Scaling in Cognitive Modelling: a Multilingual Approach to Human Reading Times

Andrea Gregor de Varda
University of Milano – Bicocca
a.devarda@campus.unimib.it

Marco Marelli
University of Milano – Bicocca
marco.marelli@unimib.it

Abstract

Neural language models are increasingly valued in computational psycholinguistics, due to their ability to provide conditional probability distributions over the lexicon that are predictive of human processing times. Given the vast array of available models, it is of both theoretical and methodological importance to assess what features of a model influence its psychometric quality. In this work we focus on parameter size, showing that larger Transformer-based language models generate probabilistic estimates that are less predictive of early eye-tracking measurements reflecting lexical access and early semantic integration. However, relatively bigger models show an advantage in capturing late eye-tracking measurements that reflect the full semantic and syntactic integration of a word into the current language context. Our results are supported by eye movement data in ten languages and consider four models, spanning from 564M to 4.5B parameters.

1 Introduction

The role of context-dependent statistical information in human language processing has received considerable attention in cognitive modelling. A solid empirical finding that has emerged from this research line is that speakers actively anticipate the upcoming linguistic material (Huettig, 2015; Staub, 2015). Indeed, behavioral and neural patterns that are diagnostic of reduced cognitive cost have been reported in response to predictable words; these emerged from the analysis of eye movements (Staub, 2015; Ehrlich and Rayner, 1981), changes in pupil size (Frank and Thompson, 2012), self-paced reading times, (Frank and Hoeks, 2019; Fernandez Monsalve et al., 2012), ERP responses (DeLong et al., 2005; Van Berkum et al., 2005; Kwon et al., 2017), frontotemporal blood oxygenation levels (Baumgaertner et al., 2002; Dien et al., 2008), and MEG data (Takahashi et al., 2021).

Inferential theories of language comprehension argue that prediction must be an intrinsic feature of an incremental probabilistic cognitive processor (Levy, 2008; Shain et al., 2022). These accounts contend that the Kullback-Leibler (KL) divergence (i.e., relative entropy) between the probabilistic state of the processor before and after observing a given word is the cause of the processing difficulty associated with that word. It has been demonstrated that the KL divergence associated with this probability shift is mathematically equivalent to the *surprisal* of that word, i.e., the negative logarithm of its probability conditioned by the preceding sentence context ($surprisal(w_i) = -\log P(w_i|w_1, w_2 \dots w_{i-1})$; Levy, 2008). Inferential theories, which predict a logarithmic linking function between contextual predictability and cognitive cost, are supported by extensive experimental evidence in the computational psycholinguistics literature (Smith and Levy, 2008, 2013; Wilcox et al., 2020; Shain et al., 2022, but see Hoover et al., 2022; Brothers and Kuperberg, 2020).

Statistical language models developed in NLP research have been of paramount importance in the evolution of inferential theories of language comprehension. Indeed, language models are usually trained to predict the upcoming word in a corpus of naturalistic text, and thus define a conditional probability distribution that can be employed to compute word surprisal. Modern computationally-derived estimates of word predictability have been shown to perform on par (Shain et al., 2022) or even better (Hofmann et al., 2022; Michaelov et al., 2022) than predictability estimates obtained with expensive human annotation (although they fail to account for the processing demands of some specific linguistic patterns, see Arehalli et al., 2022; Van Schijndel and Linzen, 2021; Hahn et al., 2022). However, given that language models display a great amount of variation in their architectures and performances, various studies have investigated

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 139–149

which models are better suited to characterize the behavioral correlates of human sentence comprehension. Seminal work has shown that the "linguistic accuracy" of a model (i.e., its ability to accurately predict the next word) is positively related to its "psychological accuracy" (namely, the capability of a surprisal estimate to explain variance in human responses, as captured by the increase in fit in a corresponding statistical model; Goodkind and Bicknell, 2018; Wilcox et al., 2020; Merkx and Frank, 2021, but see Hao et al., 2020; Kuribayashi et al., 2021).

A recent incidental finding by Shain et al. (2022) shed doubt on such conclusion. The authors reported that the GPT-2_{small} model substantially outperformed GPT-3 in predicting self-paced reading times and fixation patterns while having a parameter size smaller by three degrees of magnitude and displaying higher perplexity values in next-word prediction. The result, which suggests that the correlations between the linguistic and psychological accuracy of language models might not hold for very deep transformer-based architectures, has been promptly replicated with different GPT-2 variants (Oh et al., 2022; Oh and Schuler, 2022). This observation is at odds with the empirical scaling laws for neural language models (Kaplan et al., 2020), which show that the quality of a language model (both in terms of test loss and downstream performance, Hernandez et al., 2021) increases monotonically as the number of parameters increases (although see Lin et al., 2022).

2 Related work and motivation

Research in computational psycholinguistics has largely followed the progressive switch to the Transformer architecture that has characterized the NLP literature in the last years, with Transformer-based surprisal estimates being evaluated as predictors of processing difficulty (Wilcox et al., 2020; Hao et al., 2020; Merkx and Frank, 2021). While early studies within this research line have documented a positive relationship between the linguistic and the psychological accuracy of a model (Goodkind and Bicknell, 2018; Wilcox et al., 2020; Merkx and Frank, 2021), recent findings with decoder-only large language models have documented an opposite pattern, with larger and better-performing pre-trained Transformers providing worse psychometric estimates than their smaller counterparts (Oh et al., 2022; Oh and Schuler, 2022).

The possibility that cognitive modelling might constitute an exception to scaling laws is intriguing, but further examination is needed to warrant such claims. All the evidence in support of this view has come from the English language alone (except from Kuribayashi et al., 2021), leaving an open question as to the cross-lingual generalizability of these findings. The English-centric approach to this problem is not surprising, since inferential approaches to language processing have been primarily supported by experimental evidence in English (Aurnhammer and Frank, 2019; Frank and Bod, 2011; Frank et al., 2015; Fernandez Monsalve et al., 2012; Wilcox et al., 2020; Goodkind and Bicknell, 2018; Smith and Levy, 2013), Dutch (Frank and Hoeks, 2019; Brouwer et al., 2010) and German (Boston et al., 2008; Brouwer et al., 2021), while empirical support from non-Germanic languages is far more limited (although see Fan and Reilly, 2020; Kuribayashi et al., 2021). To the best of our knowledge, there is only one study that provided large-scale cross-lingual evidence in support of surprisal theory (de Varda and Marelli, 2022). Indeed, both NLP (Joshi et al., 2020) and cognitive science research (Blasi et al., 2022) have long over-relied on the English language to develop language processing systems and test theories of language and cognition. This tendency can lead to hasty claims of generality, and must be mitigated with cross-linguistic research efforts challenging the universality of English-specific findings.

Another potential shortcoming of the studies that reported the inverse scaling trend is that they only considered a single eye-tracking measurement as an index of processing cost (Oh et al., 2022; Oh and Schuler, 2022). This choice reflects a common tendency within the inferential language processing framework (Aurnhammer and Frank, 2019; Goodkind and Bicknell, 2018; Smith and Levy, 2013; Wilcox et al., 2020); however, natural reading is an ability composed of multiple sub-processes characterized by different levels of complexity (see for instance Plaut et al., 1996; Coltheart et al., 2001). In principle, it is reasonable to assume that different processing stages, characterized by different degrees of complexity, might be better captured by models with varying parameter sizes, with shallow processes better modelled by (relatively) simpler networks, and complex integrative operations better characterized by more complex architectures.

3 Aims

The current work aims at inspecting the relationship between the linguistic and the psychological accuracy of a neural language model across languages, testing whether previous observations on inverse scaling in cognitive modelling hold across a sample of ten languages belonging to four different families. Furthermore, our study considers different eye-tracking measures that are thought to reflect different processing stages, to examine the possibility that the relationship between the psychological and linguistic accuracy of a model might vary as a function of the computational complexity of the cognitive operations being studied.

4 Methods and materials

4.1 Data

In this study, we considered the eye movement data from the MECO-L1 corpus (Siegelman et al., 2022), a large-scale repository of eye-tracking records covering 13 languages. Participants engaged in a naturalistic reading task, and were presented with 12 texts consisting of encyclopedic entries on a handful of topics; five of the twelve original texts were translated from English to the target languages, while the other seven were non-translated texts on the same topics and with the same writing styles, comparable length, and similar difficulty. Data points that showed either very short first fixation durations (< 80 ms) or very long total fixation times (top 1% of the participant-specific distribution) were discarded. We analyzed three measures of eye movement behavior for each word w_i, which are thought to reflect early, intermediate, and late stages of processing:

1. *First fixation (FF):* the time elapsed during the first fixation on w_i. This measure is often assumed to reflect low-level oculomotor processes, early lexical access, and predictive processing (Demberg and Keller, 2008; Staub, 2015).

2. *Gaze duration (GD):* the sum of the fixations landing on w_i before the gaze leaves the word for the first time. This measure is thought to be indicative of lexical access, and possibly of early syntactic and semantic integration (Inhoff and Radach, 1998; Rayner, 1998).

3. *Total reading time (TT):* the total amount of time spent looking at w_i, including fixations returning to the word after having left it. This measure is thought to reflect full semantic integration (Radach and Kennedy, 2013) and syntactic integration and reanalysis (Meseguer et al., 2002).

4.2 Models

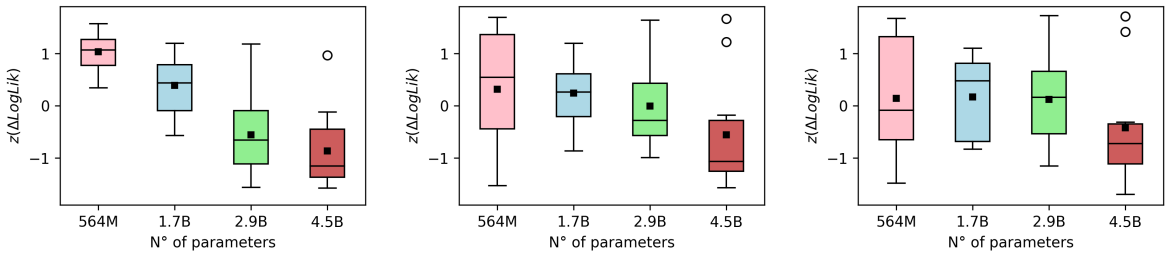
In this study, we employed the XGLM family of auto-regressive language models (Lin et al., 2021). XGLMs are Transformer-based, decoder-only language models inspired by GPT-3 (Brown et al., 2020). We considered four pre-trained models, with 564M, 1.7B, 2.9B, and 4.5B parameters, and extracted word-by-word surprisal estimates from each of them. In the case of multi-token words, we summed the log probabilities assigned to the sub-word tokens, following the chain rule.

4.3 Analyses

Of the 13 languages included in the MECO dataset we had to exclude the Hebrew, Dutch, and Norwegian data, since these languages were not included in the XGLM pre-training data. Thus, our analyses were conducted in ten languages belonging to four language families (see Appendix A). On average, there were 65,450.8 available data points for each language (SD = 19,712.2). We fit 120 linear[1] mixed-effects regression models (10 languages $\times$ 4 models $\times$ 3 fixation measurements), with random intercepts for participants and items. We included as linear covariates length, log-frequency, and their interaction relative to w_i, w_{i-1}, and w_{i-2}, to account for spillover effects. Our models also included a main effect of surprisal relative to w_i, w_{i-1}, and w_{i-2}. All the variables were standardized before being entered into the mixed-effects regression models.

To evaluate the increase in the goodness of fit due to the inclusion of surprisal as a fixed effect, we compared each model with a corresponding baseline model, which was identical except for the absence of the fixed effects of surprisal. As common practice in the literature, we calculated the difference in the log likelihood between the baseline and the experimental model (ΔLogLik; Goodkind and Bicknell, 2018; Wilcox et al., 2020; Kuribayashi et al., 2021; Oh and Schuler, 2022). In the literature we have reviewed in §1, a common approach was to correlate the perplexity of a language model with the ΔLogLik obtained by adding the surprisal terms; however, perplexity values can

[1] Our choice of fitting linear models is supported by ample evidence showing that the functional form of the effects of log-probabilities on reading times is indeed linear (see Smith and Levy, 2008, 2013; Wilcox et al., 2020; Shain et al., 2022)



(a) First fixation duration (b) Gaze duration (c) Total reading time

Figure 1: Plots of the increase in model fit (ΔLogLik) obtained by adding the surprisal estimates from XGLM models with different parameter sizes. The horizontal lines in the box plots indicate the median value obtained across languages, while the squared marker shows the mean value. Note that the ΔLogLik values were standardized by language.

be properly compared only in the context of a fixed reference vocabulary (Wilcox et al., 2020). Technically, XGLM models produce a conditional probability distribution over the same whole vocabulary, regardless of the language of the specific text they are processing. However, the models have received strong evidence during pre-training that some subportions of the vocabulary (e.g. Cyrillic tokens) should be essentially ignored while processing text in some languages (e.g. English), thus reducing their *actual* reference vocabulary. Hence, while we report the perplexity-based results in Appendix B, we focused on the link between the linguistic and psychological accuracy of the models by observing how the ΔLogLik was affected by the parameter size of the model. The choice of employing parameter size as a proxy of linguistic accuracy is supported by the results in the original XGLM paper, where the authors reported better results in almost all downstream tasks with the bigger versions of the XGLM model family (Lin et al., 2021).

The code employed in this study is publicly available[2].

5 Results

The first main finding of our study is that surprisal is a solid predictor of reading times across the languages considered, confirming the previous observation that context-dependent probabilistic processing generalizes beyond the Germanic language sample typically considered in the literature (de Varda and Marelli, 2022). The XGLM-based surprisal estimates were statistically significant in all cases when considering GD and TT, and in the vast majority of the cases when considering FF (see Appendix A).

The increase in goodness of fit that could be attributed to surprisal is displayed in Figure 1, grouped by model type and fixation measure. Concerning FF (1a), we reported a general decrease in ΔLogLik when increasing the number of parameters, with the smallest XGLM_{564M} variant outperforming the bigger models in terms of psychological accuracy. A similar trend can be observed in GD (1b), although the difference in psychological accuracy between XGLM_{564M} and $\text{XGLM}_{1.7B}$ appears to be rather small[3]. The results are different when considering TT as the dependent variable (1c), as in this case the model that provided the highest average increase in goodness of fit was $\text{XGLM}_{1.7B}$[4].

6 Discussion

In this experiment, we showed that large multilingual Transformer-based models were outperformed by their smaller variants in predicting early eye movement measurements of processing difficulty. These measurements are thought to reflect predictive processes, lexical access, and early semantic integration. This result corroborates the previous claims that cognitive modelling might constitute an exception to empirical scaling laws in NLP (Oh and Schuler, 2022). However, predictability estimates computed by *relatively* larger variants of the same architecture – but not the largest – provided surprisal estimates that better captured late

[2] https://github.com/Andrea-de-Varda/surprisal-across-languages

[3] An anonymous reviewer rightfully noted that cross-lingual variation increased in GD and TT with respect to FF; we provide a tentative explanation for this phenomenon in Appendix C.

[4] Note that a coherent pattern was observed when employing perplexity as an index of the language model's linguistic accuracy (Appendix B).

eye-tracking measurements, which are thought to reflect the full semantic and syntactic integration of a word into the phrasal context. This dissociation is in line with the observation that it is not appropriate to adopt a "one-size-fits-all" approach when studying how linguistic distributional knowledge explains different cognitive processes (Wingfield and Connell, 2022). Instead, context-dependent probabilistic information derived from different neural architectures might be more apt to model certain cognitive mechanisms, depending on the computational complexity of the processes being considered.

Limitations

This work complemented previous analyses on the link between the linguistic and psychological accuracy of a neural language model by expanding the language sample to ten typologically distinct languages. However, our sample of neural language models was limited with respect to the literature focusing exclusively on English (Oh et al., 2022; Oh and Schuler, 2022; Shain et al., 2022). This problem cannot be overcome at the present state of affairs, since there are very few available massively multilingual auto-regressive language models, and the only one with sufficient coverage of our language sample was XGLM. This problem is an expression of a general difficulty in NLP to conduct experimental research on low-resource languages, due to the extreme skewness in the distribution of available resources (Joshi et al., 2020). However, we are confident that future developments in natural language engineering will support an additional test of our hypotheses with a more representative sample of models.

References

Suhas Arehalli, Brian Dillon, and Tal Linzen. 2022. Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. *arXiv preprint arXiv:2210.12187*.

Christoph Aurnhammer and Stefan L Frank. 2019. Comparing gated and simple recurrent neural network architectures as models of human sentence processing.

Annette Baumgaertner, Cornelius Weiller, and Christian Büchel. 2002. Event-related fmri reveals cortical sites involved in contextual sentence integration. *Neuroimage*, 16(3):736–745.

Damián E Blasi, Joseph Henrich, Evangelia Adamou, David Kemmerer, and Asifa Majid. 2022. Over-reliance on english hinders cognitive science. *Trends in cognitive sciences*.

Marisa Ferrara Boston, John Hale, Reinhold Kliegl, Umesh Patil, and Shravan Vasishth. 2008. Parsing costs as predictors of reading difficulty: An evaluation using the potsdam sentence corpus. *Journal of Eye Movement Research*, 2(1).

Trevor Brothers and Gina Kuperberg. 2020. Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language*, 116.

Harm Brouwer, Francesca Delogu, Noortje J Venhuizen, and Matthew W Crocker. 2021. Neurobehavioral correlates of surprisal in language comprehension: A neurocomputational model. *Frontiers in Psychology*, 12:615538.

Harm Brouwer, Hartmut Fitz, and John Hoeks. 2010. Modeling the noun phrase versus sentence coordination ambiguity in dutch: Evidence from surprisal theory. In *Proceedings of the 2010 workshop on cognitive modeling and computational linguistics*, pages 72–80.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Max Coltheart, Kathleen Rastle, Conrad Perry, Robyn Langdon, and Johannes Ziegler. 2001. Drc: a dual route cascaded model of visual word recognition and reading aloud. *Psychological review*, 108(1):204.

Andrea Gregor de Varda and Marco Marelli. 2022. The effects of surprisal across languages: Results from native and non-native reading. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 138–144.

Katherine A DeLong, Thomas P Urbach, and Marta Kutas. 2005. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature neuroscience*, 8(8):1117–1121.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Joseph Dien, Michael S Franklin, Charles A Michelson, Lisa C Lemen, Christy L Adams, and Kent A Kiehl. 2008. fmri characterization of the language formulation area. *Brain Research*, 1229:179–192.

Susan F Ehrlich and Keith Rayner. 1981. Contextual effects on word perception and eye movements during reading. *Journal of verbal learning and verbal behavior*, 20(6):641–655.

Xi Fan and Ronan Reilly. 2020. Reading development at the text level: an investigation of surprisal and embeddingbased text similarity effects on eyemovements in chinese early readers. *Journal of Eye Movement Research*, 13(6).

Irene Fernandez Monsalve, Stefan Frank, and Gabriella Vigliocco. 2012. Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 398–408, Avignon, France. Association for Computational Linguistics.

Stefan Frank and Rens Bod. 2011. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological science*, 22(6):829–834.

Stefan Frank and John CJ Hoeks. 2019. The interaction between structure and meaning in sentence comprehension. recurrent neural networks and reading times.

Stefan Frank, Leun J Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The erp response to the amount of information conveyed by words in sentences. *Brain and language*, 140:1–11.

Stefan Frank and Robin Thompson. 2012. Early effects of word surprisal on pupil size during reading. In *Proceedings of the annual meeting of the cognitive science society*, volume 34.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)*, pages 10–18.

Michael Hahn, Richard Futrell, Roger Levy, and Edward Gibson. 2022. A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, 119(43):e2122602119.

Yiding Hao, Simon Mendelsohn, Rachel Sterneck, Randi Martinez, and Robert Frank. 2020. Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. *arXiv preprint arXiv:2009.03954*.

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. 2021. Scaling laws for transfer.

Markus J Hofmann, Steffen Remus, Chris Biemann, Ralph Radach, and Lars Kuchinke. 2022. Language models explain word reading times better than empirical predictability. *Frontiers in Artificial Intelligence*, 4:214.

Jacob Louis Hoover, Morgan Sonderegger, Steven T Piantadosi, and Timothy J O'Donnell. 2022. The plausibility of sampling as an algorithmic theory of sentence processing.

Falk Huettig. 2015. Four central questions about prediction in language processing. *Brain research*, 1626:118–135.

Albrecht Werner Inhoff and Ralph Radach. 1998. Definition and computation of oculomotor measures in the study of cognitive processes. *Eye guidance in reading and scene perception*, pages 29–53.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 2021. Lower perplexity is not always human-like. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5203–5217.

Nayoung Kwon, Patrick Sturt, and Pan Liu. 2017. Predicting semantic features in chinese: Evidence from erps. *Cognition*, 166:433–446.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Danny Merkx and Stefan L Frank. 2021. Human sentence processing: Recurrence or attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 12–22.

Enrique Meseguer, Manuel Carreiras, and Charles Clifton. 2002. Overt reanalysis strategies and eye movements during the reading of mild garden path sentences. *Memory & cognition*, 30(4):551–561.

James A Michaelov, Seana Coulson, and Benjamin K Bergen. 2022. So cloze yet so far: N400 amplitude is better predicted by distributional information than human predictability judgements. *IEEE Transactions on Cognitive and Developmental Systems*.

Byung-Doh Oh, Christian Clark, and William Schuler. 2022. Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, 5.

Byung-Doh Oh and William Schuler. 2022. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *arXiv preprint arXiv:2212.12131*.

David C Plaut, James L McClelland, Mark S Seidenberg, and Karalyn Patterson. 1996. Understanding normal and impaired word reading: Computational principles in quasi-regular domains. In *Connectionist psychology: A text with readings*, pages 367–454. Psychology Press.

Ralph Radach and Alan Kennedy. 2013. Eye movements in reading: Some theoretical context. *Quarterly Journal of Experimental Psychology*, 66(3):429–452.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Philip Levy. 2022. Large-scale evidence for logarithmic effects of word predictability on reading time.

Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, et al. 2022. Expanding horizons of cross-linguistic research on reading: The multilingual eye-movement corpus (meco). *Behavior research methods*, pages 1–21.

Nathaniel J Smith and Roger Levy. 2008. Optimal processing times in reading: a formal model and empirical investigation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 30.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Adrian Staub. 2015. The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Language and Linguistics Compass*, 9(8):311–327.

Yuta Takahashi, Yohei Oseki, Hiromu Sakai, Michiru Makuuchi, and Rieko Osu. 2021. Identifying brain regions related to word prediction during listening to japanese speech by combining a lstm language model and meg. *bioRxiv*.

Jos JA Van Berkum, Colin M Brown, Pienie Zwitserlood, Valesca Kooijman, and Peter Hagoort. 2005. Anticipating upcoming words in discourse: evidence from erps and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(3):443.

Marten Van Schijndel and Tal Linzen. 2021. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive science*, 45(6):e12988.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. *arXiv preprint arXiv:2006.01912*.

Cai Wingfield and Louise Connell. 2022. Understanding the role of linguistic distributional knowledge in cognition. *Language, Cognition and Neuroscience*, pages 1–51.

A Effects of surprisal by language and model type

We report in Table 1 the regression coefficients of surprisal (as computed on the target word w_i), the t statistic and the associated p-value, divided by language, number of parameters, and fixation measure considered. The surprisal estimates obtained from the four XGLM models were statistically significant predictors of processing times in all the language $\times$ model combinations when considering GD and TT, and in the vast majority of the cases when considering FF as the dependent variable. These result are overall more solid than the ones obtained by de Varda and Marelli (2022), who did not report significant partial effects of surprisal on FF and GD in some of the languages considered. The authors derived their probabilistic estimates employing mBERT, a bidirectional encoder. This finding highlights the importance of employing standard left-to-right causal language models when studying the effects of predictability on incremental sentence processing.

B Relationship between perplexity and ΔLogLik

The perplexity of a model (Eq. 1) is commonly considered as an intrinsic measure of a language model's linguistic accuracy. The employment of perplexity as an evaluation of a multilingual language model is not free of concerns (see §4), but for completeness and consistency with the literature we also report the relationship between perplexity and ΔLogLik.

$$exp\left[-\frac{1}{N}\sum_{i=1}^{N}\log P(w_i|w_{1...i-1})\right] \quad (1)$$

We analyzed the relationship between perplexity and ΔLogLik by fitting three generalized additive

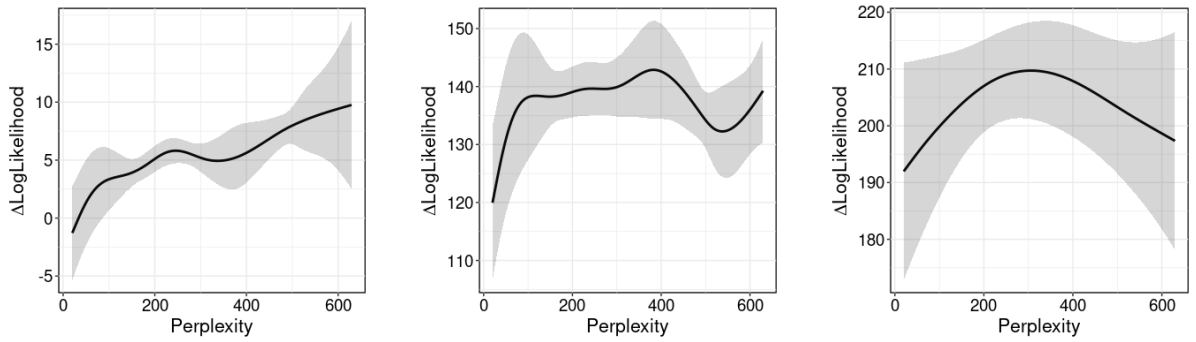
mixed models (GAMMs; one for each eye-tracking measure considered), with random slopes and intercepts for language. Note that the presence of by-language random effects mitigates the problem of comparing perplexity values with potentially different employed vocabularies.

The results are graphically depicted in Figure 2. In the case of FF (2a), we found a significant relationship between perplexity and ΔLogLik (EDF = 6.093, $F = 3.623$, $p = 0.0095$), which appears to be positive and (near)-linear from graphical inspection. In the case of GD (2b), we still found a significant partial effect of perplexity (EDF = 6.760, $F = 4.466$, $p = 0.0019$); however, the functional form of this relationship is far from linearity in this case, and is characterized by an initial growth in ΔLogLik with increasing perplexity, a local plateau, and an inversion of the trend in the 400-550 perplexity range. There is then a second inversion of the trend in the 500-600 perplexity range, although with high partial residuals. In the case of TT (2c), the relationship is clearly quadratic from graphical inspection, although the partial effect of perplexity is not statistically significant (EDF = 2.016, $F = 2.152$, $p = 0.123$).

Taken together, these results corroborate our observation that there is a negative relationship between the linguistic and the psychological accuracy of a model when considering the earliest fixation measurement, namely FF (§5); this relationship is less clear-cut when considering GD, and non-significant when considering TT. The very absence of a significant relationship between perplexity and ΔLogLik in this latter case demonstrates that the finding that smaller models outperform their over-parametrized counterparts in cognitive modelling critically depends on the computational complexity of the mental processes being analyzed.

C Cross-lingual variation in later measurements

The cross-lingual variation of our results increased with gaze duration and total reading time, in particular when considering XGLM_{564M}; our tentative explanation for this pattern is motivated by the fact that late eye-tracking measures subsume the early ones (FF < GD < TT). XGLM_{564M} is very effective at capturing early eye movement measurements (Figure 1a); some of the later measures are de facto equivalent to the earlier ones in some cases (e.g., if a word is only fixated once, FF, GD, and TT will have the same value). XGLM_{564M} might be more effective in modelling late eye tracking data in languages where these cases are more common, and less effective in languages where it is more common to refixate. This hypothesis relies on the observation in the MECO paper that refixations are more common in some languages than others (e.g., Estonian, see Siegelman et al., 2022).



(a) First fixation duration (b) Gaze duration (c) Total reading time

Figure 2: Relationship between perplexity and increase in model fit that could be ascribed to surprisal (ΔLogLik).

			First fixation duration			Gaze duration			Total reading time		
Language	Family	θ	Estimate	t	p	Estimate	t	p	Estimate	t	p
Finnish	Uralic	564M	0.0147	1.5527	0.1207	0.1118	12.5044	$3.87 \cdot 10^{-34}$	0.1567	16.8540	$2.35 \cdot 10^{-58}$
Greek	Indoeuropean	564M	0.0237	3.2279	0.0013	0.0777	9.8810	$1.66 \cdot 10^{-22}$	0.1147	15.2416	$1.11 \cdot 10^{-49}$
Korean	Koreanic	564M	0.0371	4.4060	$1.14 \cdot 10^{-05}$	0.0817	8.3948	$1.22 \cdot 10^{-16}$	0.1171	10.6387	$1.52 \cdot 10^{-25}$
Russian	Indoeuropean	564M	0.0300	3.7423	0.0002	0.0879	11.0399	$2.00 \cdot 10^{-27}$	0.1342	16.1478	$7.37 \cdot 10^{-55}$
Turkish	Turkic	564M	0.0126	1.3442	0.1791	0.0876	10.3490	$2.35 \cdot 10^{-24}$	0.1290	13.2625	$2.87 \cdot 10^{-38}$
English	Indoeuropean	564M	0.0248	3.6112	0.0003	0.0661	8.9182	$1.01 \cdot 10^{-18}$	0.0885	11.1290	$5.16 \cdot 10^{-28}$
Spanish	Indoeuropean	564M	0.0131	2.0597	0.0396	0.0554	8.5928	$1.57 \cdot 10^{-17}$	0.0713	9.9527	$6.85 \cdot 10^{-23}$
Estonian	Uralic	564M	0.0285	3.5439	0.0004	0.1437	17.0570	$7.57 \cdot 10^{-60}$	0.1764	20.9928	$3.12 \cdot 10^{-86}$
Italian	Indoeuropean	564M	0.0272	3.7723	0.0002	0.0987	13.0335	$2.37 \cdot 10^{-37}$	0.1108	13.8504	$8.62 \cdot 10^{-42}$
German	Indoeuropean	564M	0.0238	2.7832	0.0054	0.0954	10.4169	$8.55 \cdot 10^{-25}$	0.1361	15.3138	$3.39 \cdot 10^{-50}$
Finnish	Uralic	2.9B	0.0083	0.9303	0.3524	0.1073	12.7519	$2.27 \cdot 10^{-35}$	0.1530	17.5951	$5.65 \cdot 10^{-63}$
Greek	Indoeuropean	2.9B	0.0207	2.9912	0.0028	0.0744	10.0489	$3.32 \cdot 10^{-23}$	0.1037	14.5768	$8.41 \cdot 10^{-46}$
Korean	Koreanic	2.9B	0.0378	4.6397	$3.83 \cdot 10^{-06}$	0.0780	8.2755	$3.23 \cdot 10^{-16}$	0.1112	10.4415	$1.11 \cdot 10^{-24}$
Russian	Indoeuropean	2.9B	0.0209	2.7227	0.0065	0.0816	10.6812	$7.95 \cdot 10^{-26}$	0.1313	16.5508	$2.49 \cdot 10^{-57}$
Turkish	Turkic	2.9B	0.0096	1.0695	0.2850	0.0903	11.1858	$4.66 \cdot 10^{-28}$	0.1350	14.6822	$4.57 \cdot 10^{-46}$
English	Indoeuropean	2.9B	0.0180	2.7426	0.0062	0.0593	8.3905	$8.81 \cdot 10^{-17}$	0.0800	10.5037	$3.37 \cdot 10^{-25}$
Spanish	Indoeuropean	2.9B	0.0100	1.6592	0.0972	0.0474	7.7710	$1.17 \cdot 10^{-14}$	0.0600	8.8493	$1.67 \cdot 10^{-18}$
Estonian	Uralic	2.9B	0.0160	2.0700	0.0386	0.1397	17.2667	$3.91 \cdot 10^{-61}$	0.1695	20.9452	$7.97 \cdot 10^{-86}$
Italian	Indoeuropean	2.9B	0.0217	3.1766	0.0015	0.0852	11.7870	$4.51 \cdot 10^{-31}$	0.0981	12.8402	$2.24 \cdot 10^{-36}$
German	Indoeuropean	2.9B	0.0188	2.4136	0.0159	0.0849	10.1956	$7.62 \cdot 10^{-24}$	0.1278	15.9417	$5.10 \cdot 10^{-54}$
Finnish	Uralic	1.7B	0.0166	1.8208	0.0689	0.1079	12.5299	$2.91 \cdot 10^{-34}$	0.1511	16.8523	$2.44 \cdot 10^{-58}$
Greek	Indoeuropean	1.7B	0.0188	2.6711	0.0076	0.0694	9.1802	$1.05 \cdot 10^{-19}$	0.1015	13.9720	$2.18 \cdot 10^{-42}$
Korean	Koreanic	1.7B	0.0361	4.3679	$1.35 \cdot 10^{-05}$	0.0804	8.4397	$8.62 \cdot 10^{-17}$	0.1115	10.3313	$3.21 \cdot 10^{-24}$
Russian	Indoeuropean	1.7B	0.0280	3.6187	0.0003	0.0835	10.8437	$1.52 \cdot 10^{-26}$	0.1332	16.6544	$5.51 \cdot 10^{-58}$
Turkish	Turkic	1.7B	0.0160	1.7717	0.0766	0.0927	11.4122	$4.25 \cdot 10^{-29}$	0.1390	15.0582	$3.17 \cdot 10^{-48}$
English	Indoeuropean	1.7B	0.0218	3.2646	0.0011	0.0614	8.5422	$2.50 \cdot 10^{-17}$	0.0851	11.0218	$1.62 \cdot 10^{-27}$
Spanish	Indoeuropean	1.7B	0.0117	1.9096	0.0563	0.0500	8.0723	$1.11 \cdot 10^{-15}$	0.0650	9.4580	$7.24 \cdot 10^{-21}$
Estonian	Uralic	1.7B	0.0213	2.7109	0.0068	0.1427	17.4455	$2.85 \cdot 10^{-62}$	0.1757	21.5929	$1.89 \cdot 10^{-90}$
Italian	Indoeuropean	1.7B	0.0225	3.2373	0.0012	0.0905	12.3422	$8.42 \cdot 10^{-34}$	0.0990	12.7184	$9.68 \cdot 10^{-36}$
German	Indoeuropean	1.7B	0.0226	2.8348	0.0046	0.0914	10.7334	$3.47 \cdot 10^{-26}$	0.1328	16.1677	$2.01 \cdot 10^{-55}$
Finnish	Uralic	4.5B	0.0065	0.7063	0.4801	0.1082	12.3959	$1.33 \cdot 10^{-33}$	0.1539	16.9725	$4.49 \cdot 10^{-59}$
Greek	Indoeuropean	4.5B	0.0140	2.0330	0.0422	0.0671	9.0501	$3.32 \cdot 10^{-19}$	0.0979	13.7244	$4.97 \cdot 10^{-41}$
Korean	Koreanic	4.5B	0.0345	4.1755	$3.16 \cdot 10^{-05}$	0.0845	8.8906	$2.05 \cdot 10^{-18}$	0.1201	11.2087	$4.63 \cdot 10^{-28}$
Russian	Indoeuropean	4.5B	0.0189	2.4883	0.0129	0.0742	9.7792	$5.17 \cdot 10^{-22}$	0.1206	15.1986	$3.92 \cdot 10^{-49}$
Turkish	Turkic	4.5B	0.0088	0.9803	0.3271	0.0876	10.8771	$1.14 \cdot 10^{-26}$	0.1290	14.0001	$2.96 \cdot 10^{-42}$
English	Indoeuropean	4.5B	0.0234	3.5854	0.0003	0.0591	8.3770	$9.82 \cdot 10^{-17}$	0.0790	10.3938	$1.01 \cdot 10^{-24}$
Spanish	Indoeuropean	4.5B	0.0092	1.5360	0.1247	0.0446	7.3498	$2.75 \cdot 10^{-13}$	0.0569	8.4442	$5.21 \cdot 10^{-17}$
Estonian	Uralic	4.5B	0.0232	2.9332	0.0034	0.1446	17.5013	$1.14 \cdot 10^{-62}$	0.1787	21.8217	$3.29 \cdot 10^{-92}$
Italian	Indoeuropean	4.5B	0.0227	3.3715	0.0008	0.0800	11.1795	$3.31 \cdot 10^{-28}$	0.0922	12.1985	$4.12 \cdot 10^{-33}$
German	Indoeuropean	4.5B	0.0126	1.6572	0.0976	0.0791	9.6844	$1.02 \cdot 10^{-21}$	0.1210	15.3631	$1.70 \cdot 10^{-50}$

Table 1: Effects of surprisal across languages on the three fixation measurements considered. The first three columns indicate the language from which the reading data were obtained, the corresponding language family, and the number of parameters of the model considered. The following columns indicate the regression coefficients of surprisal, the t statistic and the respective p-value for FF, GD and TT.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section "Limitations" (unnumbered, page 5)

☒ A2. Did you discuss any potential risks of your work?
There are no reasonable risks in our work.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract (unnumbered), Introduction (§1), Related work and motivation (§2), Aims (§3)

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

§4.1, §4.2

☑ B1. Did you cite the creators of artifacts you used?
§4.1, §4.2

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No, due to space restrictions. However, the artifacts that we employed were publicly released for research purposes.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No; however, the artifacts were employed in accordance with their intended use.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No, but the authors of the artifacts did, and we provided a reference to the original article.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
§4.1

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
§4.3

C ☑ Did you run computational experiments?

§4, §5

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
We did report the number of parameters (§4.2) but not the computational budget or the computing infrastructure as we did not train the models ourselves.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Not applicable. Left blank.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
§5

☒ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
We used the defaults parameters of the transformers library.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.