Does multimodal pre-activation influence linguistic expectations in LLMs and humans?

Sasha Kenjeeva, Giovanni Cassani, Noortje J. Venhuizen, Afra Alishahi

Cognitive Science and Artificial Intelligence Research Centre, Tilburg University

a.b.kenjeeva@tilburguniversity.edu

Background. The meaning representations that humans construct for words capture both linguistic and multimodal sensorimotor information [1]. Rommers et al. [2] showed that a word's sensorimotor components might also be pre-activated during online sentence processing: a context in which "moon" is a highly predictable continuation resulted in facilitated processing for "tomato" (similar shape) compared to "rice". To investigate to what extent multimodal pre-activation influences linguistic expectations during sentence processing, we here describe a data-driven experimental setup—with materials normed for plausibility, visual and co-occurrence similarity—that orthogonally manipulates multimodality (sensorimotor similarity; see below) and linguistic predictability (Cloze probability). We hypothesize that high sensorimotor similarity to the likeliest Cloze completion should result in decreased processing effort, even when a word is not predictable from the linguistic context. We report processing effort in terms of LLM surprisal and additionally plan to conduct a human self-paced reading (SPR) study, to shed light on processing behaviour of humans and (multimodal) language models.

Stimuli design. We designed 37 context frames with 5 plausible continuations each (n=185), manipulating linguistic predictability in context (L dimension) using Cloze data [3], and multimodal similarity (MM dimension) to the likeliest Cloze completion – the target – according to word vectors derived from the Lancaster sensorimotor norms [4] and ViSPA, a psychologically-validated computer vision model [5]. Refer to Table 1 for an example of a context: the critical prediction is that the pre-activation of the visual features of the referent of the likeliest completion ("watch") will result in decreased processing difficulty of "compass" (L-MM+) compared to "dog" (L-MM-), due to the visual similarity between a compass and a watch. Figure 1a shows that MM+/- is reflected by high and low visual similarity to the target word, respectively, in both Lancaster norms and ViSPA. As shown in Figure 2a, this effect is partially confounded by simple linguistic similarity, as measured by word2vec [6].

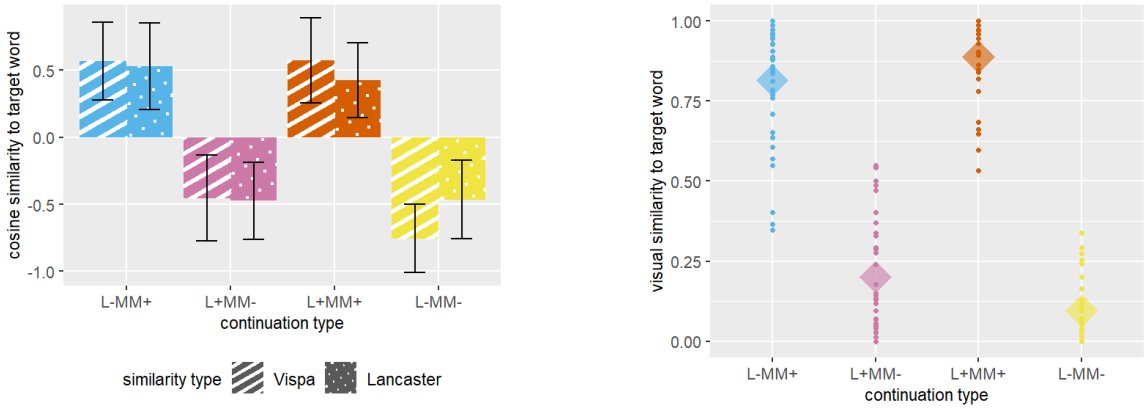
Norming studies. To validate our stimuli design, we collected plausibility, visual similarity and co-occurrence similarity ratings. For each study, we collected $\sim$20 data points per sentence. Participants (n=97) rated most sentences (170/185) as plausible. In the visual (n=145) and co-occurence similarity (n=229) studies, participants viewed the context frames with the target word and rated which of two words on a slider was more visually similar/more likely to appear in similar sentences to the target word. Results from both studies mirrored the similarity patterns of our data-driven measures (see Figures 1b and 2b), thus validating our stimuli design.

LLMs experiment. To test whether visual similarity might lower surprisal in a model trained on text and images, we fitted human ratings for visual and co-occurrence similarity to predict surprisal in two auto-regressive LLMs: GPT2 (language-only [7]) and QWEN2-VL (dual-stream vision-language [8]). Contrary to our prediction, visual similarity did not significantly predict surprisal in either model for the critical comparison (L-MM+ vs L-MM-). Co-occurrence similarity also had no effect on surprisal, indicating that the surprisal estimates of both LLMs seem to follow Cloze predictability patterns, regardless of being trained with visual information or not.

SPR experiment. We are currently conducting a self-paced reading study to investigate whether, contrary to the LLMs, visual similarity influences human reading times. We will then determine whether humans' online processing of plausible sentences entails a multimodal dimension that goes beyond Cloze predictability.

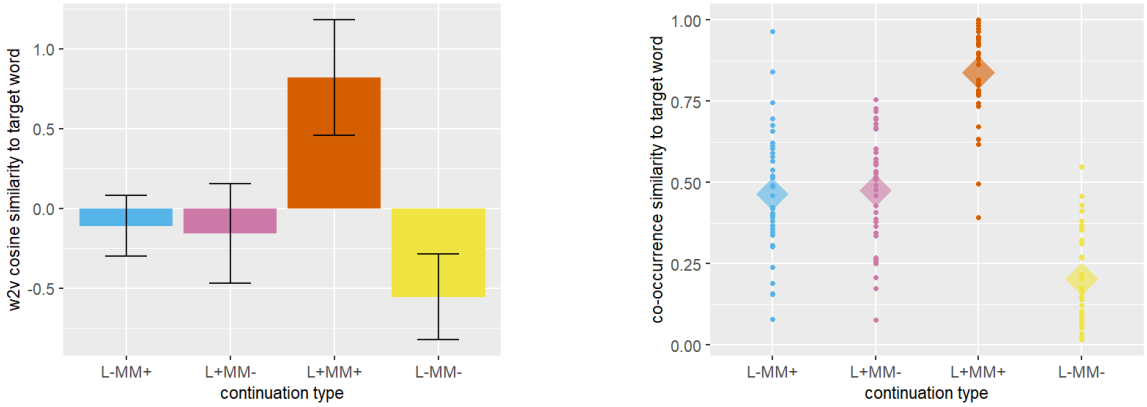
Context	Completion	Continuation Type	Manipulation
The impatient man kept glancing at his …	*watch*	target	highest Cloze
	compass	L-MM+	zero Cloze / high visual similarity
	wife	L+MM-	low Cloze / low visual similarity
	phone	L+MM+	low Cloze / high visual similarity
	dog	L-MM-	zero Cloze / low visual similarity

Table 1: Example of a context frame with its five continuations



(a) Computational visual similarity **(b)** Human-rated visual similarity (mean and per context)

Figure 1: Visual similarity of each continuation type to the target word



(a) Co-occurrence similarity from word2vec **(b)** Human co-occurrence similarity (mean and per context)

Figure 2: Co-occurrence similarity of each continuation type to the target word

References. [1] Davis, C. P., & Yee, E. (2021). Building semantic memory from embodied and distributional language experience. *Wiley Interdiscip. Rev. Cogn. Sci., 12*(5), e1555.

[2] Rommers, J., Meyer, A. S., Praamstra, P., & Huettig, F. (2013). The contents of predictions in sentence comprehension: Activation of the shape of objects before they are referred to. *Neuropsychologia, 51*(3), 437–447.

[3] Peelle, J. E., Miller, R. L., Rogers, C. S., Spehar, B., Sommers, M. S., & Van Engen, K. J. (2020). Completion norms for 3085 English sentence contexts. *Behav. Res. Methods, 52*(4), 1795–1799.

[4] Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The Lancaster Sensorimotor Norms: Multidimensional measures of perceptual and action strength for 40,000 English words. *Behav. Res. Methods, 52*(3), 1271–1291.

[5] Günther, F., Marelli, M., Tureski, S., & Petilli, M. A. (2023). ViSpa (Vision Spaces): A computer-vision-based representation system for individual images and concept prototypes, with large-scale evaluation. *Psych. Rev., 130*(4)

[6] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space.*arXiv:1301.3781.*

[7] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.

[8] Wang, P. et al. (2024). Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv:2409.12191.*