Not lost in the garden path: large language models navigate their way to syntactic knowledge

Zeping Liu, Chien-Jer Charles Lin, Nazbanou Nozari (Indiana University Bloomington)

Introduction: Large language models (LLMs) have shown remarkable ability in processing complex linguistic structures [1], prompting the question of whether they possess human-like syntactic knowledge. Unlike humans, LLMs are trained on massive text datasets without explicit syntactic rules, raising doubts about whether their strong performance reflects a true understanding of syntactic structures akin to that of humans or merely the modeling of statistical patterns in word sequences. To address this question, this study focuses on garden-path sentences, which create temporary syntactic ambiguities that are resolved eventually in favor of the initially dispreferred parse. Previous studies suggest that human readers and LLMs show similar patterns of processing difficulty when interpreting these sentences. However, LLMs struggle to fully capture the magnitude of difficulty observed in humans across different garden-path constructions [2-3]. In light of this, this study investigates whether LLMs' sensitivity to syntax can effectively capture human parsing behaviors without presupposed syntactic knowledge, focusing on garden-path disambiguation in English and in Mandarin to examine their cross-linguistic potential in syntactic processing.

Methods: We examined 3 garden-path types (MVRR, NPS, NPZ) in English (**Table 1)**. Human word-by-word reading times (RTs) for these sentences (24 pairs per type with lexical items controlled, each with both ambiguous and unambiguous versions) were collected from 2,000 participants [3]. We used two transformer-based models, GPT2-large [4] and Llama-3.2-1B, to calculate surprisal values (the negative log probability of a word given prior context [5]) for each word in the sentences. In Mandarin, we focused on the *V+N1+DE+N2* fragment, which features a structural reanalysis from an initially preferred relative clause (RC) interpretation to a complement clause (CC) interpretation [6-7]. RTs for 13 sentence pairs were collected from 40 Mandarin speakers, with surprisal values derived from gpt2_wiki40b_zh-cn [8] and Chinese-llama-2-1.3b. Mandarin sentences were manually segmented into words to match the RT data.

Analysis: Garden-path effects in human reading data typically spread over multiple words, with RTs peaking at the word immediately following the point of disambiguation (e.g., *several* in *needed several more*). In contrast, LLMs respond to syntactic disambiguating information immediately at the first disambiguating word (e.g., *needed*), without the delayed reaction observed in humans (**Fig. 1**). Previous studies addressed this discrepancy by fitting a linear relationship between RTs and surprisal on filler items to predict target RTs, which were compared to empirical data [3-4]. This approach is indirect and less intuitive because the linear relationship between RTs and surprisal may not be the same for syntactically complex and simpler sentences. Given this, we used two pre-critical regions (e.g., *the truck*) as a baseline to estimate syntactic disambiguation relative to a consistent pre-disambiguation context. Our analysis focuses on the difference between critical regions (e.g., *needed several*) and pre-critical regions to compare human and LLMs performance on syntactic disambiguation in English and Mandarin.

Results: Model prediction shows sensitivity to syntactic disambiguation difficulty in the English dataset (greater cost for ambiguous than unambiguous items). It also captures the relative difficulty across different types, showing that both MVRR and NPZ are more challenging than NPS in ambiguity resolution (the difference between MVRR and NPZ may be due to noise, **Fig. 2**). This provides evidence that pure exposure to linguistic data without a priori syntactic framework, as seen in LLMs, can lead to the implicit learning of syntactic knowledge. In Mandarin, both models predict a garden-path effect in the opposite direction from the empirical one. One potential source of the models' failure is the misalignment of sentence segmentation between human RT data and model results (even with manual segmentation). Given this, we analyzed a subset of sentences where model segmentation matches the RT data. However, LLMs still did not capture the RC/CC disambiguation, leaving open questions regarding the model's ability to generalize across languages and its limitations in processing syntactic ambiguity in Mandarin.

Table 1. Garden-path sentences in English and Mandarin.

		baseline	critical
Matrix Verb/Reduced Relative Clause (MVRR)			
Ambiguous:	The mechanic brought	the truck	**needed several** more hours to repair it.
Unambiguous:	The mechanic who was brought	the truck	**needed several** more hours to repair it.
Noun Phrase/Sentential Complement (NPS)			
Ambiguous:	The mechanic observed	the truck	**needed several** more hours to be repaired.
Unambiguous:	The mechanic observed that	the truck	**needed several** more hours to be repaired.
Noun Phrase/Main Clause Subject (NPZ)			
Ambiguous:	Because the mechanic stopped	the truck	**needed several** more hours before …
Unambiguous:	Because the mechanic stopped,	the truck	**needed several** more hours before …

Relative Clause/Complement Clause (RC/CC)

Ambiguous: 虐待 / 孩子 / 的 / 保姆 / **之后，** / **那对** / 夫妻 …
mistreat / child / DE / nanny / **after,** / **that** / couple …
"After mistreating the child's nanny, that couple…"
Garden-path (RC analysis): the nanny mistreated the child …

Unambiguous: 虐待 / 恶毒 / 的 / 保姆 / **之后，** / **那对** / 夫妻 …
mistreat / evil / DE / nanny / **after,** / **that** / couple …
"After mistreating the evil nanny, that couple…"

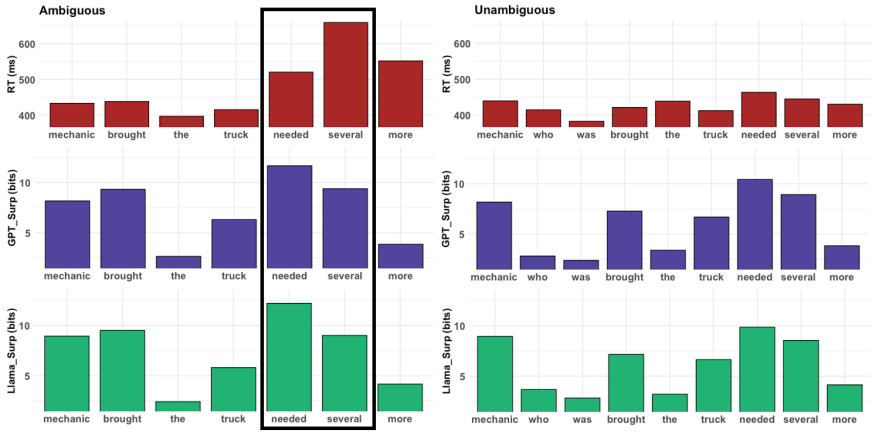

Fig. 1 RTs (red) and surprisal (purple: GPT; green: Llama) distributed across words in the MVRR ambiguity. The circled pattern is also observed in the NPZ and NPS ambiguities. Bars represent mean values across all items, with an annotated example sentence for illustration.

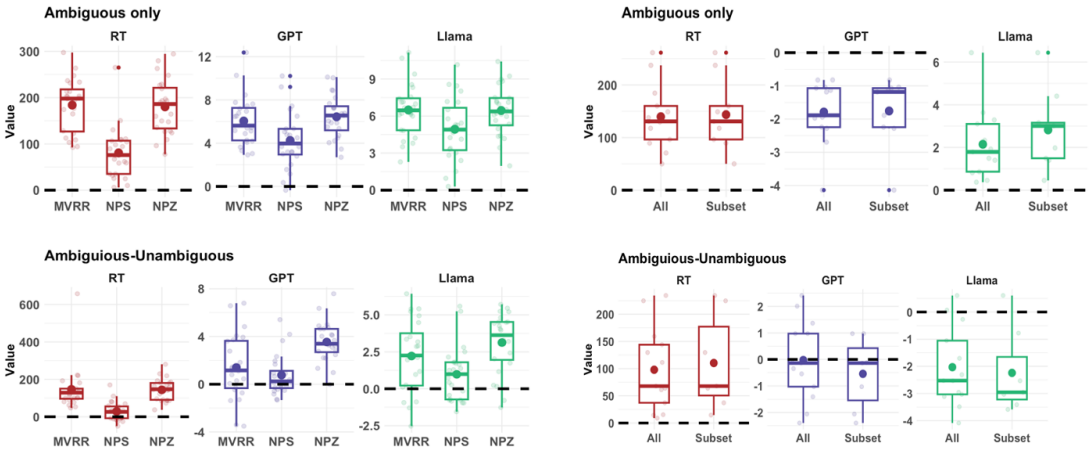

Fig 2. Human RTs and LLMs surprisal for three garden-path constructions in English (left) and Mandarin (right). The Y-axis is the difference between averaged critical regions and averaged baseline regions. In Mandarin results, the subset focuses on sentence items where model segmentation matches the RT data.

References: [1] Futrell et al., 2019. In *Proceedings of NAACL*. [2] van Schijndel & Linzen, 2021. *Cognitive Science.* [3] Huang et al., 2024. *Journal of Memory and Language.* [4] Hale, 2001. In *Proceedings of NAACL*. [5] Radford et al., 2018. OpenAI blog. [6] Ng & Wicha, 2014. *Journal of Memory and Language.* [7] Hsieh et al., 2015. *Journal of Psycholinguistic Research*. [8] Xu et al., 2023. In *EMNLP*.