The Representational Alignment between Humans and Language Models is implicitly driven by a concreteness effect

Cosimo Iaia
Goethe University Frankfurt
iaia@psych.uni-frankfurt.de

Bhavin Choksi
Goethe University Frankfurt
choksi@em.uni-frankfurt.de

Emily Wiebers
Goethe University Frankfurt

Gemma Roig
Goethe University Frankfurt
Center for Brains, Minds and Machines, MIT
Hessian.AI

Christian J. Fiebach
Goethe University Frankfurt
Brain Imaging Center

Abstract

The nouns of our language refer to either concrete entities (like a table) or abstract concepts (like justice or love). Cognitive psychology has established that concreteness influences how words are processed. Accordingly, understanding how concreteness is represented in our mind and brain is a central question in psychology, neuroscience, and computational linguistics. While the advent of powerful language models has allowed for quantitative inquiries into the nature of semantic representations, it remains largely underexplored how they represent concreteness. Here, we used behavioral judgments to estimate semantic distances implicitly used by humans, for a set of carefully selected abstract and concrete nouns. Using Representational Similarity Analysis, we find that the representational similarity space of participants and the semantic representations of language models are significantly aligned and that both are implicitly aligned to an explicit representation of concreteness, which was obtained from our participants using an additional concreteness rating task. Importantly, using ablation experiments, we demonstrate that the human-to-model alignment is substantially driven by concreteness– not by other important word characteristics established in psycholinguistics, such as word frequency.

Keywords: Representational Alignment; Language Models; Concreteness

Introduction

Word concreteness, typically measured through human subjective ratings (Kanske & Kotz, 2010; Brysbaert, Warriner, & Kuperman, 2014), refers to the extent to which a word refers to concepts related to sensory experience (Reilly et al., 2024). Psycholinguistic research has established that whether words refer to concrete or abstract concepts influences their processing both at the behavioral and neural level (Solovyev, 2020; Montefinese, 2019). For example, concrete words are better remembered than abstract words (Fliessbach, Weis, Klaver, Elger, & Weber, 2006). Furthermore, concrete and abstract words show different patterns of brain activations (Fiebach & Friederici, 2004; Bucur & Papagno, 2021). More recently, it has been shown that concreteness estimates can be automatically generated from language models by predicting this information from word embeddings (Köper & Im Walde, 2016; Wartena, 2024) or by probing a generative model (Martínez et al., 2025). However, whether humans and language models have a shared representation of concreteness is an unexplored line of research. In this work, we address this gap by asking three questions: i) is there an alignment of single word meaning between language models representations and human mental representations? ii) Do humans and language models implicitly represent concreteness? iii) Can concreteness explain the degree of alignment between humans and models? To address these questions, we ran a behavioral experiment with 40 participants consisting of two tasks: 1) in a first task, participants were asked to rate 9880 triplets in an odd one out task (Turini & Võ, 2022) generated from all possible combinations of 40 German nouns varied along the concreteness axis. Participants selected the odd word in a triplet of words enabling us to build an implicitly derived representational space; 2) in a second task, the same participants were asked to explicitly rate concreteness of the same nouns. We used Representational Similarity Analysis (RSA) (Kriegeskorte, Mur, & Bandettini, 2008) to compare the representational spaces derived from the odd one out task and from word embeddings coming from German versions of 5 language models (fastText, word2vec, BERT base, BERT large, GPT2). Following Oota, Çelik, Deniz, and Toneva (2024), we performed ablation experiments to investigate whether the representational alignment between humans and language models is driven by a concreteness effect.

Methods

Stimuli generation: We selected 40 nouns by using spectral clustering of word vectors (fastText) from a larger sample of 811 nouns. All words were selected from one cluster, ensuring semantic similarity among items. Words were selected to vary along the concreteness axis, and, as control condition, in word frequency. For concreteness values we used the automatically generated ratings provided by Köper and Im Walde (2016). Word frequency, i.e., how frequent a word is in a representative corpus of a language, was extracted by Subtlex-de (Brysbaert et al., 2011). Additionally, we controlled for the number of characters of each word (between 5 and 9), and

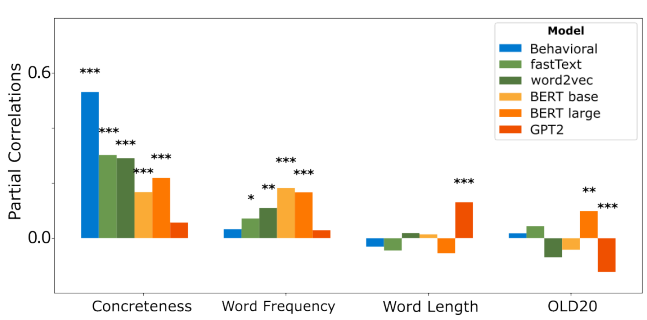

Figure 1: **Partial correlations for the behavioral model (odd-one-out) and the computational models (language models):** The representational space derived from the odd-one-out (in blue) is only correlated to the rated concreteness space, while language models (other colors) are aligned to other feature spaces as well. The representational spaces derived from all language models but GPT2 (in red) show alignment not only to concreteness but also to word frequency. GPT2, instead, is correlated to word length and OLD20. *(*** $p < .001$, ** $p < .01$, * $p < .05$).*

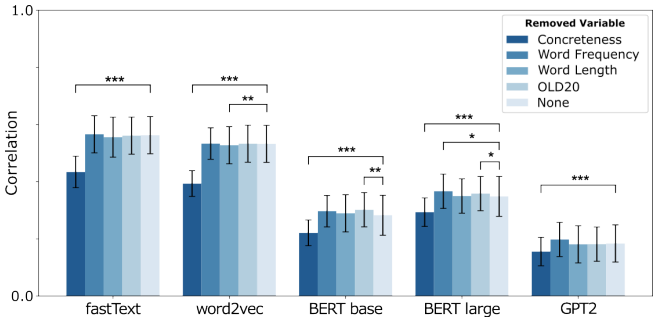

Figure 2: **Representational Similarity Analysis after removing each feature:** Compared to the original correlations between the non-ablated computational representation (lightest shade of blue) and the representation derived from the odd-one-out task, the biggest drop is observed when removing concreteness (dark blue) for all language models. *(Williams' test, *** $p < .001$, ** $p < .01$, * $p < .05$)*

Orthographic similarity (i.e., OLD20, Yarkoni, Balota, and Yap (2008)).

Behaviorally-derived Representational Dissimilarity Matrix (RDM): It was constructed by averaging the triplet ratings that contain the specific word pair (for more details see Turini and Võ (2022)). For language models, all computational RDMs were derived using the pairwise cosine distances between word embeddings. For each variable of interest (concreteness, word frequency etc.) we additionally built explicit feature spaces using euclidean distance.

RSA: it was performed using Spearman correlations between the vectorized lower triangles of the RDMs. For the ablation approach, individual variables were removed from word vectors using a residual approach (Oota et al., 2024), by predicting word embeddings from the ablated variable, training a ridge regression, and subtracting predictions from original vectors.

Results

The RSA revealed that the behavioral RDM generated from the odd-one-out similarity task and the computational RDMs generated from the language models are significantly aligned (fastText, ρ = .51, word2vec ρ= .53, BERT base ρ= .24, BERT large ρ= .37, GPT2 ρ= .14). To assess whether and how strongly humans and language models, considered independently, represent the different word features, we calculated partial correlations with the respective feature while controlling for all other features. Only concreteness (ρ= .53) was represented in the behaviorally based similarity space, but not with any of our control variables (word frequency, word length, and orthographic similarity). In contrast, concreteness and word frequency are represented in all computational models except GPT2, and word length and orthographic similarity show little to no representation across models (see Fig. 1).

When selectively removing word features from word embeddings and rerunning the RSA with the Behavioral RDM, we find a consistent drop of alignment when concreteness is removed for all language models (on average, 20%). No other variable shows a comparable drop in the alignment when ablated from word vectors (see Fig. 2).

Discussion

By investigating the representational alignment between humans and language models, we here show that: i) both systems are independently aligned to a representational space based on explicit concreteness ratings (Fig 1), ii) removing the concreteness feature from the semantic spaces of the language models decreases their alignment with the human data (Fig. 2), and iii) removal of 'lower-level' orthographic or lexical features has no comparable influence on the human-model alignment (Fig. 2). Taken together, these results show a concreteness effect in the representational alignment between human word representations and language models. While we show that concreteness is a relevant dimension organizing semantic representations in the human mind, also implicitly, and that language models represent concreteness in highly similar ways, the question arises how the abstract-concrete dimension is organized in the human brain. Thus, extending the present approach to functional neuroimaging data and model-to-brain alignment will be highly fruitful to improve our understanding of the organization of semantic representations in the human brain.

Acknowledgments

The authors gratefully acknowledge the funding support of the Deutsche Forschungsgemeinschaft (DFG) - DFG Research Unit FOR 5368 (project number 459426179) for GR (DFG RO 6458/2-1) and CJF (DFG FI 848/9-1).

References

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect. *Experimental psychology*.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, *46*, 904–911.

Bucur, M., & Papagno, C. (2021). An ale meta-analytical review of the neural correlates of abstract and concrete words. *Scientific reports*, *11*(1), 15727.

Fiebach, C. J., & Friederici, A. D. (2004). Processing concrete words: fmri evidence against a specific right-hemisphere involvement. *Neuropsychologia*, *42*(1), 62–70.

Fliessbach, K., Weis, S., Klaver, P., Elger, C. E., & Weber, B. (2006). The effect of word concreteness on recognition memory. *NeuroImage*, *32*(3), 1413–1421.

Kanske, P., & Kotz, S. A. (2010). Leipzig affective norms for german: A reliability study. *Behavior research methods*, *42*, 987–991.

Köper, M., & Im Walde, S. S. (2016). Automatically generated affective norms of abstractness, arousal, imageability and valence for 350 000 german lemmas. In *Proceedings of the tenth international conference on language resources and evaluation (lrec'16)* (pp. 2595–2598).

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*, 249.

Martínez, G., Molero, J. D., González, S., Conde, J., Brysbaert, M., & Reviriego, P. (2025). Using large language models to estimate features of multi-word expressions: Concreteness, valence, arousal. *Behavior Research Methods*, *57*(1), 1–11.

Montefinese, M. (2019). Semantic representation of abstract and concrete words: A minireview of neural evidence. *Journal of neurophysiology*, *121*(5), 1585–1587.

Oota, S. R., Çelik, E., Deniz, F., & Toneva, M. (2024, August). Speech language models lack important brain-relevant semantics. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 8503–8528). Bangkok, Thailand: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.acl-long.462/ doi: 10.18653/v1/2024.acl-long.462

Reilly, J., Shain, C., Borghesani, V., Kuhnke, P., Vigliocco, G., Peelle, J. E., ... others (2024). What we mean when we say semantic: Toward a multidisciplinary semantic glossary. *Psychonomic bulletin & review*, 1–38.

Solovyev, V. (2020). Concreteness/abstractness concept: State of the art. In *International conference on cognitive sciences* (pp. 275–283).

Turini, J., & Võ, M. L.-H. (2022). Hierarchical organization of objects in scenes is reflected in mental representations of objects. *Scientific Reports*, *12*(1), 20068.

Wartena, C. (2024). Estimating word concreteness from contextualized embeddings. In *Proceedings of the 20th conference on natural language processing (konvens 2024)* (pp. 81–88).

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond coltheart's n: A new measure of orthographic similarity. *Psychonomic bulletin & review*, *15*(5), 971–979.