Dependency length shaped by strategic memory allocation: A corpus study in 11 languages

Weijie Xu & Richard Futrell (UC Irvine)

The processing of nonlocal syntactic dependencies requires working memory (WM) to encode, store, and retrieve the representation of heads and dependents[1-2]. This creates an evolutionary pressure for languages to be structured in a way that keeps the subparts of a syntactic dependency closer to each other, an efficiency principle termed *dependency locality*[3-4]. Given this basic finding, the natural next step is to see how far this efficiency account can go with a more and more realistic characterization of the nature and constraints of WM. We propose strategic memory allocation as one such constraint. Specifically, we propose that the pressure of this locality principle can be modulated by the surprisal of the antecedent (i.e. the first part of a nonlocal dependency) due to strategic memory allocation.

Hypothesis. Novel and unpredictable information should receive prioritized WM resources for encoding and storage in memory, yielding more robust memory representation against interference and decay[5-7]. If this is the case, when an antecedent carries novel and unpredictable information, it can tolerate longer dependency length, with more intervening materials before the other subpart of the dependency is integrated.

Method. We examine the hypothesis in 11 languages (Fig. 1), using corpora from the Universal Dependencies project (UD)[8]. We obtained from the GPT-3 language model[9] the surprisal of each word (the negative log probability of a word given a context $-\ln p(w|c)$). For each word, its context includes the longest preceding text that the model can take in the corresponding document. We then collected all the syntactic dependencies contained in each sentence. Dependency length is calculated as the number of intervening words between the head and the dependent. We analyze the relation between dependency length and antecedent surprisal for: 1) the full dataset with all the dependency types included; 2) a subset that only includes dependencies with subject relations; 3) a subset that only includes object relations.

Statistical models. For analyses on the full dataset, we fit a linear mixed effect model as in (1) for the data of each language. We include antecedent surprisal as the critical fixed effect, with random slopes and intercepts by dependency relation (e.g. subj, obj). We also include three control variables: sentence position in the text), antecedent position in the sentence, and sentence length (word counts). For the analyses on the two subsets of data with subject or object relations only, we fit a linear model as in (2), with the same critical variable and control variables as above.

Results. For the full dataset that includes all the dependency types (Fig.1A), 8 out of 11 languages numerically exhibit a positive correlation between antecedent surprisal and dependency length, among which DA, DE, EN, ES, and IT reach significance. For subject relations (Fig.1B), 7 out of 11 languages (AM, CN, DA, EN, ES, IT, RU) exhibit significant positive effect of antecedent surprisal on dependency length; DE and TR show significant negative effect; the effect for JA and KO does not reach significance. For object relations (Fig.1C), no language shows a significant positive antecedent surprisal effect; AM, CN, DE, KO, RU show a significant negative effect; the effect for DA, EN, ES,IT, JA, TR does not reach significance.

Conclusion. We find that antecedents with higher surprisal exhibit longer dependency length in general. This is consistent with our hypothesis that more surprising information is prioritized for WM resources, receiving more robust representation against decay. Moreover, a closer look at the dependencies that only contain core arguments (i.e. subject/object relations) shows that this positive effect of antecedent surprisal on dependency length is mainly driven by the subject relations.

(1) dep-length $\sim$ sent-pos + antec-pos + sent-len + antec-surpr + (antec-surpr|dep-rel)

(2) dep-length $\sim$ sent-pos + antec-pos + sent-len + antec-surpr

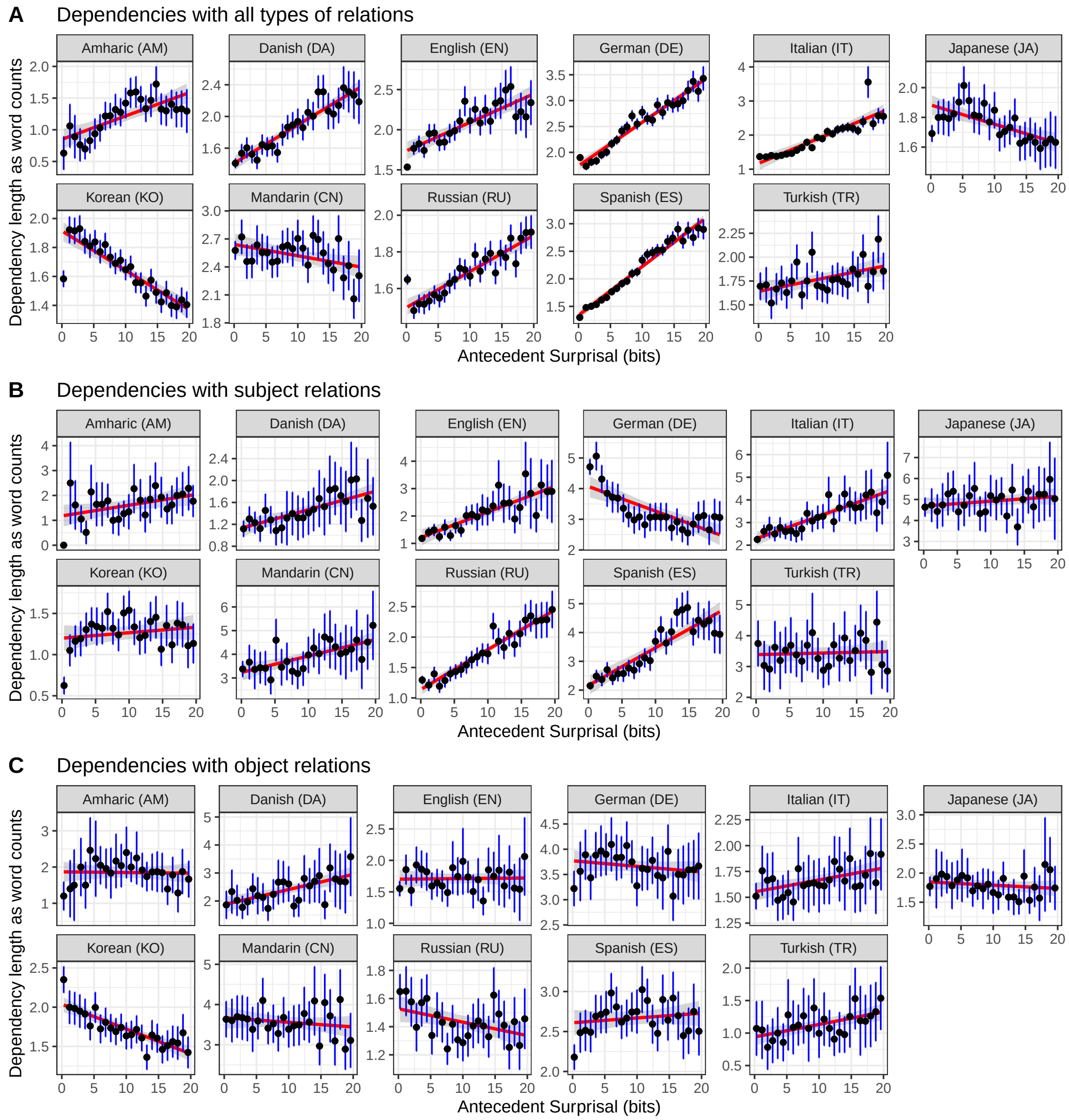

Figure 1: Average dependency length as a function of antecedent surprisal. Surprisal is binned into 25 categories, and the mean dependency length within each category is shown in black dots with 95% confidence interval in blue bars. A linear fit to these points is shown in red.

References. [1] Gibson (1998) *Cognition*; [2] Grodner & Gibson (2005) *Cog Sci*; [3] Gibson et al. (2019) *Trends in Cog Sci*; [4] Futrell et al. (2020) *Language*; [5] Hahn et al. (2022) *PNAS*; [6] Hofmeister (2011) *Lang Cogn Process*; [7] Xu & Futrell (2022) *AMLaP*; [8] Nivre et al. (2016) *LREC*; [9] Brown et al. (2020) *NeurIPS*