**Linguistics Should Not Stop Worrying About Languages Other Than English: A Commentary on Futrell and Mahowald (2025)**

*Bjarki Ármannsson, Iris Edda Nowenstein and Einar Freyr Sigurðsson*

Although we agree with Futrell and Mahowald's (F&M) overall stance as we understand it—that language models (LMs) can serve as tools for investigating features of human language and that theoretical and computational linguistics can inform each other going forward—we question the validity and generalizability of some of F&M's conclusions cross-linguistically. Just as the language sciences (not least early work in generative grammar) have been criticized for basing assumptions primarily on examples from English (Evans and Levinson 2009, Blasi et al. 2022), some of F&M's key takeaways (notably, that LMs' success "strengthens the case for existing usage-based language theories") are almost exclusively based on research on model capabilities in English, impacting the validity of LMs as a language-agnostic model of human language. This is (briefly) recognised as a "caveat" at the outset, but no effort is made to engage with the obvious argument that spending years building statistical models that optimize primarily over English text data can easily lead to overfitting to English linguistic structure (an argument made at least as early as Bender (2011) about machine-learning systems in general). This argument is not only theoretical: previous work has produced examples where LMs show a clear preference for English-like grammars over non-English-like ones (Dyer et al. 2019, Davis and van Schijndel 2020, Davis 2022). Most egregiously, in section 4.4, the terms "real English text" and "human language" are practically treated as equivalent when discussing the learning of "impossible languages".

This clearly impacts the discussion of whether LMs can be considered suitable models of human language. While F&M discuss how autoregressive LMs "show a bias towards information locality", obtained through their next-word prediction training set-up, and that this bias "helps language learning and is likely shared with humans", it is by no means clear that all languages are equally well-suited to the next-word prediction objective. Indeed, recent work shows that certain very robust features of Icelandic are not effectively learned by LMs trained on more Icelandic data than a human is exposed to during the language acquisition phase, likely by orders of magnitude.

In Ármannsson et al. (2026), we find that state-of-the-art LMs (as of 2024) fall remarkably short of human accuracy for robust noun-phrase internal predicate-agreement patterns in Icelandic when rating clearly grammatical/ungrammatical sentences, often preferring ungrammatical gender agreement with a pre-adjectival subject in the presence of a post-adjectival predicate. While native speakers showed 99% acceptance for the grammatical sentences and 0.3% for ungrammatical ones, the best-performing model recorded 57% accuracy and almost all fell below chance (Fig. 1). In addition, assuming that children are exposed to roughly seven to ten million words per year (Gilkerson et al. 2017), this agreement pattern seems robust in children 'trained' on no more than twenty million words, presumably a fraction of the models' Icelandic training data (the largest corpus of Icelandic available online, the Icelandic Gigaword Corpus (IGC; Barkarson et al. 2022), contains approximately 2.6 billion words).

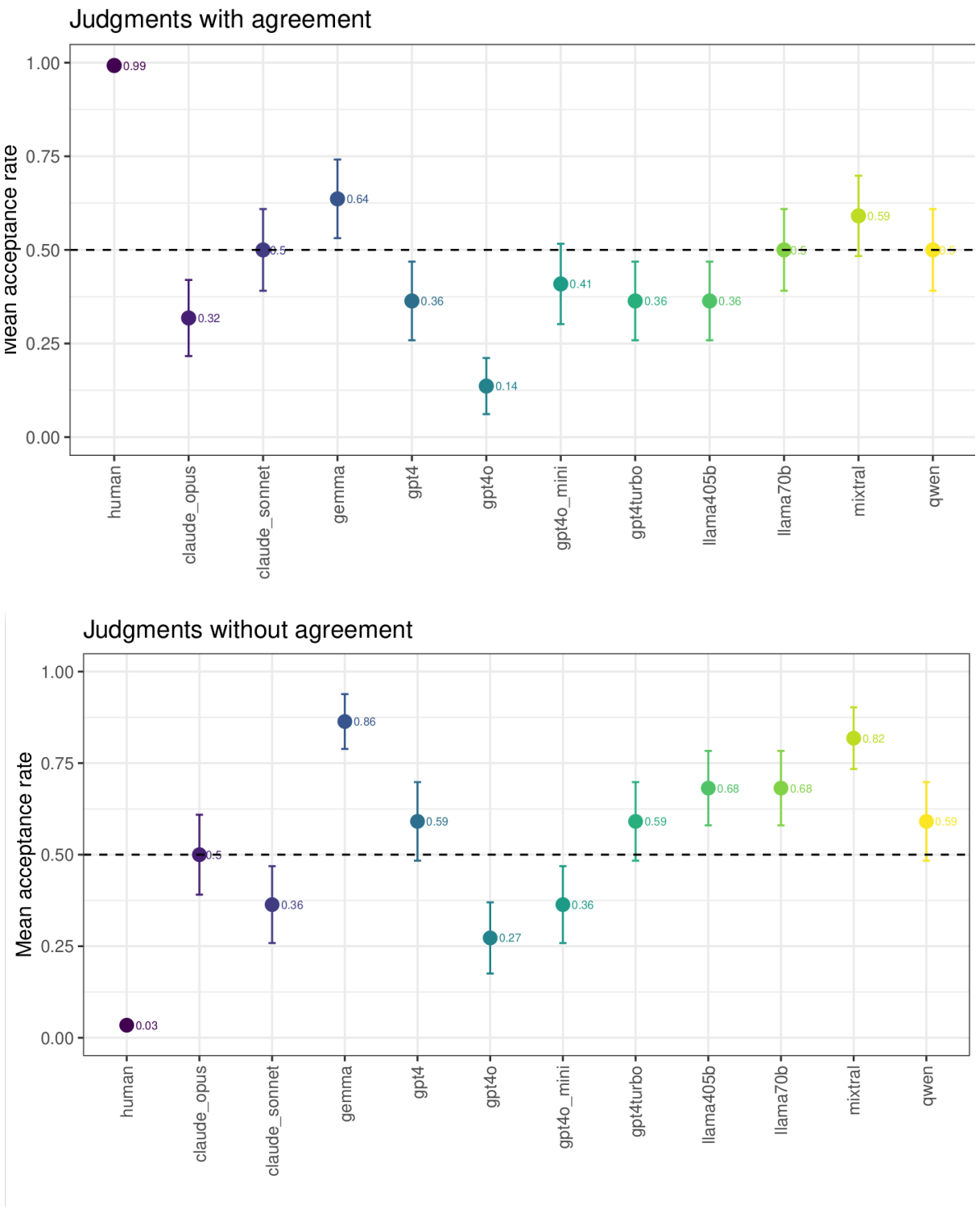


*Figure 1: The mean acceptance rate of sentences in the gender agreement grammaticality judgment tasks in Ármannsson et al. (2026), shown for only those sentences that contained noun-phrase internal predicate agreement (above) and those that did not (below). Native speakers categorically accept sentences containing noun-phrase internal predicate agreement and reject those without, while no model reliably matches the judgments of human participants.*

While the suitability of LMs as models of human language obviously does not hinge on this or any other single experiment, it is worth contemplating that F&M, and much of the literature they refer to, focus on work which shows LMs successfully modelling—to various degrees—"nontrivial formal linguistic patterns" in English (often quite rare constructions, e.g., Misra and Mahowald 2024). Meanwhile, examples like ours show massive models failing to successfully generalize a ubiquitous pattern which appears to be both *categorical* in native speakers and *unanimous* in corpora of Icelandic (we found zero counterexamples in our search of the morphologically-tagged IGC). This is in direct contrast with, e.g., the commonly cited work of Linzen et al. (2016), who probe models through an English pattern which is known to be variable in speakers (or in F&M's framing, "human accuracy in producing the right verb forms [...] is 85%") and dependent on various syntactic and semantic factors (den Dikken 2001).

Overall, F&M make an obvious and commendable effort to balance their arguments with the necessary caveats, which is why we find it surprising that they eventually arrive at the previously referenced "proof of concept" conclusion. Usage-based theories of language, as all theories of language, are meant to be language-agnostic. (They are also meant to account for how language is learned with the time and input available to a child learner, which remains the most urgent obstacle to treating LMs as theories of language. In our view, this significant issue is not addressed well enough by F&M but we leave this line of argument to other commentators.) There is a major leap between the well-argued claim that LMs "have learned nontrivial formal linguistic patterns better than many expected was possible" and the conclusion that their success "strengthens the case for existing usage-based language theories based on gradient representations". In order to make this leap, F&M would need to present considerably more cross-linguistic evidence than they do and extend their argument against modern computational and neurolinguistic work rooted in generative linguistics (e.g., Belth et al. 2021, Payne et al. 2021, Gwilliams et al. 2025). This work is cited to some extent, but not included under the umbrella of generative linguistics, perhaps because it does not neatly fit into a more simplistic binary characterization pitting a usage-based/statistical tradition against a generative/"anti-statistical" one.

LMs' successful performance in processing and outputting English text can lead to overhyped claims about their approximation of the human language acquisition process in general, across any language. Similarly, it can lead to claims that LMs provide us with clear answers in debates opposing usage-based and generative theories of language, despite a growing body of work showing parallels between various aspects of linguistic theory, including generative ones, and LM performance (e.g., Caucheteux and King 2022, Desbordes et al. 2023, Simon et al. 2024). Hopefully, future work will focus more on the nuances present in F&M's target article and less on reductive claims about LM performance and its impact on modern linguistic theory.

**References**

Ármannsson, B., Ingimundarson, F. Á., Nowenstein, I. E., & Sigurðsson, E. F. (2026) Icelandic Gender Agreement Shows LLMs' Failure to Effectively Generalise. To appear in *University of Pennsylvania Working Papers in Linguistics* 32(1).

Barkarson, S., Steingrímsson, S., & Hafsteinsdóttir, H. (2022). Evolving large text corpora: Four versions of the Icelandic Gigaword Corpus. *Proceedings of the Thirteenth Language Resources and Evaluation Conference,* 2371–2381. European Language Resources Association. https://aclanthology.org/2022.lrec-1.254/.

Belth, C., Payne, S., Beser, D., Kodner, J., & Yang, C. (2021). The greedy and recursive search for morphological productivity. *The 43th Annual Meeting of the Cognitive Science Society (CogSci),* 2869–2875.

Bender, E. M. (2011). On Achieving and Evaluating Language-Independence in NLP. *Linguistic Issues in Language Technology* 6(3), 1-26. https://doi.org/10.33011/lilt.v6i.1239.

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences* 26(12), 1153–1170. https://doi.org/10.1016/j.tics.2022.09.015.

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Commun Biol* 5, 134. https://doi.org/10.1038/s42003-022-03036-1

Davis, F. L. (2022). *On the limitations of data: Mismatches between neural models of language and humans.* Unpublished doctoral dissertation, Cornell University.

Davis, F., & van Schijndel, M. (2020). Recurrent neural network language models always learn English-like relative clause attachment. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1979-1990. Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.179/.

den Dikken, M. (2001). “Pluringulars”, pronouns and quirky agreement. *The Linguistic Review* 18, 19-41. https://doi.org/10.1515/tlir.18.1.19.

Desbordes, T., Lakretz, Y., Chanoine, V., Oquab, M., Badier, J.-M., Trébuchon, A., Carron, R., Bénar, C.-G., Dehaene, S., & King, J.-R. (2023). Dimensionality and ramping: Signatures of sentence integration in the dynamics of brains and deep language models. *Journal of Neuroscience* 43(29), 5350–5364. https://doi.org/10.1523/JNEUROSCI.1163-22.2023.

Dyer, C., Melis, G., & Blunsom, P. (2019). A critical analysis of biased parsers in unsupervised parsing. arXiv:1909.09428v. https://arxiv.org/pdf/1909.09428.

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences* 32(5), 429–448. https://doi.org/10.1017/S0140525X0999094X.

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Kimbrough Oller, D., Hansen, J. H. L., & Paul, T. D. (2017). Mapping the Early Language Environment Using

All-Day Recordings and Automated Analysis. *American journal of speech-language pathology*, *26*(2), 248–265. https://doi.org/10.1044/2016_AJSLP-15-0169.

Gwilliams, L., Marantz, A., Poeppel, D., & King, J.-R. (2025). Hierarchical dynamic coding coordinates speech comprehension in the human brain. *Proc. Natl. Acad. Sci. U.S.A.* 122 (42) e2422097122. https://doi.org/10.1073/pnas.2422097122.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics* 4, 521-535. https://aclanthology.org/Q16-1037/.

Misra, K., & Mahowald, K. (2024). Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 913–929. Association for Computational Linguistics. https://aclanthology.org/2024.emnlp-main.53/.

Payne, S., Kodner, J., & Yang, C. (2021). Learning morphological productivity as meaning-form mappings. *Proceedings of the Society for Computation in Linguistics 2021*, 177-187. Association for Computational Linguistics. https://aclanthology.org/2021.scil-1.17/.

Simon, P. J. D., d'Ascoli, S., Chemla, E., Lakretz, Y., & King, J.-R. (2024). A Polar coordinate system represents syntax in large language models. *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=x2780VcMOI.