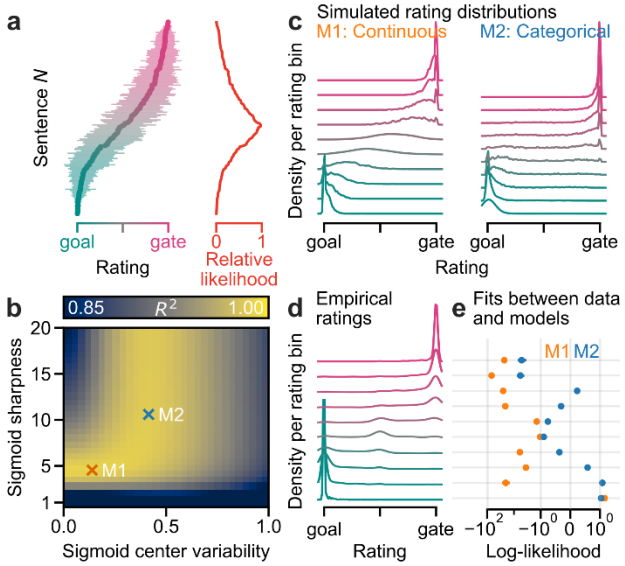

Figure 1: Semantic disambiguation behavior. **a** Average sentence ratings, mean ± SD, and relative meaning likelihood quantified as 1-|rating|. **b** R^2 between empirical and simulated average ratings. **c** Kernel density estimates (KDEs) of simulated single-trial ratings under two models. **d** KDEs of empirical single-trial ratings. **e** Average log-likelihood that empirical single-trial ratings are drawn from the model KDEs in *c*. Larger values indicate better fit.

Semantic Disambiguation Behavior is Categorical. The observed group-averaged ratings (Fig. 1a) could stem from single-trial ratings that either represent meaning likelihood continuously (i.e., capture its graded differences) or are categorically biased towards one meaning. To directly contrast these models, we simulated individual participants' ratings as sigmoid curves with varying sharpness (bias towards extremes) and center variability (mean bias). As expected, average empirical ratings were fit equally well

by simulations with a broad range of sharpness values (Fig. 1b), including such where simulated single-trial ratings clearly distinguished the continuous model from the categorical one (Fig. 1c). Crucially, empirical single trial ratings were fit significantly better by the categorical model than the continuous model (Fig. 1d, e; $t(19) = -2.48$, $p = 0.02$). Thus, single-trial ratings are largely insensitive to continuous meaning likelihood[1]; instead, they are categorically biased towards either meaning.

Neural Encoding of Meaning Likelihood is Continuous. We then asked whether the neural representation of meaning likelihood is categorical or continuous. Under both models, target representations are similar whenever its meanings are equally likely. Once a particular meaning becomes more likely (i.e., relative meaning likelihood decreases; red curve in Fig. 1a), the continuous representation model predicts a gradual decrease in representational similarity, whereas the categorical representation model predicts a sharp non-linear decrease.

To distinguish between these models, we decoded the target word (*Blatt* vs. *Tor*) from neural activity recorded with MEG while participants ($N = 19$) passively listened to the sentences. Data were split into non-overlapping training and test sets, each containing sentences biased towards one of the two possible meanings for each target word. Therefore, above-chance decoding generalization accuracy reflects the similarity of neural representations between the two alternative target meanings. Using cross-validated ridge regression, we then examined how single-trial generalization accuracy is influenced by meaning likelihood.

During target presentation, cross-meaning generalization was above chance (Fig. 2a, black curve). This is expected since the auditory word forms of each target are similar regardless of meaning. Furthermore, around 800 ms before target onset, cross-meaning generalization was more accurate whenever target meanings approached equal likelihood (Fig. 2a, red curve), as predicted by both models. Thus, the representation of target meaning is activated before the acoustic onset of the target word.

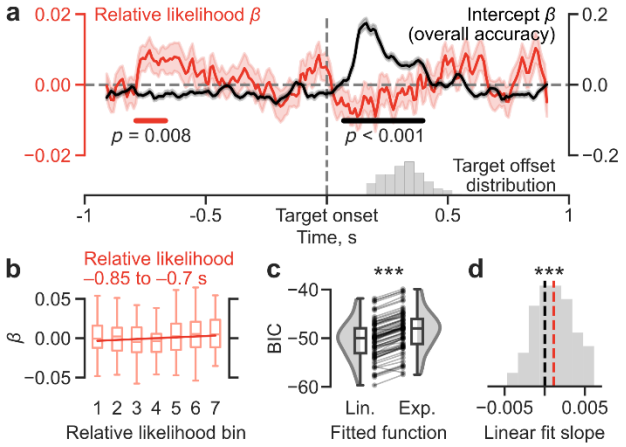

Figure 2: Neural encoding of meaning likelihood. **a** *β* coefficients from the model of cross-meaning generalization accuracy. **b** Cross-meaning generalization accuracy (*β* coefficients) per bin of relative meaning likelihood. Bold line shows an average linear fit. **c** Bayesian Information Criterion (BIC) of linear and exponential fits to the coefficients in *b*. Lower values indicate better fit. **d** Slopes of the linear fit in *c*.

Finally, to test whether generalization accuracy in this time window changes linearly (as expected under the continuous model) or exponentially (as expected under the categorical model), we estimated generalization accuracy separately for seven bins of relative meaning likelihood (~20 sentences per bin and participant; Fig. 2b). Linear function showed a significantly better fit to the binned generalization accuracy than an exponential one (Fig. 2c, $t(75) = -28.79$, $p < 0.0001$) and had a slope significantly above 0 (Fig. 2d, $t(75) = 3.99$, $p = 0.0002$). Thus, the improvement in generalization accuracy with larger relative meaning likelihood is best described by a linear function, in line with the continuous encoding model.

Conclusions

Although semantic judgments are largely insensitive to continuous meaning likelihood, the brain nonetheless encodes it shortly before we hear an ambiguous word in context. Specifically, its neural representation changes continuously as a function of contextual meaning likelihood. This might underlie the alignment between the representations of ambiguous words in LLMs and in humans.

[1] Of note, the empirical single-trial ratings also show a narrow peak around the neutral option. We are currently developing a model that includes a neutral category.

Acknowledgements

The authors thank Boehringer Ingelheim Fonds and the International Max Planck Research School for the Mechanisms of Mental Function and Dysfunction (IM-PRS-MMFD) for supporting Andrey Zyryanov.

References

Caucheteux, C., & King, J. R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*, 134. https://doi.org/10.1038/s42003-022-03036-1

Duffy, S. A., Kambe, G., & Rayner, K. (2001). The effect of prior disambiguating context on the comprehension of ambiguous words: Evidence from eye movements. In Gorfein, D. S. (Ed.), *On the consequences of meaning selection: Perspectives on resolving lexical ambiguity* (pp. 27–43). American Psychological Association. https://doi.org/10.1037/10459-002

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., … Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, *25*(3), 369–380. https://doi.org/10.1038/s41593-022-01026-4

Rodd, J. M. (2020). Settling Into Semantic Space: An Ambiguity-Focused Account of Word-Meaning Access. *Perspectives on Psychological Science*, *15*(2), 411–427. https://doi.org/10.1177/1745691619885860