

ChatGPT and Cartography: prompting the Left Periphery of Italian and Italian varieties

Angelapia Massaro (University of Siena)

Giuseppe Samo (Beijing Language and Culture University/ Geneva)

angelapia.massaro@unisi.it, samo@blcu.edu.cn

The Left Periphery of the clause (since Rizzi 1997), is the portion of the syntactic architecture hosting items bearing scope-discourse properties, such as topics, as given in *Il tuo libro, lo ho comprato* ‘Your book, I bought it’ (Italian, Rizzi 1997: 289, 15a).

The patterns represent a marked word order, reorderings of canonical clauses, realized via A'-movement of the dislocated element leaving a gap within the clause (cf. filler-gap). Their parsing creates asymmetries across populations of speakers (developmental grammars and language pathology; Friedmann et al. 2009 and related works). In a similar vein, filler-gap dependencies do not represent hard challenges for large language models (LLMs, see also Linzen & Baroni 2021 for an overview) trained on a large set of (tendentially) non-annotated, raw data. In a recent paper published on Linguistic Inquiry, Wilcox et al. (2022) proposes that these LLMs might provide evidence against the Argument of the poverty of the stimulus (see also Katzir 2023). If ChatGPT (OpenAI 2023; see an overview and critical discussions in Bang et al. 2023), is prompted with a question (see also ChatGPT as informant, Haider 2023) about an activated Left Periphery of Italian, the provided answer sounds somewhat “creative”.

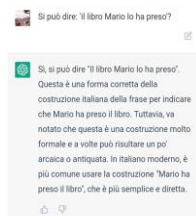


Figure 1: User: can you say (lit.) ‘The book, Mario it has taken’? Reply: Yes, you can say (lit.) “The book, Mario it has taken”. This is a correct form of Italian construction to indicate that Mario took the book. However, it should be noted that this is a **very formal construction and it may result archaic or antiquated**. In Modern Italian, it is more common to adopt the construction “Mario took the book”, which is simpler and direct.

In our presentation, we aim to evaluate such behavior by raising and investigating two research questions (1) bias effects and (2) what research questions can be raised exploring LLMs. We first run a study on bias, following assumptions from Quantitative Computational Syntax (Merlo 2016 and related works). We queried four treebanks annotated under the guidelines of Universal Dependencies (UD, Zeman et al. 2023; ISDT, VIT, TWITTIRO, PoSTWITA) and fronted arguments in the LP are not rare (4% all arguments, 23% of XP arguments; Binomial test: $p < .000001$; $z\text{-p} = z = -88.508914$). What emerges is that, quantitatively speaking, reorderings in Italian are not una forma arcaica o antiquata ‘archaic or antiquated form’. The second part of the presentation is devoted in prompting additional examples such as clitic resumptions, subject topicalization in Southern Italian varieties, from those who are more represented in training data (e.g. Neapolitan, Basile 2022) to smaller varieties (e.g. the Apulian dialect of San Marco in Lamis). We believe that we are far away from exploring LLM for grammaticality (cf. Haider 2023 on English), at least, and even, for Standard Italian.

Keywords: Cartography; Italian; LLMs; ChatGPT, Left Periphery.