# Dependency Length, Syntactic Complexity & Memory: A Reading Time Benchmark for Sentence Processing Modeling

Nina Nusbaumer[1], Corentin Bel[2], Iria de-Dios Flores[3], Guillaume Wisniewski[1], Benoît Crabbé[1]

[1] Université Paris Cité;[2] ENS-PSL & Neurospin, CEA;[3] Universitat Pompeu Fabra

Human sentence processing is influenced, among other factors, by working memory constraints that penalize the retrieval of displaced arguments in syntactic structures involving long-distance dependencies [1, 2, 3, 4, 5]. According to the Dependency Locality Theory [6] [1], such configurations increase cognitive load, slowing down comprehension as integration costs rise with distance. Cue-based retrieval accounts [2] further propose that processing difficulty arises from decay effects and interference from structurally or semantically similar items during retrieval. To support the evaluation of language models on human-like processing behavior, we introduce a dataset of self-paced reading times collected for English sentences that systematically vary in dependency length and syntactic complexity. Existing large English reading-time datasets –such as Natural Stories [7], Franck et al. [8], GECO [9], Dundee [10], ZuCo [11], and the Syntactic Ambiguity Benchmark [12]– offer naturalistic data but lack the control needed to isolate dependency-length and structural effects. In contrast, controlled psycholinguistic datasets precisely manipulate these factors but remain too small and specialized for comprehensive model evaluation. Addressing these gaps, our dataset is explicitly designed to manipulate syntactic variables that drive processing difficulty – namely dependency length, and embedding depth – within a highly controlled, yet lexically diverse and plausible environment. Sentences span from adjacent S–V structures to subject and object center-embedded relative clauses (see Table 1). This design allows us to model key loci of working memory cost [1] while controlling for lexical frequency. By incorporating individual working memory scores, our dataset enables fine-grained modeling of inter-individual variability, a dimension absent from most existing resources. The corpus includes 360 sentences (6 conditions × 60 sets) generated with LLaMA 3.2 (100B) [13] through a grid search varying theme (e.g., school, health, weather) and subject number (singular/plural). The six conditions systematically vary dependency length between the main subject and verb (0, 4, or 9 words) and syntactic complexity, using intervening prepositional or relative clauses. The main verb is in the present tense, embedded verbs in the past, and all noun phrases preceding the main verb share number to prevent interference. Sentences were manually reviewed for contextual plausibility and to reduce antilocality effects. The experimental materials was divided into 6 lists following a Latin square design. Reading times were collected with a self-paced reading paradigm [14] from 510 native English speakers recruited via Prolific. Comprehension questions followed half the sentences to check attention and processing. The reading task was followed by an operation span task to measure working memory capacity, allowing us to correlate processing costs with individual cognitive capacities. Preliminary analyses using linear mixed-effects models confirm that reading times increase with dependency length, modulated by syntactic complexity and working memory capacity. The dataset provides a benchmark for probing language model sensitivity to human-like processing difficulty, under controlled manipulations of these factors. While not intended to test theory directly, it creates conditions where processing difficulty from integration and retrieval demands naturally emerges in both humans and model behavior. We release this resource to support the psycholinguistics and NLP communities in building cognitively grounded, structure-sensitive models of human language processing – models capable of capturing integration cost, individual variation, and structural generalization.

Table 1: Experimental conditions manipulating dependency length and syntactic complexity.

| Condition | Dependency Length (words) | Example Sentence | Clause Type |
|---|---|---|---|
| 1 | 0 | *The **violinist leaves** the stage before the audience loudly applauds the great performance.* | Baseline |
| 2 | 4 | *The **violinist** in the large orchestra **leaves** the stage before the audience loudly applauds.* | PP |
| 3 | 4 | *The **violinist** that followed the conductor **leaves** the stage before the audience loudly applauds.* | SRC |
| 4 | 4 | *The **violinist** that the conductor followed **leaves** the stage before the audience loudly applauds.* | ORC |
| 5 | 9 | *The **violinist** that followed the conductor that worked in the orchestra **leaves** the stage.* | 2×SRC |
| 6 | 9 | *The **violinist** that the conductor that the whole orchestra admired followed **leaves** the stage.* | 2×ORC |

## References


[1] Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, *2000*, 95–126.

[2] Lewis, R. L., & Vasishth, S. (2005). Activation-based models of sentence processing. *Cognitive Science*, *29*(3), 375–419. https://doi.org/10.1207/s15516709cog0000_25

[3] King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, *30*(5), 580–602.

[4] Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(6), 1411.

[5] Holmes, V. M., & O'Regan, J. K. (1981). Eye fixation patterns during the reading of syntactically ambiguous sentences. *Journal of Verbal Learning and Verbal Behavior*, *20*(4), 417–430.

[6] Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1–76.

[7] Futrell, R., Wilcox, E., Morita, T., & Levy, R. (2021). The natural stories corpus: A reading-time corpus of english texts containing rare syntactic constructions. *Language Resources and Evaluation*, *55*(1), 33–49.

[8] Frank, S. L., Monsalve, I. F., Thompson, R. L., & Vigliocco, G. (2013). Reading time data for evaluating broad-coverage models of english sentence processing. *Behavior Research Methods*, *45*, 1182–1190.

[9] Cop, U., Dirix, N., Drieghe, D., & Duyck, W. (2016). The geco corpus: Eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, *49*, 602–615.

[10] Kennedy, A., & Pynte, J. (2003). The dundee corpus. *Language and Cognitive Processes*, *18*(5-6), 777–806.

[11] Hollenstein, N., Zhang, C., & Langer, N. (2018). Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 1237–1243.

[12] Huang, S., Wilcox, E., Dillon, B., & Levy, R. (2024). A large-scale investigation of syntactic processing reveals misalignments between humans and neural language models. *Transactions of the Association for Computational Linguistics*, *12*, 1–27.

[13] Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., & et al. (2024). The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

[14] Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111*(2), 228–238.