## Unsupervised text segmentation predicts eye fixations during reading

*Jinbiao Yang* [1,3], *Antal van den Bosch* [2] *and Stefan L. Frank* [3]
*1 Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands*
*2 KNAW Meertens Institute, Amsterdam, the Netherlands*
*3 Centre for Language Studies, Radboud University, Nijmegen, the Netherlands*

Words traditionally serve as the basic units in psycholinguistic and linguistic studies about sentence reading. However, recent evidence indicates that sub-word units (*e.g.*, morphemes) or supra-word units (*e.g.*, idioms) may take cognitive priority over word units. Since these cognitive units may exist at varying linguistic levels, it is difficult to identify them in strict linguistic terms. Nevertheless, humans are able to learn these flexible units unsupervised and segment text to these units in real time during reading. In previous work (Yang et al., 2020), we assumed that the cognitive units could minimize both long-term and working memory load. Based on the assumption, we designed a computational model (Less-is-Better; LiB, see Figure 1) to simulate the learning and segmentation of cognitive units. We evaluated the units generated by the LiB model and concluded that they show analogies with the cognitive units. However, the LiB-model units were not evaluated on empirical behavioral evidence.

Typical for the word-based traditional view, patterns of eye fixations during reading are usually interpreted as a word-by-word process, and the exceptions are interpreted as that some words are re-fixated or skipped. In present work, we assume that eye fixations during reading more straightforwardly reveal the locations of the cognitive units, and attempt to predict the eye fixations by the LiB-model units on both English and Dutch text. The results show that the LiB-model units can predict eye fixations better than the baselines assuming that fixations are based on words, better than the Chunk-Based Learner model (McCauley & Christiansen, 2019), and comparable with the Adaptor Grammar model (Johnson et al., 2007).

The unsupervised text segmentation models do not aim to simulate the eye movements of humans, nor they are trained on eye-fixation data. However, our findings indicate that the units generated by the models, especially the LiB model, *can* predict eye fixations during reading. Considering that the LiB model was designed to discover cognitive units, we interpret the predictive ability as that: 1. the LiB-model units are analogous to the cognitive units (as already shown in our previous work and further supported by the current empirical evidence) and 2. the cognitive units control the fixation locations.

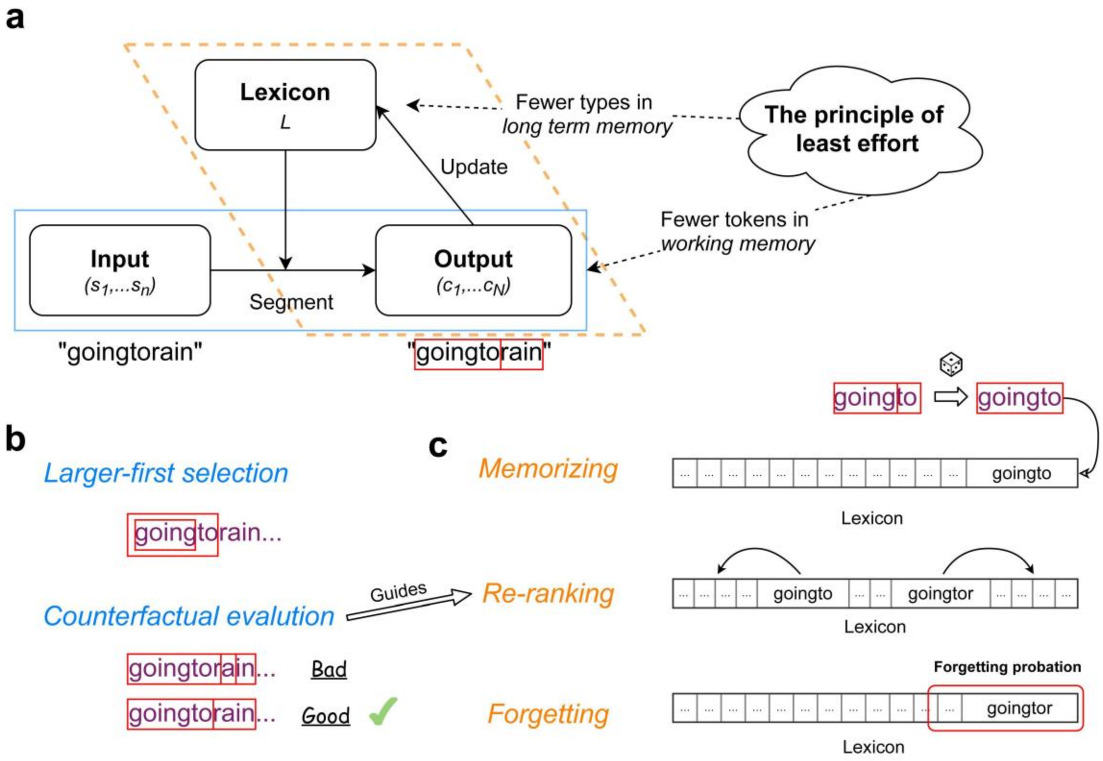

Figure 1:
Illustration of the LiB model: a) information flow in the LiB model; b) the mechanisms in the text segmentation module; c) the mechanisms in the lexicon update module.

## References

- Yang, J., Frank, S. L., and van den Bosch, A. (2020b). Less is better: A cognitively inspired unsupervised model for language segmentation. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon* (Online: Association for Computational Linguistics), 33–45
- McCauley, S. M. and Christiansen, M. H. (2019). Language learning as language use: A cross-linguistic model of child language development. *Psychol. Rev.* 126, 1–51. doi:10.1037/rev0000126
- Johnson, M., Griffiths, T. L., Goldwater, S., and Others (2007). Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. *Adv. Neural Inf. Process. Syst.* 19, 641