Evaluating Syntactic Generalization in Multilingual Language Models through Targeted Test Suites

Marie Dewulf[1,2]

[1] University of Antwerp; [2] Ghent University

Marie.Dewulf@uantwerpen.be

Low- and medium-resource language scenarios pose significant challenges for the development and evaluation of neural language models (LMs), particularly in capturing syntactic generalizations beyond surface-level regularities. Specifically, training corpora are predominantly in English, and existing evaluation methods may be biased towards English, potentially overlooking limitations in multilingual linguistic knowledge of models. This work proposes an evaluation framework inspired by [1, 2] to systematically assess syntactic competence in French, a morphosyntactically rich and medium-resource language in the context of large-scale model pretraining. Building on targeted syntactic evaluation methodologies [3, 4], I develop diagnostic test suites that assess models' sensitivity to morphosyntactic and syntactic phenomena in French, extending existing English-focused resources.

My research aims to answer the following questions: (1) To what extent do multilingual and French-specific language models acquire abstract syntactic representations in a medium-resource training context? (2) How consistent are these representations across morphosyntactic and syntactic features? (3) Does morphosyntactic competence primarily depend on model size or training data?

I constructed a benchmark of 500 minimal pairs based on representative examples in grammar books specifically targeted towards English-speaking L2 learners of French. These minimal pairs contrast grammatical and ungrammatical sentences and cover multiple syntactic phenomena such as agreement (within verb and noun phrases), argument structure (e.g., direct vs. indirect object), tense and mood, question formation, negation and preposition use. For example, in the subject-verb agreement suite, verbs are varied to match or mismatch with the subject in person and number (e.g., *Tu travailles* vs. **Tu travaillent*). Each test item appears in controlled conditions that isolate syntactic dependencies. Surprisal values, calculated as the negative log probability of a token given the preceding context, are computed at the sentence-level. The syntactic sensitivity of the model is quantified by the extent to which predicted surprisal differences match theoretical expectations: grammatical sentences should be less surprising than ungrammatical ones.

Results show that CroissantLLMBase [5], a smaller bilingual model trained on balanced English-French datasets, outperforms larger multilingual models such as Llama-2 (7B and 13B) [6] on several syntactic dimensions, particularly argument structure, negation and preposition usage. This demonstrates that linguistically accurate behavior in LMs is not solely determined by model size or architecture but rather depends on the quantity and linguistic diversity of the training data.

References

[1] Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8, 377–392. https://doi.org/10.1162/tacl_a_00321

[2] Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. (2020). A Systematic Assessment of Syntactic Generalization in Neural Language Models. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1725–1744). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.158

[3] Marvin, R., & Linzen, T. (2018). Targeted Syntactic Evaluation of Language Models. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 1192–1202). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1151

[4] Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 32–42). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1004

[5] Faysse, M., Fernandes, P., Guerreiro, N. M., Loison, A., Alves, D. M., Corro, C., Boizard, N., Alves, J., Rei, R., Martins, P. H., Casademunt, A. B., Yvon, F., Martins, A. F. T., Viaud, G., Hudelot, C., & Colombo, P. (2024). *CroissantLLM: A Truly Bilingual French-English Language Model* (No. arXiv:2402.00786). arXiv. https://doi.org/10.48550/arXiv.2402.00786

[6] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … Scialom, T. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models* (No. arXiv:2307.09288). arXiv. https://doi.org/10.48550/arXiv.2307.09288