Structural sensitivity does not entail grammaticality: assessing LLMs against UHFH

Tommaso Sgrizzi[1],[2], Asya Zanollo [1],[2] , Cristiano Chesi[1],[2]
[1]University School for Advanced Studies IUSS Pavia; [2]NEtS - IUSS
asya.zanollo@iusspavia.it

The study explores whether Large Language Models (LLMs) can generalize the universal hierarchy of functional heads (UHFH), a cross-linguistic syntactic pattern argued to be rooted in human cognition (Cinque 1999). In formal linguistics, the consistent word order of some expressions across languages has been argued to originate from the cognitive layout of our language faculty (see also Scontras et al. 2017). A standard case in point is the relative order between epistemic and aspectual adverbs, which across languages, obey the order Epistemic $>$ Aspectual (Cinque 1999).

(1) a. *John **probably** has **again** read the book.*

b. **John **again** has **probably** read the book.*

Other domains of grammar express similar ordering constraints, as the nominal domain (Cinque 2005), and the verbal domain (Cinque 2006). We focus on Italian Restructuring Verbs (RVs), which exhibit a distinct syntactic behavior from control verbs (CVs), particularly concerning clitic climbing (movement of an object clitic from the infinitive to the matrix verb). RVs (unlike CVs) are theorized to occupy fixed positions within UHFH (e.g., aspectual > modal > mood; cf. Cinque 2006).

Four LLMs were evaluated: Mistral-7B-v0.3, GPT2-small, GePpeTto, and Minerva-7B-base-v1.0. GePpeTto and Minerva-7B-base-v1.0 were trained on Italian corpora, while Mistral-7B-v0.3 and GPT2-small were primarily trained on English data. The study addressed three research questions: **RQ1:**To what extent do LLMs generalize the verb ordering hierarchy proposed by Cinque (2006) for RVs? **RQ2:** Can LLMs differentiate the underlying structural ambiguity inherent in restructuring versus control verb constructions (Wurmbrand 2001)? **RQ3:** What is the syntactic structure assigned by LLMs to novel verbs which introduce non-finite complements? To investigate these questions, 14 minimal pair experiments were designed, targeting clitic placement, auxiliary selection, and multiple sequences of restructuring verbs. The experiments manipulated restructuring environments, matrix verb types (RVs, CVs, and pseudo-verbs), and structural distance between RVs. Three pseudo-verbs (*grabbare, drommare a, trellare di*) were introduced to test how LLMs assign syntactic behavior to novel lexical items. Specific experiments were designed to address each research question: **RQ1** was investigated by constructing minimal pairs of verb sequences that either respected or violated Cinque's UHFH (Exp. 1, 3). Exp. 2 and 4 included proclitic clitics to evaluate the influence of clitic placement. **RQ2** was explored by pairing CVs with RVs in different orders, examining clitic placement, as CVs block clitic climbing (Exp. 5, 6). **RQ3** was investigated by combining pseudo-verbs with restructuring and CVs, assessing whether LLMs treated novel verbs as compatible with restructuring based on proclisis (Exp. 7, 8, 9, 10, 11, 15). Auxiliary selection with pseudo-verbs was also tested (Exp. 12, 13, 14). The evaluation used the *LM-eval* platform, generating 610,500 minimal pairs. A generalized linear mixed model was fitted to analyze model performance. We present results from Exp.3 and 4., which indicate that LLMs, particularly GePpeTto, show sensitivity to the UHFH proposed by Cinque. Model behavior aligned with hierarchical orderings, as sentence preference correlated with greater structural distance between verbs, regardless grammaticality (Fig.1 and 2). Failure in selecting the grammatical option increased

especially when clitic climbing was involved (Tab.1). These results suggest that hierarchical awareness acts more as a heuristic than as a diagnostic of grammatical well-formedness. Notably, GePpeTto, a smaller GPT-2-style model trained on Italian data, outperformed all larger models, including 7B-parameter Mistral and Minerva architectures (Tab.1). This finding challenges the common assumption that larger size universally yields better syntactic abstraction and suggests that language-specific training and vocabulary alignment may play a more decisive role in domains relying on typologically grounded syntactic contrasts. The study concludes that LLMs can learn structural tendencies aligned with linguistic hierarchies, but these are not necessarily associated with a proper grammaticality judgment.

This page serves for Tables, Figures, and References.

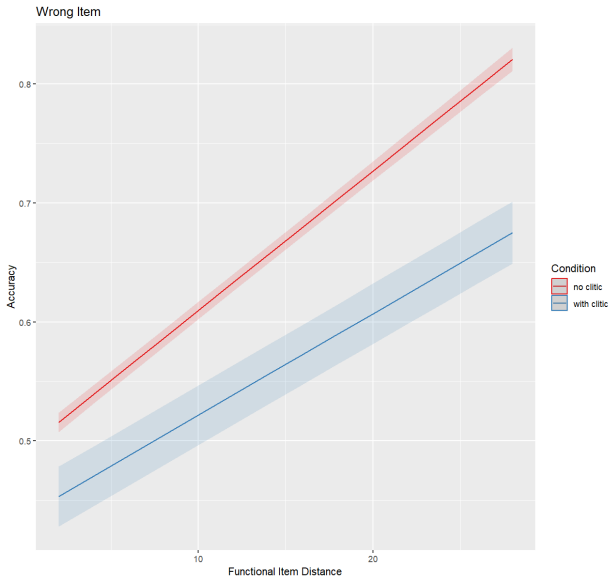

Figure 1: Acc. on ungrammatical triplets (wrong item)

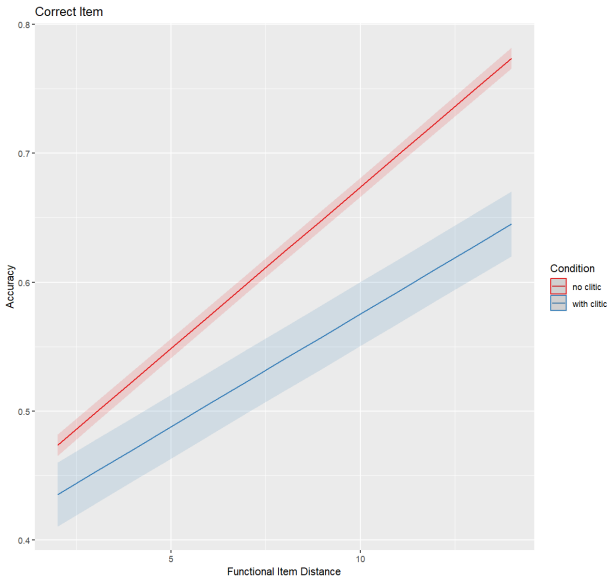

Figure 2: Acc. on grammatical triplets (correct item)

Contrast	Estimate	SE	z	p
GePpeTto vs. Mistral-7B	0.0710	0.00364	19.522	< .0001
GePpeTto vs. GPT2	0.0247	0.00364	6.794	< .0001
GePpeTto vs. Minerva-7B	0.4399	0.00364	120.809	< .0001
GePpeTto vs. TinyLlama-1.1B	0.1518	0.00364	41.728	< .0001
No clitic (Exp. 3 vs. 4)	0.1100	0.0136	8.117	< .0001
Distance effect (correct item)	0.0834	0.0129	6.441	< .0001
Distance effect (wrong item)	0.0919	0.0131	7.004	< .0001

Table 1: Summary of fixed effects from Exp. 3 and 4.

References Cinque, G. (1999). *Adverbs and functional heads*. Oxford University Press. * Cinque, G. (2005), Deriving Greenberg's Universal 20 and its exceptions, *Linguistic Inquiry* 36, 315-332 * Cinque, G. (2006). *Restructuring and functional heads*. Oxford University Press. * Wurmbrand, S. (2001). *Infinitives: Restructuring and clause structure*. De Gruyter Mouton. * Linzen, T., Dupoux, E., & Goldberg, Y. (2016). *Assessing the ability of LSTMs to learn syntax-sensitive dependencies*. Transactions of the Association for Computational Linguistics. * Hu, J., Gauthier, J., Qian, P., Wilcox, E. & Levy, R. (2020). *A systematic assessment of syntactic generalization in neural language models*. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. * Hale, J., & Stanojević, M. (2024). *Do LLMs learn a true syntactic universal?*. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. * Scontras, G., Degen, J, Goodman, N. D. (2017), Subjectivity predicts adjective ordering preferences, *Open Mind* (1), 53-66