Experiential Semantic Information and Brain Alignment: Are Multimodal Models Better than Language Models?

Anna Bavaresco (a.bavaresco@uva.nl), **Raquel Fernández (raquel.fernandez@uva.nl)**
University of Amsterdam
Institute for Logic, Language and Computation
Science Park 107, 1098 XG Amsterdam, The Netherlands

Abstract

A common assumption in AI is that multimodal models learn language in a more human-like way than language-only models, as they can ground text in images or audio. However, empirical studies checking whether this is true are largely lacking. We address this gap by comparing word representations from contrastive multimodal models vs. language-only ones in the extent to which they capture experiential information—as defined by an existing norm-based 'experiential model'—and align with human fMRI responses. Our results indicate that, surprisingly, language-only models are superior to multimodal ones in both respects. Additionally, they learn more unique brain-relevant semantic information beyond that shared with the experiential model. Overall, our study highlights the need to develop computational models that better integrate the complementary semantic information provided by multimodal data sources.

Keywords: semantics; language modelling; fMRI; multimodality

Introduction

The relationship between abstract linguistic representations and the real-world entities they refer to is central to the academic discourse around semantics—the 'symbol-grounding problem' (Harnad, 1990; Bender & Koller, 2020). While some researchers view word meanings as purely symbolic (Fodor, 1983), a great body of cognitive and neuroscientific work inspired by embodied cognition (Barsalou, 2008) emphasises that words have meanings *precisely because* they are linked to specific entities, experiences or notions.

These ideas have motivated a line of computational work aiming to create more human-like language models by learning text representations from sources other than text, such as images or audio. The early efforts in this direction (Bruni, Tran, & Baroni, 2014; Silberer & Lapata, 2012) were characterised by 1) a focus on developing human-aligned computational models of meaning and 2) limited computational modelling resources available. By contrast, more recent works (Deitke et al., 2024; Liu, Li, Li, & Lee, 2024) share 1) a focus on solving, or improving performance on, downstream tasks (e.g., image captioning, visual question answering, visual reasoning), and 2) the availability of massive datasets and large models with billions of parameters. Despite the differences, all these efforts have presented multimodality as a *desideratum*, assuming that images and audio provide additional semantic information that cannot be learnt from text alone; however, there is little to no work investigating *which* these semantic aspects are. Here, we aim to fill this gap by addressing the following question: *Do recent multimodal models learn some facets of meaning related to perceptual experiences that language-only models cannot capture?*

To approach this issue, it is necessary to first to operationalise the 'extra-linguistic' information that multimodal models allegedly learn. We tackle this challenge by relying on a norm-based semantic model introduced by Fernandino, Tong, Conant, Humphries, and Binder (2022) to capture 'experiential information'. By comparing word representations from multimodal and language-only models against the experiential semantic model and fMRI responses, we shed light on the semantic information they capture and expose their limitations as cognitive models of human semantics. For a more detailed description of the present study, see Bavaresco and Fernández (2025).

Experiments

Methods

In this study, we compare five AI models in their ability to 1) reflect experiential semantic information and 2) align with human fMRI responses to single words.

The set of linguistic stimuli we focus on includes 320 nouns, half of which refer to *objects* and the other half to *events*. For these words, we consider an experiential semantic model—Exp48—created by asking crowdworkers to rate each word on 48 predefined dimensions (e.g., *Vision*, *Hand action* or *Unpleasant*) aimed at capturing people's experience of the content described by words. This Exp48 model represents each word as a 48-dimensional array where each entry corresponds to averaged human ratings.

fMRI responses for the same word stimuli were collected by recording brain activity in a 'semantic network ROI', as defined by Binder, Desai, Graves, and Conant (2009), from 36 participants, who viewed each noun in isolation and were instructed to rate it according to the frequency with which they experienced the corresponding entity in daily life. Both Exp48 and the fMRI responses were introduced and made publicly available by Fernandino et al. (2022).

Through representational similarity analysis (RSA) (Kriegeskorte, Mur, & Bandettini, 2008), we compare word representations extracted from five AI models against Exp48 representations and fMRI responses. The AI models we consider comprise three contrastive models and two non-contrastive ones, included as a baseline. The contrastive models are: SimCSE (Gao, Yao, & Chen, 2021), a transformer-based (Vaswani et al., 2017) language-only model trained on pairs of Wikipedia sentences with different drop-out masks applied; MCSE (Zhang, Mosbach, Adelani, Hedderich, & Klakow, 2022), a vision-language model trained contrastively on image-caption pairs; CLAP (Wu et al., 2023), an audio-language model trained contrastively on audio-caption pairs. All three models were pretrained with similar, contrastive learning objectives and share the same BERT-based (Devlin, Chang, Lee, & Toutanova, 2019) architecture as language encoder. For reference, we also evaluate BERT and its vision-language extension VisualBERT (Li, Yatskar, Yin, Hsieh, & Chang, 2019).

Since all these AI models were trained to output contextualised word representations from input text sequences, we note that single words may be an out-of-distribution input. To address this issue, we embed words in five neutral sentence templates (e.g., `Someone mentioned the [word]`)

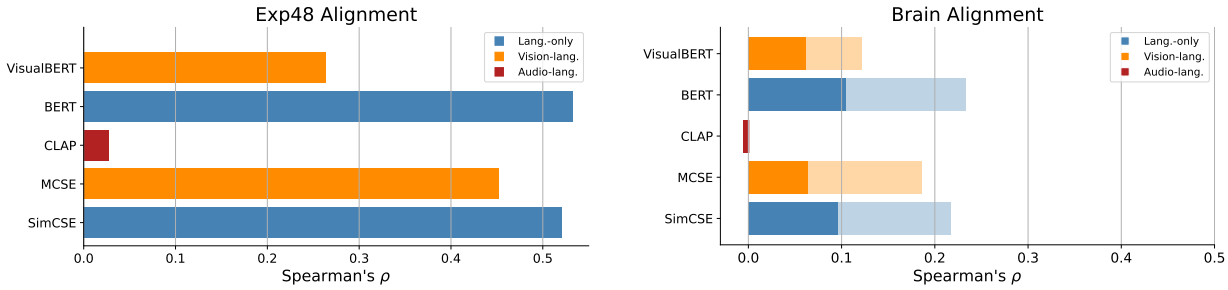

Figure 1: Results from RSA. On the left, Spearman correlations between EXP48 representations and model representations. On the right, initial (lighter shades) and partial (darker shades) correlations between model representations and fMRI responses.

when passing them to the models. To obtain a single vector representation for each word, we first isolate the hidden states of the target-word tokens (discarding those from the rest of the sentence template); next, we average them; finally, we average these target-word hidden states across templates. We extract representations from all model layers and report RSA results considering an average of the representations from the three layers yielding the highest alignment individually. For RSA, we compute all pairwise distances using the cosine metric and measure the alignment between representational spaces as Spearman correlation.

Results

Results from our experiments are reported in Figure 1. We ensure that differences between models are statistically significant by conducting appropriate statistical tests with Bonferroni corrections for multiple comparisons.

RSA against EXP48 (Figure 1's left panel) indicates BERT as the most aligned model (ρ = 0.53); SimCSE and MCSE also display moderate correlations with EXP48 (ρ = 0.52 and ρ = 0.45, respectively). In contrast, CLAP's representations are poorly aligned with EXP48, exhibiting a correlation of just 0.03. A comparison between vision-language models (MCSE and VisualBERT) and their unimodal counterparts (SimCSE and BERT) reveals that the former, surprisingly, reflect *less* experiential information than the latter.

Regarding alignment with brain responses (Figure 1's right panel, light-shade bars), BERT is again the best model (ρ = 0.23), although remaining less brain-aligned than EXP48 (ρ = 0.27). All other models display positive correlations, with the exception of CLAP, whose correlation is not statistically significant (ρ = 0.00, p = 0.70). Similarly to the EXP48-alignment results, here we find the language-only models BERT and SimCSE to be more brain-aligned than their vision-language extensions VisualBERT and MCSE.

To further assess how much of each model's brain alignment is attributable to independently-acquired semantic information as opposed to semantic knowledge shared with EXP48, we conduct a partial correlation analysis where EXP48's representational dissimilarity matrix (RDM) is regressed out from each model's RDM. An interesting result revealed by this analysis is that, although MCSE is more brain-aligned than VisualBERT, their unique contribution without EXP48 is the same in absolute value (ρ = 0.06); in other terms, 50% of VisualBERT's brain alignment is due to unique information, while in MCSE it is 32%. Regarding BERT and SimCSE, the majority of their initial brain alignment is eroded when regressing out EXP48; however, the asymmetry is not substantial, and the unique contribution accounts for more than 40% of the initial brain alignment in both models. As for CLAP, it exhibits a weak negative correlation that is not statistically significant, confirming that the model does not contribute any brain-relevant information.

Discussion and Conclusions

While multimodal models are often expected to learn additional semantic aspects that language-only models cannot learn, our results reveal that their word representations are *less* aligned with EXP48 and fMRI responses than those by language-only models. Moreover, within multimodal models, the vision-language ones show moderate positive correlations with EXP48 and fMRI responses, while the audio-language one correlates weakly with EXP48 and does not yield a significant correlation with brain responses. A potential explanation for our findings is that the dimensions used to create EXP48 are moderately abstract, whereas the extra information learnt by multimodal models may concern lower-level features or patterns of co-occurrence.

Altogether, our study invites caution against assuming that multimodal models are necessarily more human-like than language-only ones, and indicates that there is significant room for improving current computational language models so that they learn the brain-relevant experiential information they currently lack—how to concretely achieve this remains an open question.

Acknowledgments

This project was funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819455).

References

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.*, *59*(1), 617–645.

Bavaresco, A., & Fernández, R. (2025). Experiential semantic information and brain alignment: Are multimodal models better than language models? *arXiv preprint arXiv:2504.00942*.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5185–5198).

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral cortex*, *19*(12), 2767–2796.

Bruni, E., Tran, N.-K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of artificial intelligence research*, *49*, 1–47.

Deitke, M., Clark, C., Lee, S., Tripathi, R., Yang, Y., Park, J. S., ... others (2024). Molmo and PixMo: Open weights and open data for state-of-the-art multimodal models. *CoRR*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N19-1423/ doi: 10.18653/v1/N19-1423

Fernandino, L., Tong, J.-Q., Conant, L. L., Humphries, C. J., & Binder, J. R. (2022). Decoding the information structure underlying the neural representation of concepts. *Proceedings of the National Academy of Sciences*, *119*(6), e2108091119.

Fodor, J. A. (1983). *The modularity of mind*. MIT press.

Gao, T., Yao, X., & Chen, D. (2021, November). SimCSE: Simple contrastive learning of sentence embeddings. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 6894–6910). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-main.552 doi: 10.18653/v1/2021.emnlp-main.552

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1-3), 335–346.

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

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., & Chang, K.-W. (2019). VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Liu, H., Li, C., Li, Y., & Lee, Y. J. (2024, June). Improved baselines with visual instruction tuning. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)* (p. 26296-26306).

Silberer, C., & Lapata, M. (2012). Grounded models of semantic representation. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 1423–1433).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., & Dubnov, S. (2023). Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *Icassp 2023 - 2023 ieee international conference on acoustics, speech and signal processing (icassp)* (p. 1-5). doi: 10.1109/ICASSP49357.2023.10095969

Zhang, M., Mosbach, M., Adelani, D., Hedderich, M., & Klakow, D. (2022, July). MCSE: Multimodal contrastive learning of sentence embeddings. In M. Carpuat, M.-C. de Marneffe, & I. V. Meza Ruiz (Eds.), *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 5959–5969). Seattle, United States: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.naacl-main.436 doi: 10.18653/v1/2022.naacl-main.436