Context matters: History and informativity affect understanding of opaque references

Veronica Boyce (vboyce@stanford.edu), Ben Prystawski, Alvin Tan, Michael C. Frank
Stanford University

Background: Word predictability (often measured as surprisal) is a predictor of language processing times, but how rapidly do comprehenders adapt to changing contexts that may license expressions that are globally uncommon? Does the use of an unfamiliar shorthand become unsurprising with repeated exposure (based on very local context), or does it stay surprising based on overall unexpectedness?

We use transcripts from iterated reference games as a way to create this globally unfamiliar, locally predictable structure. In Boyce et al. 2024[1], groups of people played an iterated reference game where a speaker described a target image so that listeners could pick out that target image from a set of 12 images. Over repeated trials, all images were described 6 times. Descriptions from later rounds were shorter and often used opaque descriptions distilled from successful earlier descriptions. Descriptions were in English.

Methods: For the present experiment, we selected 10 games with above average accuracy, and above average reduction to in-group nicknames, to test how well outsiders could understand the shorthand descriptions and examine their reading time profile. We recruited 198 English-speaking participants who each read trial transcripts and guessed the intended target. Participants first read the transcript by uncovering it word by word in a modified self-paced reading paradigm, and then selected a target image (Figure 1). In the **yoked** condition, a participant saw all the trials from one game in the order they occurred (72 trials, 6 descriptions of each of 12 targets). In the **shuffled** condition, a participant saw all the trials from one game in a randomized order. Materials, data, and code at `github.com/vboyce/tg-matcher`.

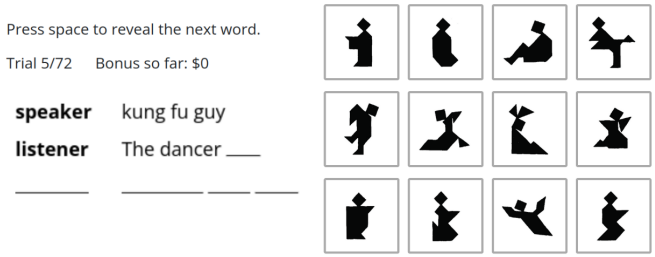
Results: Participants had an average target selection accuracy of 80% in the yoked condition, and 70% the in shuffled condition (Figure 2, Table 1).

What predicts word-by-word reading times in this task? In addition to word length and (log) frequency, we considered 3 measures of a word's informativity: **1)** surprisal from a base large language model (LLM, LLama 3.1 8B)[2], **2)** surprisal from a vision language model (VLM, Llama 3.2 11B)[2], conditioned on grid of 12 possible targets, and **3)** the word-wise change (operationalized as KL divergence) in predicted target distribution from a joint image and text model (CLIP)[3] with a finetuned classifier head (Figure 3). This last measure represents the task-relevant information content of a word based on how much it shifts the model's distribution over possible targets. We residualized the measures of informativity and normalized all word-level predictors, and then modeled RTs as a function of these predictors for the current and 3 prior words, in addition to condition-level predictors (Figure 4).

Shorter and more frequent words were faster to read, as expected. Unexpectedly, current word LLM surprisal was not predictive, and both VLM surprisal and KL divergence were predictive of shorter reading times, in contrast to the typical positive relationship between surprisal and RTs. LLM surprisal, VLM surprisal, and KL all measure different sources of information with increasing task-specificity, and may have different impulse response functions, leading to different patterns of lag.

Discussion: Language is comprehended in many contexts, in pursuit of many different goals. Here, we examined reading times from participants trying to identify the targets of referential expressions produced in a previous experiment. Judging by their accuracies, participants were able to adapt to the unfamiliar shorthand, especially when they saw the interaction history in order. The RT paradigm used is noisy and prone to spillover effects, but RTs were correlated with word-level predictors such as word length, word frequency, and measures of information. Further work could refine the measures using more sensitive techniques such as eye-tracking, analyses that better account for time course, and measures of information that take prior informativity into account.

Figure 1: Experimental set-up



In the current experiment, we sampled the game transcripts from 10 games from Boyce et al. 2024[1]. We recruited new participants to each see the transcripts from one game, either with the trials in the same order (yoked) or a random order (shuffled) for 72 trials. Participants revealed each trial transcript word by word and then selected the image they thought was being described.

Figure 2: Accuracy

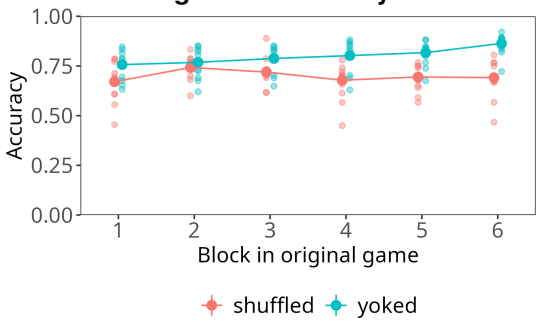
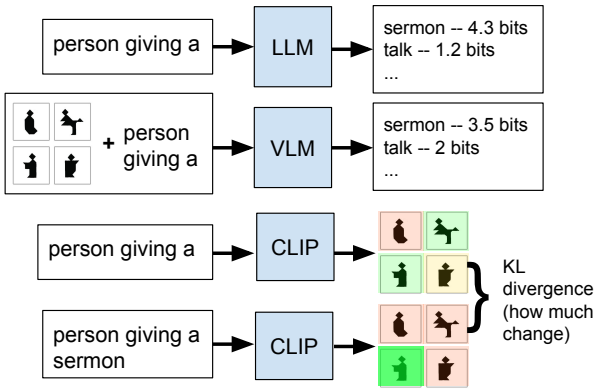


Table 1: Accuracy model

Term	Odds Ratio	95% CrI
Intercept	1.574	[.891 – 2.834]
Original Block	.986	[.954 – 1.020]
Condition (yoked)	2.196	[1.634 – 3.002]
Block x Condition	.944	[.889 – 1.002]
Viewing order	1.019	[1.016 – 1.021]

Logistic Model: *Accuracy* $\sim$ *original-block* $\times$ *condition + order-viewed + (1 | game) + (1 | target) + (1 | participant)* Coefficients are presented as odds ratios (how much more likely correct responses are).

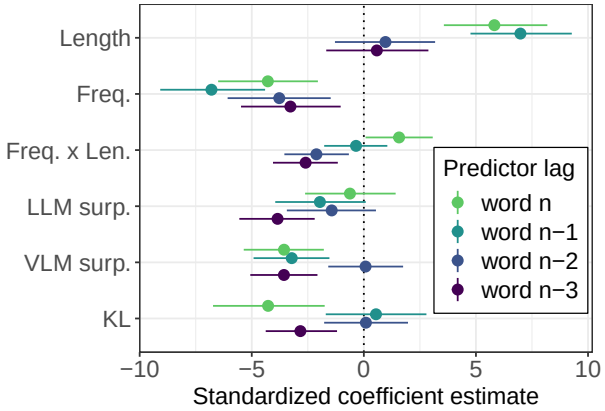
Figure 3: Informativity Predictors

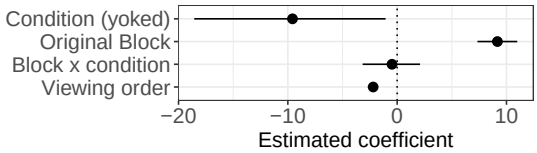


We considered 3 different sources for the information content of a word, here shown for the word “sermon”. We used Llama 3.1 8B for the LLM, and Llama 3.2 11B for the VLM. The CLIP model used a finetuned classifer head trained on the Boyce et al. 2024[1] corpora.

LLM surprisal was residualized off of word length and word frequency. VLM surprisal and KL divergence were residualized off of word length, word frequency, and LLM surprisal. All word-level predictors were normalized.

Figure 4: RT model





Model: *RT* $\sim$ *[word-length* $\times$ *frequency + LLM + VLM + KL] + original-block* $\times$ *condition + order-viewed + (1 | game) + (1 | target) + (1 | participant)* Bracketed predictors were included for current word and 3 prior words to account for spillover.

References:

1. Boyce, Hawkins, Goodman, Frank. PNAS, 2024. • 2. Grattafiori et al., arXiv, 2024. arXiv:2407.21783 • 3. Radford et al. arXiv, 2021. arXiv:2103.00020