Syntactic Maps or Surface Hacks?
Testing Restructuring Verb Order and Clitic Placement in LLMs

Tommaso Sgrizzi
NETs
IUSS
Pavia, Italy
tommaso.sgrizzi
@iusspavia.it

Asya Zanollo
NETs
IUSS
Pavia, Italy
asya.zanollo
@iusspavia.it

Cristiano Chesi
NETs
IUSS
Pavia, Italy
cristiano.chesi
@iusspavia.it

Abstract

This paper investigates how LLMs generalize the Universal Functional Hierarchy (UFH; Cinque, 1999), a concept proposed to arise from general, language-specific cognitive constraints (Cinque and Rizzi, 2012). Focusing on Italian restructuring verbs (Cinque, 2006), we test LLM sensitivity to verb ordering, clitic climbing, and auxiliary selection using minimal pairs. We also introduce pseudo-verbs to assess how models handle novel lexical items and the syntactic distinction between restructuring and control verbs (Landau, 2024; Wurmbrand, 2001). Three models were evaluated: GPT2-small, GePpeTto, and Minerva-7B-base-v1.0, with the latter two trained on Italian. Results indicate LLMs, particularly smaller Italian-trained models like GePpeTto, show some sensitivity to the UFH, but this is often heuristic, and doesn't translate to grammatical competence. On the other hand, models can in effect generalize restructuring and control syntax to novel verbs.

1 Introduction

Large language models (LLMs) have achieved remarkable success across a wide range of natural language understanding tasks, reigniting interest in their syntactic abilities and sparking a vigorous debate regarding the cognitive plausibility of the linguistic generalizations they acquire from data (Linzen et al., 2016, a.o.). Recent research has begun to probe the extent to which LLMs implicitly encode hierarchical syntactic structure (Goldberg, 2019; Hu et al., 2020; Wilcox et al., 2018), examining their sensitivity to phenomena such as long-distance dependencies and subject-verb agreement. This paper contributes to this growing body of work by investigating whether LLMs are sensitive to a crosslinguistically robust constraint governing the hierarchical distribution of functional verbs in Italian (Cinque, 2006; Grano, 2015). Given the broad cross-linguistic relevance of this phenomenon (Wurmbrand, 2001; Wurmbrand, 2015), our investigation directly addresses the question of the coherence of linguistic structural representations in LLMs. The research questions (RQs) that guide this study can be framed as following: **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? **RQ3:** What is the syntactic structure assigned by LLMs to novel verbs which introduce non-finite complements?

2 The empirical domain: Italian Restructuring verbs

The empirical domain under investigation are RVs in Italian. In this section we want to emphasize the relevance of the cartographic enterprise as valid heuristics to test the cognitive plausibility of syntactic generalizations. In formal linguistics, the cartographic approach refers to the effort to systematically map out the functional structure of the clause. Much like a geographical map reveals detailed topography, syntactic cartography seeks to uncover the fine-grained architecture of language, identifying a universal and richly articulated hierarchy of functional projections that determine the order of constituents in natural language (Cinque & Rizzi, 2012). Cartographic universals are not merely typological observations; they reflect deep structural constraints on human language, likely rooted in cognitive and interface-driven pressures such as learnability, interpretability, and communicative efficiency (see a.o. Biberauer, 2017; G. Ramchand and Svenonius, 2014; G. C. Ramchand, 2018). As such, they offer a highly structured benchmark for evaluating whether LLMs reflect the underlying principles of natural language cognition or simply reproduce surface-level statistical patterns. Assessing cartographic generalizations in LLMs thus becomes another valuable diagnostic tool for determining whether their internal representations exhibit the kind of compositional and hierarchical structure found in human language. A particularly revealing case study for testing structural representations from a cartographic perspective in LLMs comes from the domain of RVs in Italian (Cinque, 2006; Grano, 2015). RVs—such as *potere* 'can', *dovere* 'must', *volere* 'want', *continuare* 'continue', *cominciare* 'begin', are verbs that, despite selecting an infinitival complement, do not behave as if they embed a full clause (cf. Olivier et al., 2023; Wurmbrand, 2001; Wurmbrand, 2015, a.o.). Instead, they participate in a monoclausal structure, lacking the full complement of functional projections found in fully embedded (i.e., biclausal) contexts. This has observable syntactic consequences: only RVs permit movement of the object clitic from the complement position of the infinitive up to the matrix verb (e.g., *Marco lo vuole mangiare* 'Marco wants to eat it'), while control verbs, which are superficially similar, do not (e.g., **Marco lo decide di mangiare* 'Marco decides to eat it'). Clitic placement (Clitic Climbing; CC) thus offers a fruitful diagnostic for the underlying syntactic structure of a restructuring configuration. More specifically, the working hypothesis that we are adopting here (Cinque, 2006; Grano, 2015) views RVs as functional heads occupying a fixed hierarchy (e.g., aspectual > modal > mood), with each verb spelling out a specific functional projection (Fig. 1) rooted in the cartographic representation of the inflectional domain. Control verbs, on the other hand, are lexical verbs, projecting argument structure and part of a different syntactic category.

MoodP$_{\text{speech act}}$ > MoodP$_{\text{evaluative}}$ > MoodP$_{\text{evidential}}$ > MoodP$_{\text{epistemic}}$ > TP(Past) > TP(Future) > MoodP$_{\text{irrealis}}$ > ModP$_{\text{aletic}}$ > AspP$_{\text{habitual}}$ > AspP$_{\text{repetitive(I)}}$ > AspP$_{\text{frequentative(I)}}$ > ModP$_{\text{volitional}}$ AspP$_{\text{celerative(I)}}$ > TP(Anterior) > AspP$_{\text{terminative}}$ > AspP$_{\text{continuative}}$ > AspP$_{\text{retrospective}}$ AspP$_{\text{proximate}}$ > AspP$_{\text{durative}}$ > AspP$_{\text{generic/progressive}}$ > AspP$_{\text{prospective}}$ > ModP$_{\text{obligation}}$ ModP$_{\text{permission/ability}}$ > AspP$_{\text{completive}}$ > VoiceP > AspP$_{\text{celerative(II)}}$ > AspP$_{\text{repetitive(II)}}$ > AspP$_{\text{frequentative(II)}}$

Figure 1: Cinque, 2006:12

3 Syntactic Generalization in LLMs

Despite the impressive performance of state-of-the-art LLMs, it remains an open question whether their enhanced predictive capabilities reflect genuine syntactic knowledge. The issue of LLMs's grammatical knowledge is approached in the linguistic community through different approaches relying on controlled experimental settings, probing LLMs' performances on minimal pair sentences (Warstadt et al., 2020), and evaluating the internalization of deep hierarchical dependencies of the underlying linguistic structures in in specific benchmarks (Srivastava et al., 2022). In this context, the empirical domain of RVs provides an ideal testing ground for disentangling linear generalizations from structural rules. On the one hand, RVs follow specific linear orderings that could, in principle, be learned from surface patterns in the training data. On the other hand, their ordering can either permit or block some syntactic phenomena making linear order a surface reflex of deeper structural constraints. Capturing the relevant syntactic generalizations in this domain therefore requires more than sensitivity to word order—it demands an understanding of the underlying hierarchical structure.

4 Experiments

We designed 15 minimal-pair experiments targeting clitic placement, auxiliary selection, and verb-verb complementation. We manipulated restructuring presence, matrix verb type (restructuring, control verbs, pseudo-verbs), and verb distance in the hierarchy. We coded 15 restructuring verbs (reflecting Cinque's

2006 hierarchy, from *andare a* ‘to go’ [1] to *solere* ‘to be used to’ [15], i.e. from the structurally lowest to the structurally highest) and 15 *control* verbs (cf. Landau 2024), arbitrarily numbered: e.g., *andare a* [1], *correre a*, *cominciare a* [2],*salire a*).

We added three pseudo-verbs: *grabbare*, *drommare a*, and *trellare di*, to test whether LLMs treat them as restructuring or *control* verbs. The presence of the preposition for the latter two mirrors the behaviour of aspectual (restructuring verbs) such as *cominciare a* 'to begin' [2]. On the contrary, when control verbs select *di* or *a*, these are complementizers. Another surface similarity between control and restructuring verbs which reflects important structural differences.

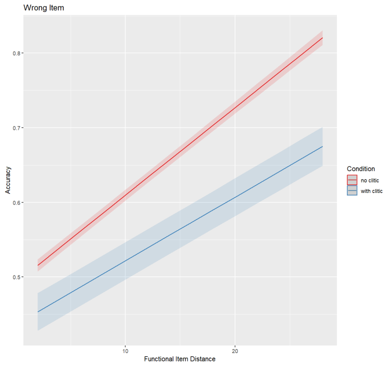
Exp. 1 contrasts hierarchy-respecting vs. violating restructuring verb pairs (combinations of 2 verbs); Exp. 2 adds proclitics. Exp. 3–4 extend these with 3-verb sequences, with (Exp. 4) or without (Exp. 3) clitics. Exp. 5–6 pair restructuring+*control* vs. *control*+restructuring verbs (both with clitics); only enclisis is grammatical, testing clitic domain blocking. Exp. 7–8 pair pseudo-verbs with restructuring verbs (pseudo+restructuring vs. restructuring+pseudo), probing proclitic preference as evidence for restructuring generalization. Exp. 9–10 do the same with *control* verbs; since proclisis is ungrammatical here, this tests whether models reject restructuring-like behavior for both control and pseudo verbs. Exp. 11 pairs pseudo-verbs with each other to investigate their behavior with proclisis, and enclisis; Exp. 12 tests their auxiliary selection (*have*/*be*) when selecting unaccusative verbs. Exp. 13–14 apply the same to restructuring and *control* verbs, respectively. Exp. 15 presents pseudo-verbs with proclitic/enclitic alternations, testing for restructuring effects modulated by prepositions *di*/*a*. For each condition internal to each experiment, we generated 100 lexical variants displaying different lexical items as subjects, infinitival verbs, and objects (when present).

4.1 LLMs Evaluation

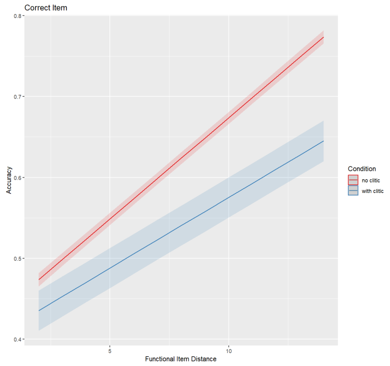
We employed one larger models ($\tilde{7}$ billion parameters) based on the Mistral architecture - Minerva-7B-base-v1.0; and two small models using a GPT2-style architecture- GPT2-small and GePpeTto. Among these, Minerva-7B-base-v1.0 and GePpeTto are models trained on Italian corpora, hence they allow us to assess whether exposure to Italian during training enhances syntactic generalization in a typologically relevant domain. This setup enables a direct comparison between architectures trained on different languages (specifically English and Italian), as well as between larger and smaller models, in their ability to internalize the structural dependencies necessary to abstract the relevant generalizations. All models are available on Hugging Face. The *LM-eval* platform (Gao et al., 2024) was adopted to perform minimal pair tests. The evaluation used the LM-eval platform, generating 610,500 minimal pairs. A generalized linear mixed model was fitted to analyze model performance.

5 Results and Discussion

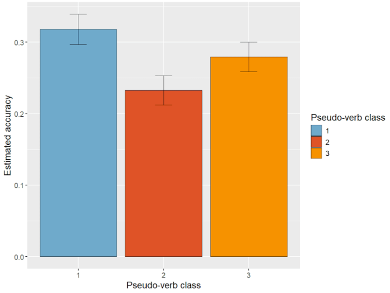
We present results from Exp. 3, 4, and 7. The former two indicate that LLMs, particularly GePpeTto, show sensitivity to the UFH proposed by Cinque. Model behavior aligned with hierarchical orderings, as sentence preference correlated with greater structural distance between verbs, regardless grammaticality (Fig. below). Failure in selecting the grammatical option increased especially when clitic climbing was involved. These results suggest that hierarchical awareness acts more as a heuristic than as a diagnostic of grammatical well-formedness. In Exp. 3 and 4, *GePpeTto* outperformed all other models tested. The estimated accuracy advantage of *GePpeTto* over *GPT2-small* it was $\hat{\beta} = 0.0247$ (SE = 0.00364, $z = 6.794$, $p < .0001$); and over *Minerva-7B-base-v1.0* it was $\hat{\beta} = 0.4399$ (SE = 0.00364, $z = 120.809$, $p < .0001$). 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. Results from Exp. 7 (see Fig. below) suggest that LLMs vary systematically in how readily they assign a restructuring analysis to novel pseudo-verbs, as reflected in their preferences for clitic placement (higher accuracy means a preference for the enclitic options). Across models, we observe that restructuring is most strongly disfavored with bare pseudo-verbs like *grabbare*, more likely with *a*-marking (*drommare a*), and disfavored again with *di*-marking (*trellare di*). Importantly, because pseudo-verbs do not appear in the training data and are not intrinsically marked as grammatical or ungrammatical, any preference for clitic placement must reflect the model's structural generalizations rather than memorization of specific lexical items. The study concludes that LLMs can learn structural aspects aligned with cartographic hierarchies, and extend them to novel contexts, but these are not necessarily translate to proper grammaticality judgments.



Summary of fixed effects Exp. 3–4



Eclitic selection in the 3 pseudo-verbs



Eclitic selection in the 3 pseudo-verbs

6 References

Biberauer, T. (2017). *Peripheral significance: A phasal perspective on the grammaticalisation of speaker perspective*. *Jung*, 93. * Cinque, G. (1999). *Adverbs and functional heads: A cross-linguistic perspective*. Oxford University Press. * Cinque, G. (2006). *Restructuring and functional heads*. Oxford University Press. * Cinque, G., & Rizzi, L. (2012). The cartography of syntactic structures. *CISCL Working Papers on Language and Cognition*, 2, 43–59. * Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., Le Noac'h, A., et al. (2024). *The language model evaluation harness*. https://doi.org/10.5281/zenodo.12608602 * Goldberg, Y. (2019). *Assessing BERT's syntactic abilities*. https://arxiv.org/abs/1901.05287 * Grano, T. (2015). *Control and restructuring*. Oxford University Press. * Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. (2020). A systematic assessment of syntactic generalization in neural language models. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1725–1744. * Landau, I. (2024). *Control (Elements)*. LingBuzz. https://doi.org/lingbuzz/008204 * 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. * Olivier, M., Sevdali, C., & Folli, R. (2023). Clitic climbing and restructuring in the history of French. *Glossa*, 8(1), 1–45. * Ramchand, G., & Svenonius, P. (2014). Deriving the functional hierarchy. *Language Sciences*, 46, 152–174. * Ramchand, G. C. (2018). *Situations and syntactic structures: Rethinking auxiliaries and order in English* (Vol. 77). MIT Press. * Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. (2022). *Beyond the imitation game: Quantifying and extrapolating the capabilities of language models*. *arXiv preprint arXiv:2206.04615*. * 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. * Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018). *What do RNN language models learn about filler-gap dependencies?* https://arxiv.org/abs/1809.00042 * Wurmbrand, S. (2001). *Infinitives*. De Gruyter Mouton. * Wurmbrand, S. (2015). *Restructuring cross-linguistically*. LingBuzz. https://doi.org/lingbuzz/002514