Investigating the Correlation Between Human Predictability Judgements and Computational Estimates from Text- and Audio-Based Models

Wei Xue, Julius Steuer, Dietrich Klakow, Bernd Möbius (Saarland University)

weixue@lst.uni-saarland.de

Previous research has shown that human language comprehension improves when an upcoming word is predictable within its context (Pickering and Garrod, 2007). Statistical language models (LM), trained to predict the next word in a sequence, offer probabilistic estimates of word predictability (de Varda et al., 2023). These estimates (e.g., surprisal) have been found to correlate well with human comprehension performance measures such as self-paced reading times (de Varda et al., 2023). In this study, we investigate whether the computational estimates from LMs and Automatic Speech Recognition (ASR) systems align with human judgments on the predictability of target words in sentence pairs. We focus on two estimates: surprisal and entropy. Surprisal reflects how unexpected a word is given its preceding context, while entropy quantifies the uncertainty in predicting the next word in a sequence.

To investigate the alignment, we first conducted a multiple-choice experiment where participants judged which context fit the target word best in paired sentences[1], resulting in binary predictability judgements for the target trigram of target words. An example of sentence pairs is shown in Table 1. We then compared the judgments to the surprisal and entropy estimates derived from the LM and ASR models. We hypothesized that a larger difference in estimates $\Delta_{\text{Estimates}}$ on the target word w given the two contexts correlate with the difference of preference in the human judgements. The difference in estimates is calculated following formula (1), where C_{easy} and C_{hard} refer to the contexts that make the target word easier or more difficult to predict, respectively. The difference of preference in the human judgements $\Delta_{\text{Preference}}$ is calculated following formula (2), where $P(w, C_{\bullet})$ represents the number of participants who judged the context to be more or less predictable. We then correlate $\Delta_{\text{Estimates}}$ and $\Delta_{\text{Preference}}$ over sentence pairs, namely trigrams.

Figure 1 shows a clear difference in surprisal and entropy estimates from LMs between predictable and unpredictable sentential context and type of trigrams. After excluding seven sentence pairs from the total thirty pairs for which there was no agreement in the human judgements (i.e., with $\Delta_{\text{Preference}}$ values smaller than 20), we found significant correlations (r = 0.50, p = 0.0152) of LM entropy from English translations of the stimuli with human judgments and of Dutch LM surprisal summed over whole sentences (r = 0.47, p = 0.022). During the conference, we aim to additionally present the correlation of surprisal and entropy with human predictability judgments in a cross-lingual setting. To this end, we would present native speakers of a language other than Dutch (i.e., German and English) with the Dutch stimuli and ask them to translate the target word. We hypothesize that lexical similarity affects the prediction of the target word if there is a high similarity between the Dutch context and a translated context, and therefore surprisal and entropy at the target word are low.

[1]We first selected 15 target words that are cognates in Germanic languages (i.e., Dutch, German, and English). Then we extracted one high-surprisal (i.e., only preposition phrases, PP) and one low-surprisal (i.e., only noun phrases; NP) trigram for each target word from trigram monolingual LMs trained on CGN (Schuurman et al., 2003), ukWaC, and deWaC (BARONI et al., 2009). Note that phrase type and trigram being high or low surprisal are tangled to ensure this setting is cross-lingual. For each trigram, we constructed two sentences where the target word is more predictable given the context in one sentence than the other, leading to four sentences per target word.

$$\Delta_{\text{Estimates}}(w, C_{\text{easy}}, C_{\text{hard}}) = | - \log_2 p(w|C_{\text{easy}}) + \log_2 p(w|C_{\text{hard}})| \quad (1)$$

$$\Delta_{\text{Preference}}(w, C_{\text{easy}}, C_{\text{hard}}) = \frac{|P(w, C_{\text{easy}}) - P(w, C_{\text{hard}})|}{P(w, C_{\text{easy}}) + P(w, C_{\text{hard}})} \quad (2)$$

Predictability	Trigram Surprisal	Sentence
low	high	De jongen raakte de bal met de **arm**. (English translation "The boy touched the ball with the **arm**.")
high	high	Hij maakte een mooie beweging met de **arm**. (English translation: "He made a nice movement with the **arm**.")
low	low	Hij masseerde zachtjes zijn andere **arm**. (English translation "He gently massaged his other **arm**.")
high	low	Ze toonde trots zijn andere **arm**. (English translation: "She proudly showed his other **arm**.")

Table 1: An example of a paired sentence given a selected trigram of the target word.

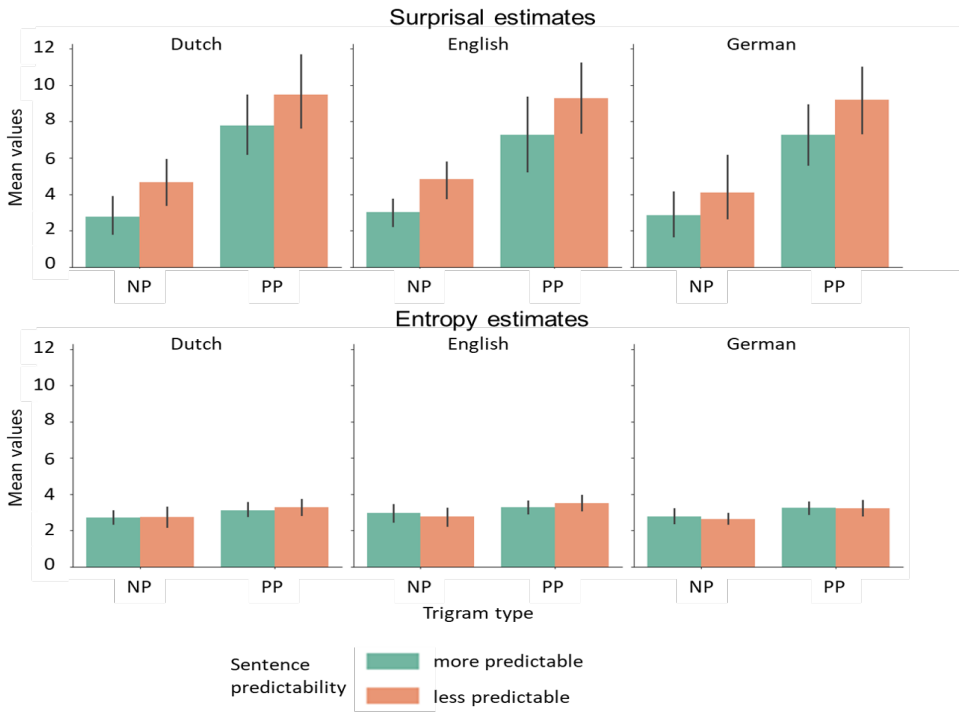

Figure 1: Mean surprisal and entropy estimates from Dutch, English, and German LMs.

References

BARONI, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The wacky wide web: A collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, *43.3*, 209–226.

de Varda, A. G., Marelli, M., & Amenta, S. (2023). Cloze probability, predictability ratings, and computational estimates for 205 english sentences, aligned with existing eeg and reading time data. *Behavior Research Methods*, 1–24.

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *rends in cognitive sciences*, *11*(3), 105–110.

Schuurman, I., Schouppe, M., Hoekstra, H., & van der Wouden, T. (2003). CGN, an annotated corpus of spoken Dutch. *Proceedings of 4th International Workshop on Linguistically Interpreted Corpora (LINC-03) at EACL 2003*. https://aclanthology.org/W03-2414