Hierarchic syntax improves reading time prediction
Marten van Schijndel & William Schuler (The Ohio State University)
vanschm@ling.osu.edu

Previous studies of eye movements during reading have debated whether humans use hier-
archic syntax during processing [2, 1]. This study demonstrates that hierarchic syntax predicts
reading times even over a strong baseline. Further, this work introduces a simple method to im-
prove language models for future studies.

This study fits linear mixed effects models to reading times from the Dundee corpus. Prior
to evaluation, the first and last fixation of each sentence and of each line, and fixations after
saccades of more than 4 words are filtered out to avoid wrap-up effects and track-loss. During
reading, a person’s eye can saccade over multiple words each time it moves; this study refers
to that span of words as a region. All evaluations in this study used sentence position (sentpos),
word length (wlen), region length (rlen), whether the previous word was fixated (prevfix), and 5-
gram log probability of the current word given the preceding context (5-gram) as independent
variables. Interpolated 5-grams were computed from the Gigaword 4.0 corpus (2.96 billion words).
Each model contains random intercepts for subjects and words, and all independent predictors
are centered and scaled before fitting. Likelihood ratio testing was used to measure significance.

Evaluation 1 – Language Model Improvement: It is common for psycholinguistic models to
include a measure of \( n \)-gram frequency for each fixated word conditioned on its context, but unless
probabilities for words between fixations are also included, the probabilities used in this calculation
are not probabilities of complete word sequences and may miss words that are parafoevially fixated
or simply inferred. To address this, a better metric (cumu-5-gram) was generated by summing the
5-gram log probabilities over each region. To test this new metric, a baseline was created with fixed
factors for sentpos, wlen, rlen, prevfix and random by-subject slopes for all fixed factors, 5-grams,
and cumu-5-grams. Over this baseline, the following fixed effects showed significant improvement:
5-grams (\( p<0.01 \)), cumu-5-grams (\( p<0.001 \)), and both 5-gram factors (\( p<0.002 \) over each model
with a single 5-gram fixed effect).

Evaluation 2 – Hierarchic Syntax: A new model was fit using all above factors as fixed effects
and as by-subject random slopes and with Penn Treebank (PTB) PCFG surprisal as a by-subject
random slope. Over this baseline, a fixed effect for PCFG surprisal significantly improved reading
time predictions (\( p<0.001 \)) suggesting people use more than just sequential information during
sentence processing. Unexpectedly, a cumulative version of surprisal was unable to improve over
the baseline, suggesting only local hierarchic syntactic information affects reading times.

Evaluation 3 – Long-distance Hierarchic Syntax: To confirm the above finding, the effect
of PCFG surprisal was computed using a generalized categorial grammar (GCG) that represents
long-distance dependencies [3]. A new model was fit using all above factors as fixed effects and
as by-subject random slopes and with GCG PCFG surprisal as a by-subject random slope. GCG
PCFG surprisal was a significant fixed effect predictor over this baseline even though PTB PCFG
surprisal was also included as a fixed effect (\( p<0.01 \)). This result suggests that people use non-
local hierarchic structure during reading, though Evaluation 2 suggests that a rich grammar that
explicitly represents long-distance dependencies is needed to observe this effect.

References
[1] Victoria Fossum and Roger Levy. Sequential vs. hierarchical syntactic models of human incremental sentence
[2] Stefan Frank and Rens Bod. Insensitivity of the human sentence-processing system to hierarchical structure.
Psychological Science, 2011.
[3] Luan Nguyen, Marten van Schijndel, and William Schuler. Accurate unbounded dependency recovery using gener-

1Both centered first pass and go-past durations yield comparable results.