MultiplEYE Final Conference
August 20-21, 2026
Skopje, North Macedonia

Effects of context length and text genre on Turkish language models' prediction of human reading: A lossy-context surprisal analysis

Onur Keleş*[1], Sezen Bektaş Yüksel[2], Servet Arda Kayhan[1], Feyza Budan[1], İris Beyza Deveci[1], Nazik Dinçtopal Deniz[2,3]

[1]Department of Linguistics, Boğaziçi University
[2]Department of Foreign Language Education, Boğaziçi University
[3]M.A. Program in Cognitive Science, Boğaziçi University

*contact: onur.keles1@bogazici.edu.tr

Surprisal theory predicts slower reading for less predictable words (Levy, 2008). Psychometric predictive power (PPP; Goodkind & Bicknell, 2018) tests this by examining the log-likelihood gain in predicting human reading behavior through a language model's (LM) surprisal value. Kuribayashi et al. (2022) showed that LMs can become more human-like when prior context is restricted, suggesting that full-context Transformers use information that readers do not always maintain online. We compare full and restricted context LM surprisal in Turkish, an agglutinative SOV language, and ask whether their fit to human eye-movement data depends on model size, text genre, and comprehension accuracy. We used eye-tracking data from 57 native Turkish readers who read 10 MultiplEye passages (Jakobi et al., 2025) from Literary, Institutional, Argumentative, and Popular Science genres. Seven Turkish causal LMs were screened by tokenizer fertility (mean subwords/word; Rust et al., 2021). Kanarya (Safaya et al., 2022) had the lowest fertility (1.35); so, we analysed Kanarya-750m and Kanarya-2b. For each model we computed word surprisal under full context and under lossy n-gram contexts retaining the last 20, 10, 7, 5, 3, or 2 words (where 20 words correspond to broad local context and 2 words correspond to extremely limited context), and examined whether surprisal predicted human eye-tracking measures: log first fixation duration (FFD), gaze duration (GD), total duration (TD), re-reading duration (RRD), total incoming regressions (TRC_{in}), and probability of regression out (PRO). Frequency and word length were fixed effects; subjects and stimulus texts were random effects. Separate PPP models were fit for each context length. We report the results for full-context vs. 2-gram and full-context vs. 7-gram contrasts because they test extreme and intermediate context limitation, respectively. These contrasts were tested by paired permutation tests on by-token squared residuals. Comprehension accuracy was overall 65.3%. It was the highest for Institutional (69.3%) and Argumentative (69.7%) texts and the lowest for Literary texts (62.8%); accuracy for Popular Science texts (66.7%) was in the middle; $\beta = -.069$, $p = .001$. Per-word GD was slower for Institutional and Literary than Argumentative texts, both $p < .001$. Figures 1 and 2 summarize the GD PPP results (reported as ΔLogLik $\times 10^{-3}$) by context length and genre. The context-length effect differed by model size. For Kanarya-750m, predictive power was highest, for GD, when context was not limited (PPP = 5.77) and lost its predictive power decreased in 2-gram contexts ($\Delta = -0.95$; permutation test: $p < .002$). The full context was favored in Kanarya-750m for TD ($p < .001$), RRD ($p = .012$), and PRO ($p < .001$), and in Kanarya-2b for PRO ($p < .001$), as well. Kanarya-750m also showed no 7-gram advantage on any measure (ps $\geq .09$). However, for Kanarya-2b, 7-gram context predicted GD ($\Delta = +0.12$; $p = .002$), TD ($\Delta = +0.86$; $p < .001$), and RRD ($\Delta = +0.69$; $p = .004$) better than full context. Context length effects also interacted with text genre: both full-context vs. 2-gram and full-context vs. 7-gram contrasts showed significant genre effects ($p's < .001$). Institutional was the only genre where 2-gram surprisal improved prediction for both models. ($\Delta \approx +1.5$ to $+1.8$; $p < .001$). In contrast, Literary and Popular Science texts showed the strongest full-context advantage ($\Delta \approx -2.5$ to -5.3), suggesting greater dependence on longer prior context. We also examined whether readers whose TD was more strongly related to surprisal showed higher comprehension accuracy. The results showed that readers whose total duration was more surprisal-sensitive comprehended better: the surprisal–TD relation correlated positively with comprehension accuracy in Popular Science ($r = .442$, $p < .001$) and Literary texts ($r = .267$, $p = .045$). In sum, lossy-context in Turkish was not uniformly beneficial: it improved prediction for the larger Kanarya-2b model, and for Institutional texts but full context was favored for the smaller Kanarya-750m model and for later eye-movement measures. Our results, thus, show that the restricted context effect reported for English and Japanese extends to Turkish (Kuribayashi et al., 2022), but that the useful context window varies with model size, measure, and genre.

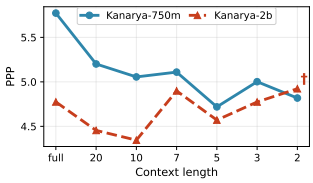

Figure 1. GD PPP by context length and model size.

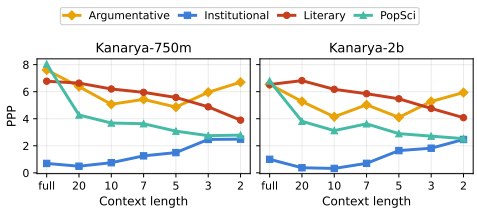

Figure 2. GD PPP by context length, model size, and genre.

References: Goodkind & Bicknell (2018); Jakobi et al. (2025); Kuribayashi et al. (2022); Levy (2008); Rust et al. (2021); Safaya et al. (2022).