Exploring Linguistic Priors in Large Language Models: An Investigation into Artificial Language Learnability Through Verb Conjugation Paradigms

Author Anonymized
Institution Anonymized

Recently, much research has been published probing linguistic knowledge from LLMs trained on natural language (Liu 2019; Manning 2020), expanding our knowledge of how computational models encode language. However, limited research probes LLMs for linguistic knowledge while controlling for language-specific idiosyncrasies. This study examines the learnability of artificial languages, investigating potential structural priors in LLMs favoring specific conjugation paradigms.

Our study investigates subject-verb agreement learnability in artificial languages. We finetune the pretrained GPT-2 model on stochastically generated sentences from three artificial languages, each representing a different verb conjugation type: one regular paradigm, non-suppletive allomorphy (in suffixes), and two verb classes, where the suffix type is unpredictable from the verb root. Accuracy is defined as the percentage of test sentences where the incorrectly conjugated sentence's perplexity exceeds that of the correctly conjugated sentence. We compare *pernum* errors (subject-verb agreement) and *class* errors (incorrect affix or allomorph class). Testing covers seen verb roots, unseen verb roots, and one-shot generalization on unseen verb roots.

Surprisingly, we find that for *pernum* errors, across all testing types, our non-suppletive-allomorphy model outperforms our one-regular-paradigm model and two-verb-classes model, with the latter two performing approximately equally well (Figure 1). Additionally, the two-verb-classes model seems unable to generalize verb conjugations in the one-shot setting, with performance decreasing with more training data (Figure 2).

These results suggest that languages whose affixes demonstrate allomorphy are more learnable by GPT architectures. We predict this is due to the allomorphs resulting in all verbs sharing similar syllable structures, providing an additional indicator for the model when targeting verbs. The regular-paradigm and two-verb-classes settings appear to be approximately equally learnable, suggesting that modeling verb classes doesn't pose additional difficulties. The two-verb-classes model's low performance in the one-shot setting suggests that these models struggle to generalize paradigms for roots not in the training data.

Figures

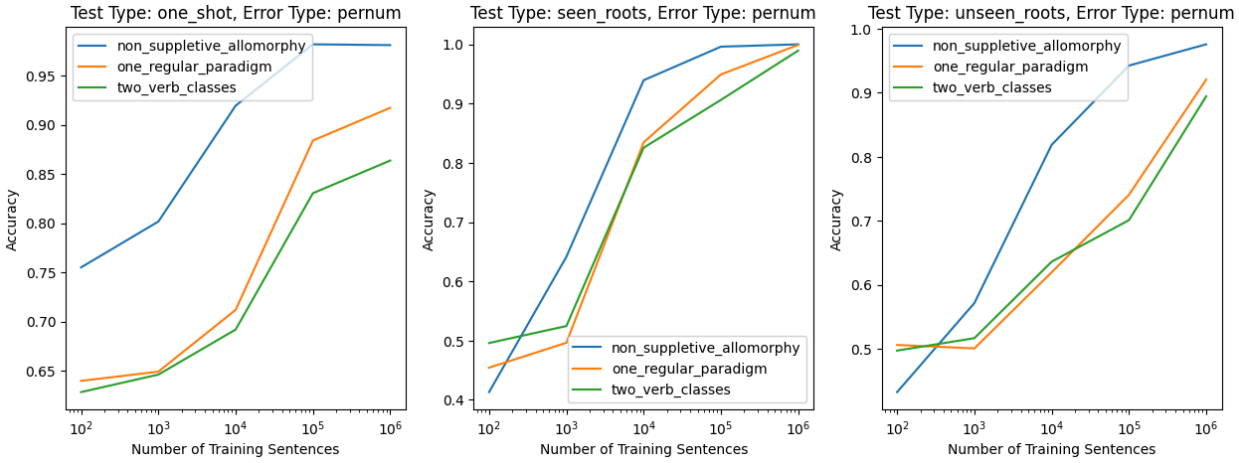

Figure 1: Accuracy for various Training Amounts for the *pernum* error type.

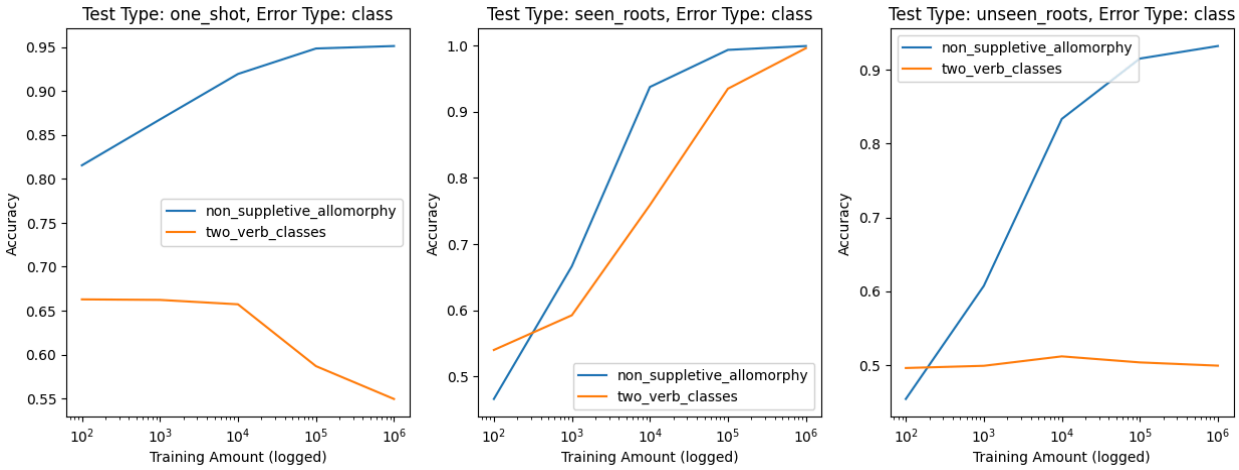

Figure 2: Accuracy for various Training Amounts for the *class* error type.

Citations

Liu, Nelson F., et al. "Linguistic knowledge and transferability of contextual representations." *Proceedings of the 2019 Conference of the North*, 25 Apr. 2019, https://doi.org/10.18653/v1/n19-1112.

Manning, Christopher D., et al. "Emergent linguistic structure in artificial neural networks trained by self-supervision." *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, 3 June 2020, pp. 30046–30054, https://doi.org/10.1073/pnas.1907367117.