Testing the Linguistic Niche Hypothesis in Large Language Models with a Multilingual Wug Test

Anh Dang[1,2,*], Limor Raviv[1,3], and Lukas Galke[1]

[*]Corresponding Author: thithaoanh.dangthithaoanh@ru.nl
[1]LEADS group, Max Planck Institute for Psycholinguistics, Nijmegen, NL
[2]CLS, Radboud University, NL
[3]cSCAN, University of Glasgow, UK

The linguistic niche hypothesis states that languages spoken by larger societies tend to have less complex morphological systems (Lupyan & Dale, 2010), which may be caused by a learnability advantage of L2 learners for less complex systems (Wray & Grace, 2007; Hudson Kam & Newport, 2009). Despite the high impact of this theory on the field of language evolution and adaptation (Gibson et al., 2019; Bentz et al., 2015, 2018; Lupyan & Dale, 2016), recent studies (Koplenig et al., 2023; Shcherbakova et al., 2023) challenge the linguistic niche hypothesis and suggest an opposite relationship between morphological complexity and population size, whereby larger societies actually have more complex morphological systems. Here, we test the underlying assumption that languages with less complex morphological systems are easier to learn for language models. To this end, we evaluate to what extent morphological generalization is influenced by linguistic complexity and population size in a new type of learner: large language models (LLMs). Testing cross-linguistic patterns of language learning in LLMs trained on large amounts of human-generated text is particularly interesting given recent findings highlighting the similarity between humans and such models with respect to language learning and processing (Galke et al., 2023; Webb et al., 2023; Srikant et al., 2022) and to the emergence of syntactic structure within the model's learned attention patterns (Manning et al., 2020). While there is little cross-linguistic work on the morphological knowledge of LLMs in relation to the degree of morphological structure, some work suggests that LLMs often fail to generate the correct inflected forms of words that are not present in the training data, regardless of the size of the training set and the target language (Liu & Hulden, 2022). As such, it is currently unknown to what extent LLMs can learn to generalize their morphological knowledge and to what extent their generalization capabilities are affected by the degree of linguistic complexity in their input.

In our study, we developed a multilingual version of the Wug Test, an artificial word completion test that is typically used to evaluate the inflectional and derivational morphological knowledge of children (Berko, 1958), and applied it to the

GPT family of large language models (Brown et al., 2020; Ouyang et al., 2022). We considered six different languages, namely German, Vietnamese, Spanish, French, Romanian, and Portuguese, which vary in their degree of morphological complexity and well as the amount of text available for them. For each language, we first asked GPT-4 to translate the questions from the original Wug Test – translations that were then evaluated and corrected by native speakers. Then, LLMs were provided with the translated questions (i.e., a sentence in which the fantasy word, e.g., 'wug', represents either a noun or a verb), and were made to respond with the inflected form (e.g,. plural form, past tense). Since the fantasy words (very likely) do not exist in the respective training data, the models needed to use their morphological knowledge of the language in order to be successful. The models' answers were then evaluated by native speakers, who judged whether the generated inflected and derived forms conform to their native language's morphological rules (see Additional File for examples).

To connect our results with the linguistic niche hypothesis, we test whether accuracy was predicted by morphological complexity and training size, taking into account Ackerman and Malouf (2013) distinction between e-complexity (the number of rules and irregularities) and i-complexity (how well are morphemes predicted by their context). E-complexity was measured using Lupyan and Dale (2010)'s original complexity scores (LNH in the table), and i-complexity was measured using Bentz et al. (2015)'s lexical diversity score, based on Shannon entropy (H_{scaled}).

Language	**%train**	**LNH**	$\mathbf{H_{scaled}}$	**Model**	**Correct**	**Unclear**	**Wrong**
German	1.68%	-12	0.4648	GPT-3.5	66%	5%	29%
				GPT-4	62%	5%	33%
Vietnamese	0.03%	-16	-1.2099	GPT-3.5	71%	0%	19%
				GPT-4	81%	0%	19%

The table shows the results for German and Vietnamese. We find that while both GPT-3.5 and GPT-4 were generally capable of generating the correct inflected forms for unknown words, they were not always able to inflect them correctly. Notably, our initial results are promising: Despite German having 50 times more representation than Vietnamese in GPT-3's training data (1.67583% compared to 0.03373%), the model scores higher on the less complex (w.r.t. LNH and H_{scaled}) Vietnamese morphological system – indicating that less complex morphological systems are learned better by LLMs, even given much less data. Our findings thus provide a first indication that multilingual LLMs satisfy the underlying assumption of the linguistic niche hypothesis – i.e., that languages with more complex morphologies are harder to learn.

References

Ackerman, F., & Malouf, R. (2013). Morphological organization: The low conditional entropy conjecture. *Language*, *89*(3), 429–464.

Bentz, C., Dediu, D., Verkerk, A., & Jäger, G. (2018). The evolution of language families is shaped by the environment beyond neutral drift. *Nature Human Behaviour*, *2*(11), 816–821.

Bentz, C., Verkerk, A., Kiela, D., Hill, F., & Buttery, P. (2015). Adaptive Communication: Languages with More Non-Native Speakers Tend to Have Fewer Word Forms. *PLOS ONE*, *10*(6), e0128254.

Berko, J. (1958). The child's learning of english morphology. *Word*, *14*(2-3), 150–177.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc.

Galke, L., Ram, Y., & Raviv, L. (2023, February). *What makes a language easy to deep-learn?* (No. arXiv:2302.12239). arXiv.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How Efficiency Shapes Human Language. *Trends in Cognitive Sciences*, *23*(5), 389–407.

Hudson Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, *59*(1), 30–66.

Koplenig, A., Wolfer, S., & Meyer, P. (2023). A large quantitative analysis of written language challenges the idea that all languages are equally complex. *Scientific Reports*, *13*(1), 15351.

Liu, L., & Hulden, M. (2022). Can a Transformer Pass the Wug Test? Tuning Copying Bias in Neural Morphological Inflection Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 739–749). Dublin, Ireland: Association for Computational Linguistics.

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PLoS ONE*, *5*(1), e8559.

Lupyan, G., & Dale, R. (2016). Why Are There Different Languages? The Role of Adaptation in Linguistic Diversity. *Trends in Cognitive Sciences*, *20*(9), 649–660.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-