

psychological
accuracy

MOCKUPS-DESIGN.COM / Building Vectors by Vecteezy

Lower Perplexity is Not Always Human-Like

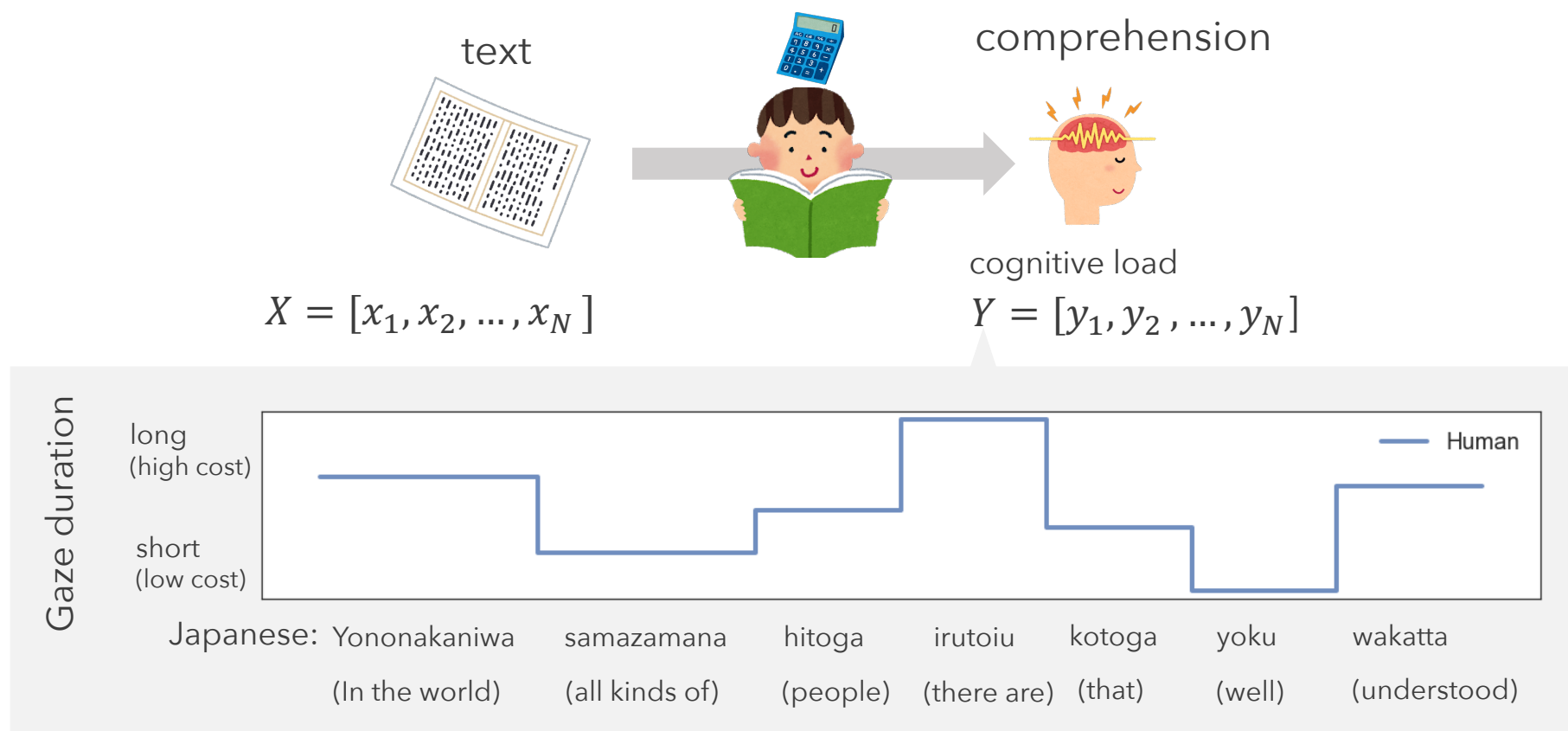
Tatsuki Kuribayashi^{1,2}, Yohei Oseki^{3,4}, Takumi Ito^{1,2}, Ryo Yoshida³,
Masayuki Asahara⁵, Kentaro Inui^{1,4}

linguistic
accuracy

1 Tohoku University, 2 Langsmith Inc., 3 University of Tokyo, 4 RIKEN, 5 NINJAL

Question

What humans incrementally compute during online sentence processing?



Question

- Can recent findings on human-like LMs be generalized across languages?
 - Recent studies have focused almost exclusively on English
 - Theories have been developed by the studies using languages with different sentence structure (e.g., dependency locality theory was developed in SVO languages, and then the anti-locality theory was proposed in SOV languages)

Question

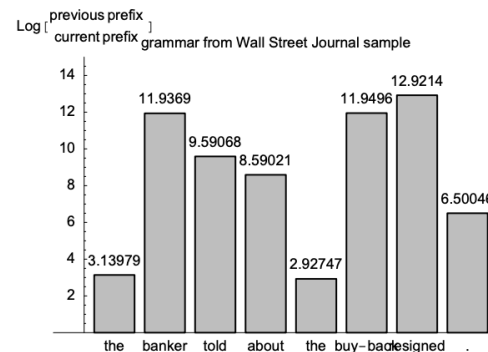
- Can recent findings on human-like LMs be generalized across languages?
 - Recent studies have focused almost exclusively on English
 - Theories have been developed by the studies using languages with different sentence structure (e.g., dependency locality theory was developed in SVO languages, and then the anti-locality theory was proposed in SOV languages)
- We specifically focus on English and Japanese
 - Typologically different from each other
 - Both languages have reliable eye-tracking data (i.e., Dundee Corpus and BCCWJ-EyeTrack)

Background

What determines the incremental processing difficulty during online sentence processing?

- ...
- Dependency locality theory [Hawkins, 1994][Gibson, 1998]...
- Anti-locality [Konieczny, 2000]...
- **Surprisals** computed from (typically) LMs [Hale, 2001][Levy, 2008][Smith&Levy, 2013]...
 - When unexpected information (segment) appears, its processing load increases.

$p(\text{segment}|\text{preceding context})$



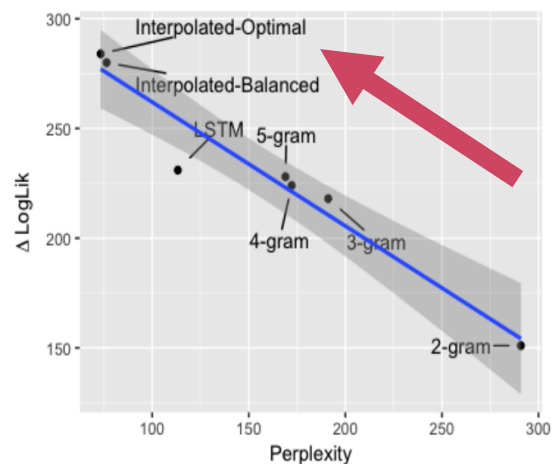
[Hale, 2001]

Background

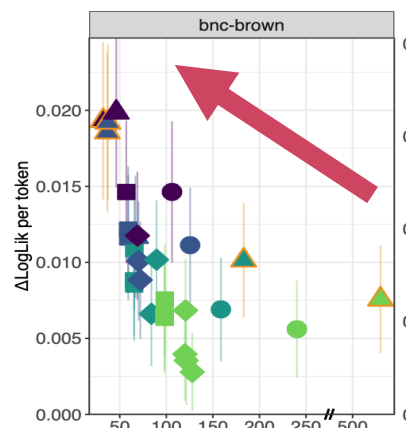
- Surprisals computed from LMs well correlate to human reading behavior
- Next question: **what type of LMs can compute surprisals better simulating the human reading behavior?** [Roark+, 2009][Frank&Bod, 2011][Fossum&Levy, 2012][Hale+, 2018][Merkx&Frank, 2020][Wilcox+, 2020]
 - hierarchical or sequential?
 - lexicalized or non-lexicalized?
 - recurrence or attention?
 - ...

Recent findings

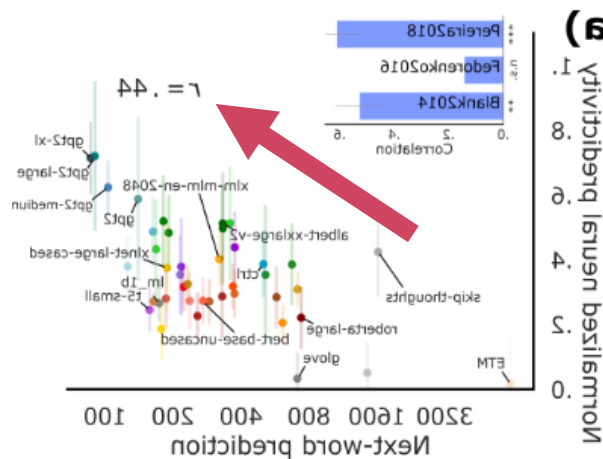
Re-examine the existing report—*LMs with lower PPL could better simulate human reading behaviors*—as a representative of the recent findings



[Goodkind&Bicknell, 2018]

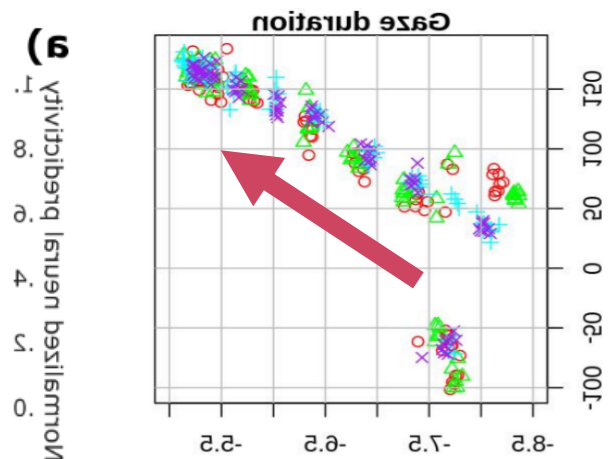


[Wilcox+, 2020]



[Schrimpf+, 2020]

(Flipped the original figure left to right)

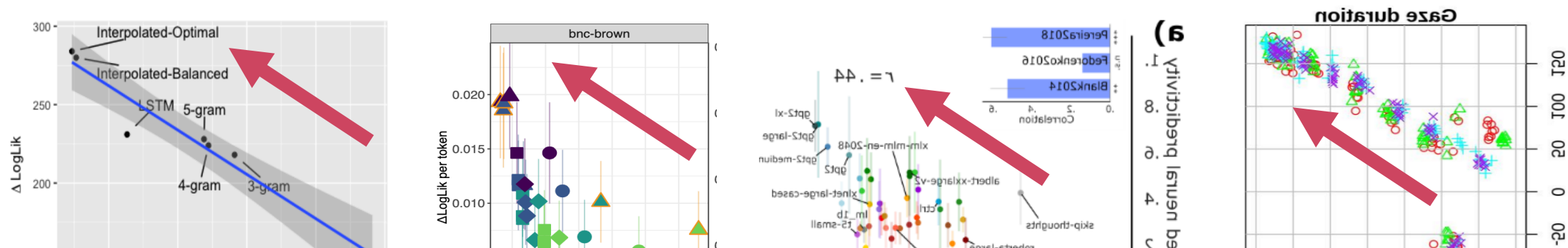


[Merks&Frank, 2021]

(Flipped the original figure left to right)

Recent findings

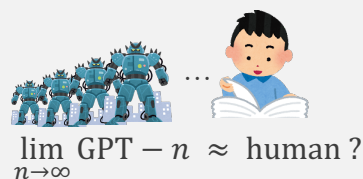
Re-examine the existing report—*LMs with lower PPL could better simulate human reading behaviors*—as a representative of the recent findings



Related question:

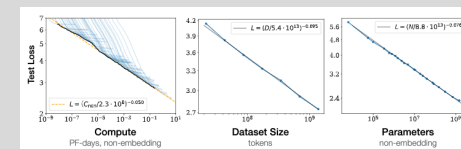
Is using more data, more parameters, and computational cost a recipe for creating human-like LMs?

(Is there a scaling law for achieving human-like LMs?)



“More data, parameters, and computational cost lead to lower PPL of LMs”

(Kaplan+, 2020)



ht)

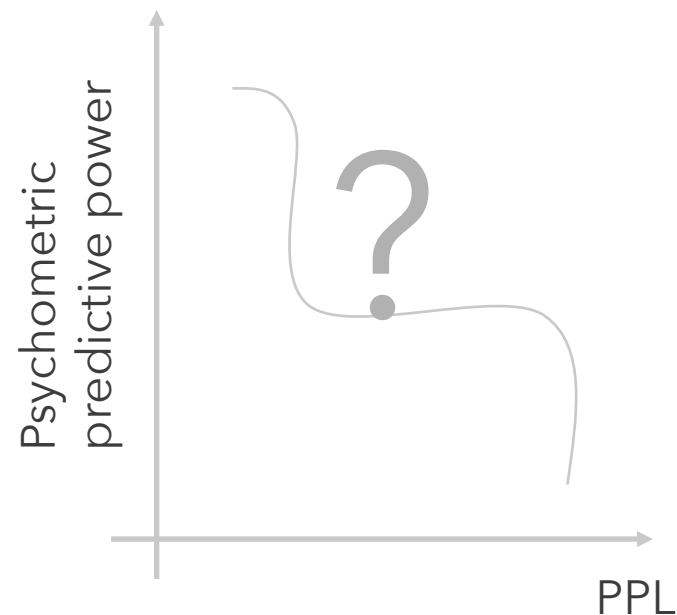
Experimental settings

Investigating the relationship between PPL and psychometric predictive power of LMs in English and Japanese

- PPL
 - evaluated on the texts from eye-tracking data
- Psychometric predictive power
 - how much surprisal contributes to modeling the gaze duration

gain ↑ Gaze duration \sim surprisal + baseline_features
Gaze duration \sim baseline_features

See Section 3.3



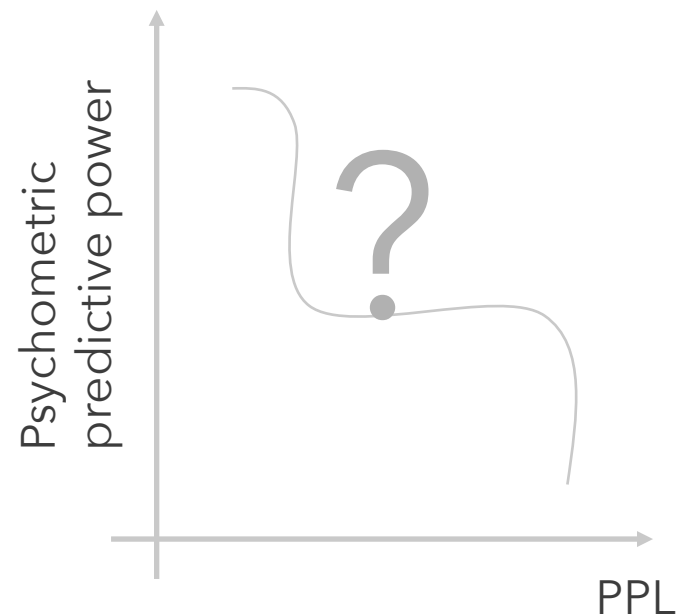
Experimental settings

Investigating the relationship between PPL and psychometric predictive power of LMs in English and Japanese

- PPL
 - evaluated on the texts from eye-tracking data
- Psychometric predictive power
 - how much surprisal contributes to modeling the gaze duration

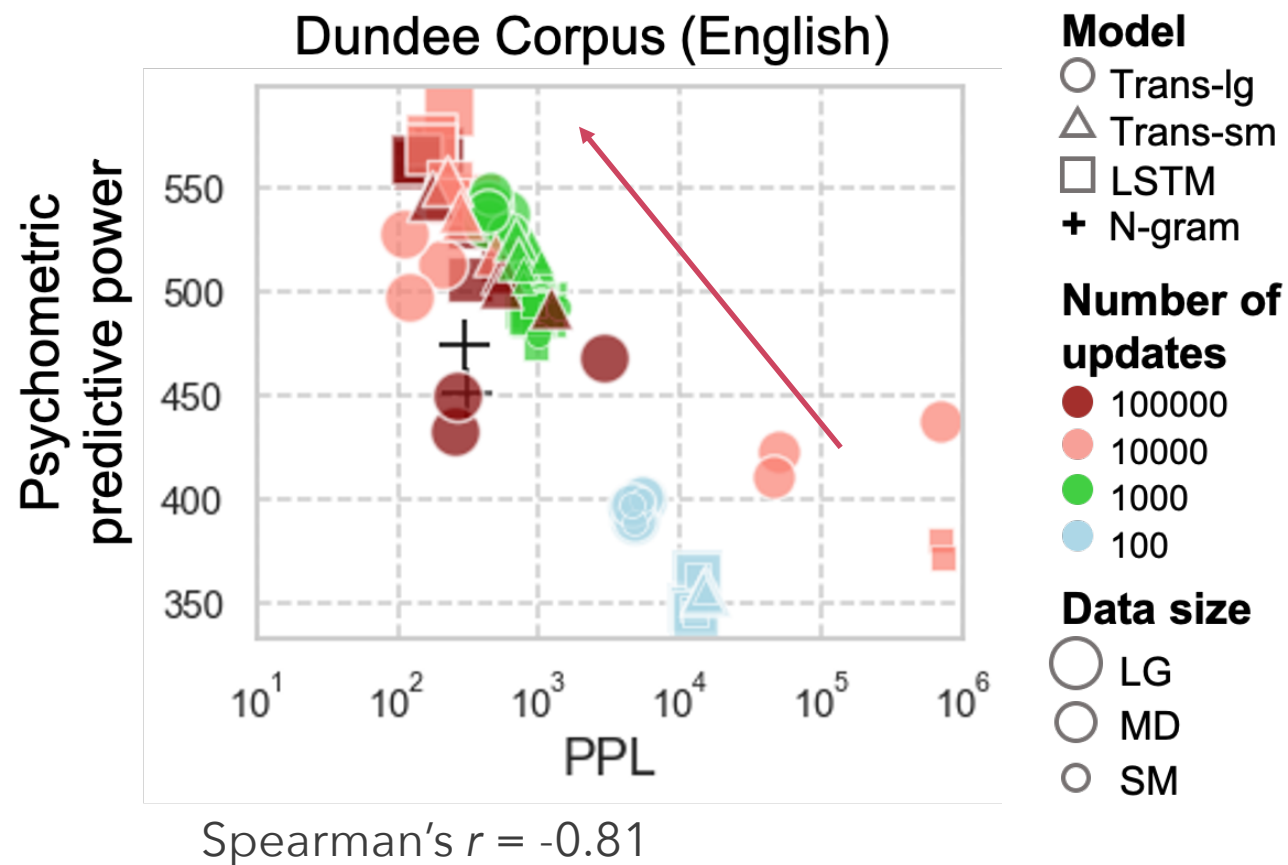
gain ↑ Gaze duration \sim surprisal + baseline_features
Gaze duration \sim baseline_features

See Section 3.3



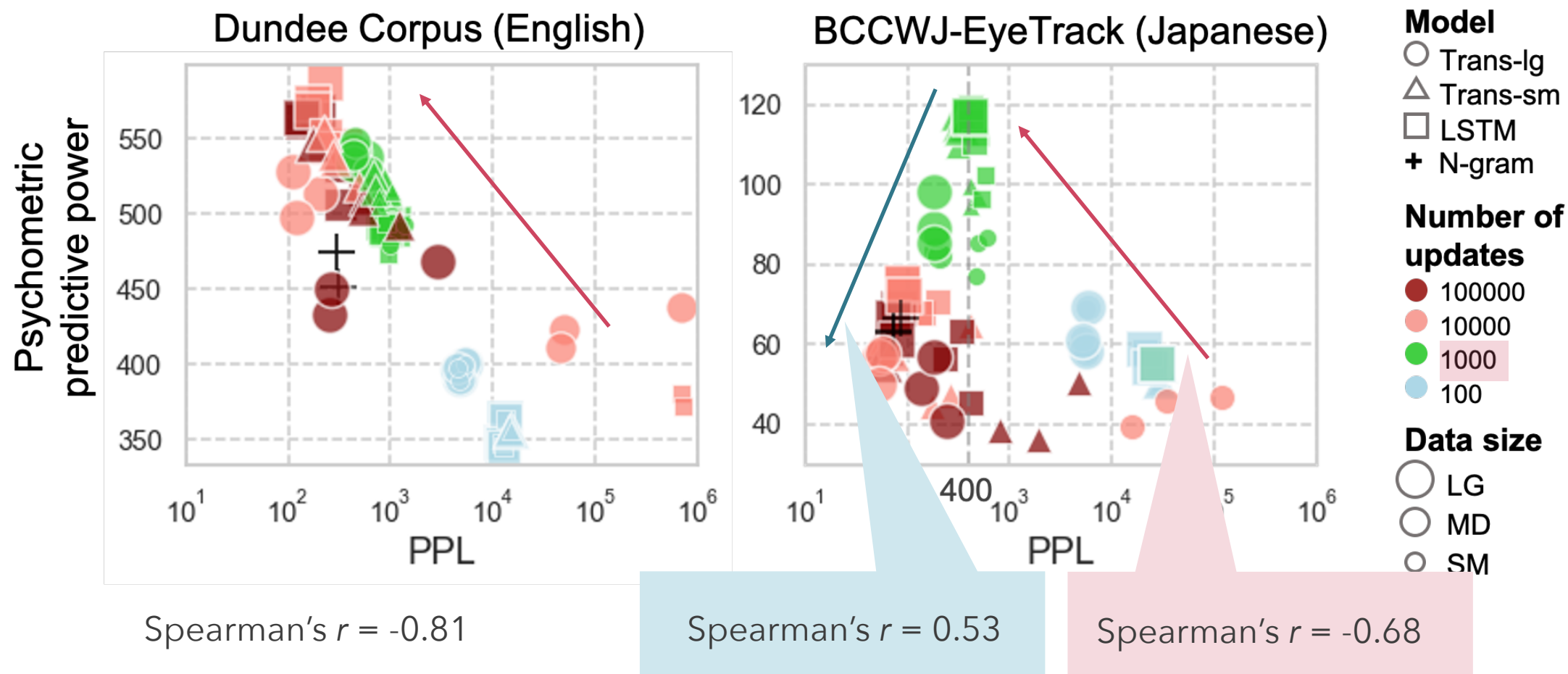
Training 111 LMs with different configurations (e.g., architecture, training data size, the number of parameter updates) for each language.

Results



We found and fixed some issues in the preprocessing for the English part of our experiments after camera ready. In this slide, we used the updated results, which are also shown in https://github.com/kuribayashi4/surprisal_reading_time_en_ja.

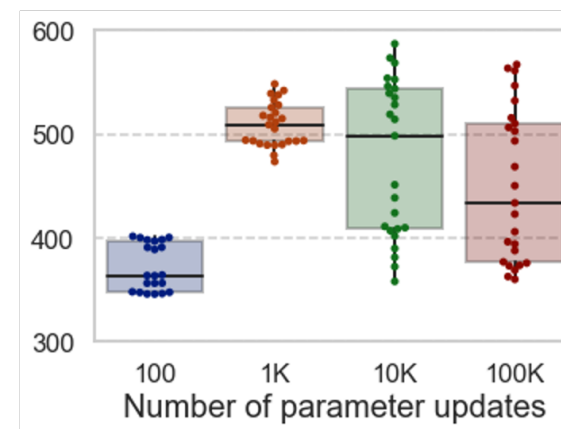
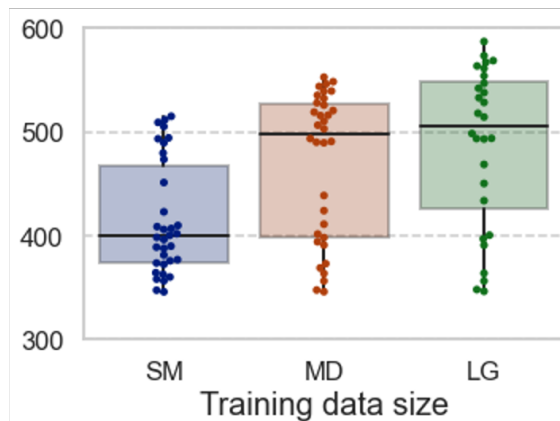
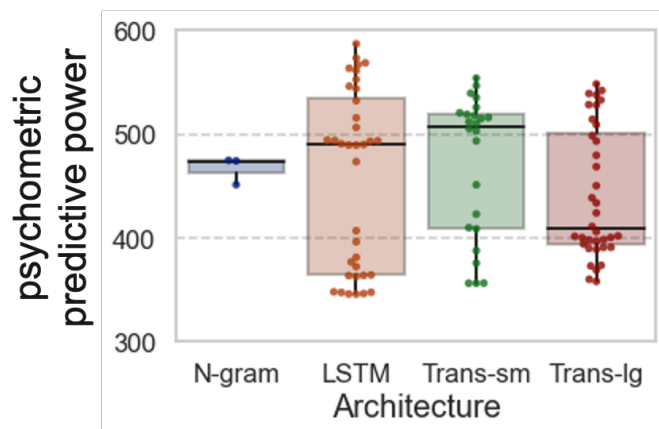
Results



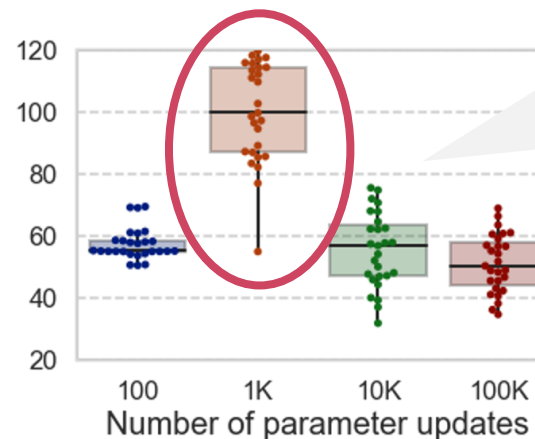
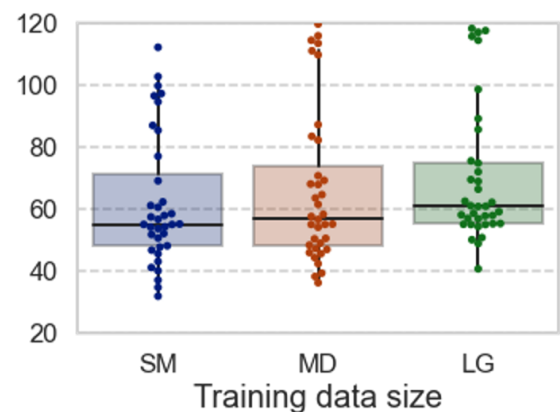
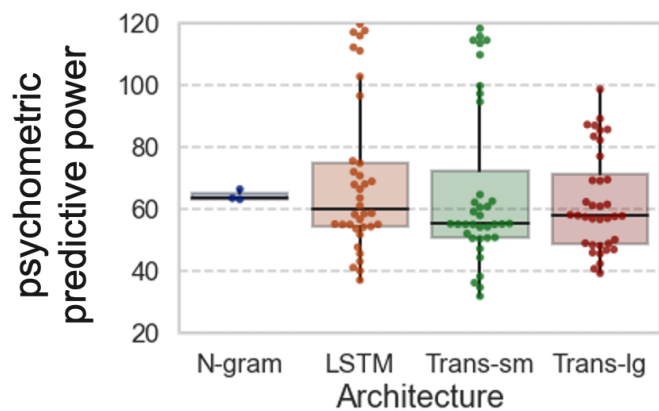
Lower PPL is not always human-like

Results

Dundee Corpus (English)



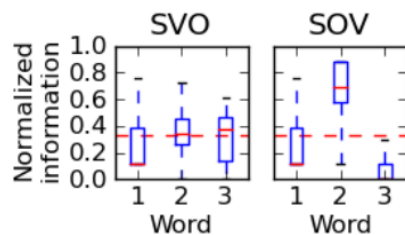
BCCEWJ-EyeTrack (Japanese)



Tuning to the LM training objective had a negative impact

Possible interpretation

- The Japanese language (SOV language) might have a **less uniformity of information density** than English.
 - [Maurits+, 2010] demonstrated that SOV language has less uniformity in information density.

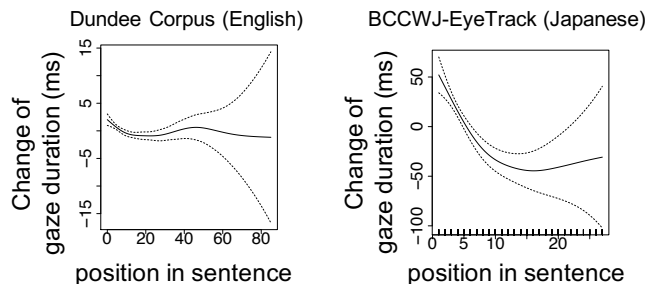


Corpus analysis using English data [Maurits+, 2010]

S O V

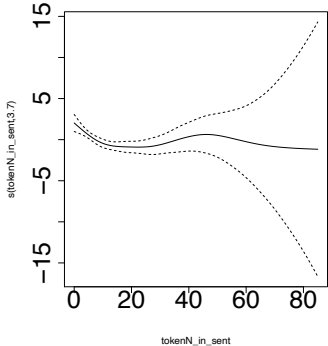
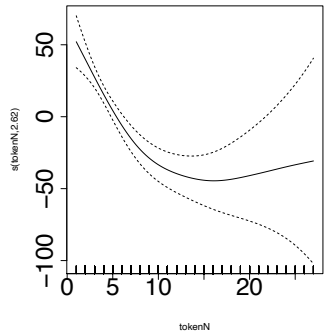


verb has less variety than nouns (S,O)
S and O facilitates the expectation

- We found that the coefficient of variation in gaze duration was 2.5 times higher in Japanese compared to English. Specifically, in Japanese, the gaze duration tended to speed up towards the end of the sentence.



Possible interpretation

The LM objective function $\sum_{i=1}^N \log p(w_i | w_{<i})$, defines that the "ideal" is to maximize all next word probabilities to 1.0 (a *uniform* goal).

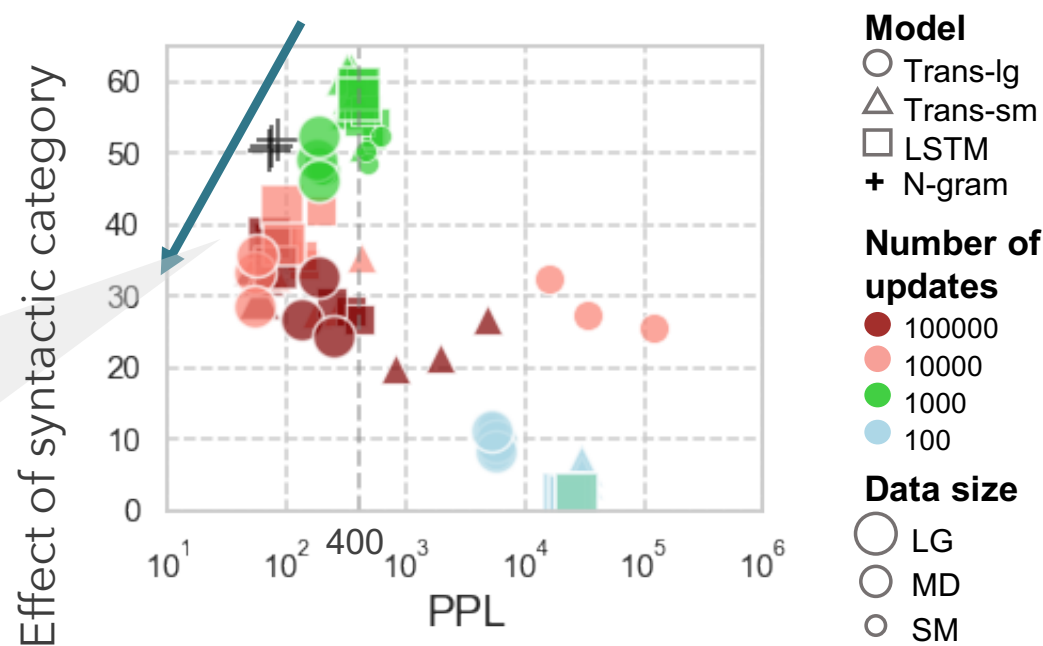
	English	Japanese
Human reading time		
The ideal of LMs	Equally, sufficiently small surprisals 	Equally, sufficiently small surprisals 

Mismatch at least
with respect to the uniformity

Analysis: probing nonuniform information density of Japanese LMs

- Does tuning the LMs to the uniform goal (LM training objective) obscure human-like dispersion in surprisal?
- We investigate whether surprisals from Japanese LMs exhibit the nonuniformity with respect to syntactic category (like part-of-speech).
 - Syntactic category was the most dominant linguistic factor for explaining the human gaze duration.

Human-like nonuniformity is obscured by tuning to the LM objective.



Summary

- Examined whether recent report on the psychometric predictive power of LMs can be generalized across languages.
- The report--the lower PPL a LM has, the more human-like the LM is--**might** lack cross-linguistic universality.
 - We couldn't fully deny the possibility that factors other than the differences in languages (e.g., corpus size, noise on eye-tracking data, experimental settings) affected our results.
- Hopefully, this study encourages researchers to further investigate the universality of human language processing across languages.