

**Statistical learning (partially) accounts for what you choose, not what you predict:
Evidence from cataphora resolution in Chinese**

Introduction: This study aims to probe to what extent language processing can be reduced to probabilistic learning. To this end, we tested the role of gender agreement and a structural constraint called **phase c-command** [1-2] during cataphora resolution in Chinese, comparing human data with predictions from the large language model (LLM) BERT with whole-word masking. Our results suggest that while BERT can simulate forced-choice judgments to some degree, it fails to capture the early structural effect observed in incremental human sentence processing [e.g., 3-4].

Methods: We crossed the factors **Structure** (Licensed vs. Unlicensed) and **Gender** (Match vs. Mismatch) in a 2 x 2 factorial design. In the Unlicensed conditions, the phase c-command constraint prohibits coreferential readings, in contrast to the Licensed conditions. Gender refers to the gender agreement between the pronoun and the postcedent. See (1) for an example. We conducted a self-paced reading (SPR) experiment (24 targets, 36 fillers, 113 participants) where each trial was followed by forced choice judgment. (We also ran an eye-tracking study with 71 participants, not reported here in detail due to space, but the results mirror those of the SPR experiment.) For online processing, we used the **gender mismatch effect (GMME)** as a diagnostic for attempts at coreferential interpretation [3-4]. To examine whether probabilistic information extracted from text can simulate human processing patterns, we derived from BERT (i) the log probability of the postcedent as a proxy for forced-choice judgments, and (ii) the surprisal of the postcedent as a proxy for reading times (RTs).

Results: **Fig.1** shows the judgment and RT patterns from the 113 participants. For the forced-choice data, we found main effects of Structure and Gender (p 's < 0.001) but no interaction, indicating that (i) the phase-based c-command constraint modulates cataphora resolution, and (ii) the structural effect is not absolute. Critically, the RT data at the postcedent region showed a main effect of Structure and a Structure $\times$ Gender interaction (p 's < 0.05). Pairwise comparisons revealed that the GMME appeared only within the Licensed conditions. This means that native speakers excluded the structurally inaccessible referent during early processing, despite gender-intrusion effects in global interpretations. The eye-tracking data showed a similar RT pattern. Interestingly, BERT partially captured the gender and structural effects in offline judgments (**Fig.2**), although the effect sizes (on a z-scale) were smaller. Critically, in contrast to the human RT results, the BERT-derived surprisal (a proxy for RTs) showed only a trending GMME, failing to fully replicate the interaction pattern. Indeed, our statistical model with surprisal as a covariable suggests that this metric alone cannot account for the observed patterns.

Conclusion: The comparison between psycholinguistic experiments and computational simulations that do not incorporate an explicit structural constraint suggests that real-time cataphora resolution cannot be fully reduced to statistical learning from text input. Furthermore, the RT results are compatible with a **defeasible filter** view of syntactic processing, where syntactic constraints “filter out” structurally inaccessible referents initially, only to be overridden later by semantic/pragmatic considerations.

(1) **An example target item is shown below.** Chinese text is transliterated verbatim to English. RC is short for relative clause. In subject relative clauses (SRCs), coreference does not violate the phase c-command constraint; in object relatives (ORCs), coreference is degraded.

a. **Licensed:** that [_{SRC} **t_i** by **his/her** teacher scolded Comp] **boy_i** was reviewing homework.

b. **Unlicensed:** that [_{ORC} **his/her** teacher scolded **t_i** Comp] **boy_i** was reviewing homework.

Translation: ‘The boy who (was scolded by)_{SRC} his/her teacher (had scolded)_{ORC} was...’

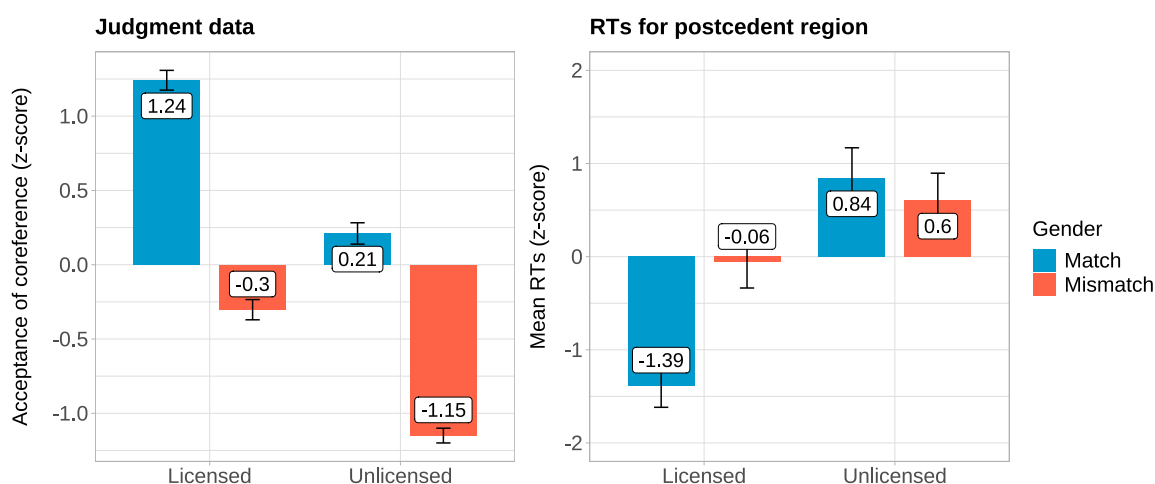


Fig.1: Forced-choice judgments and RTs at the postcedent region in the SPR experiment.

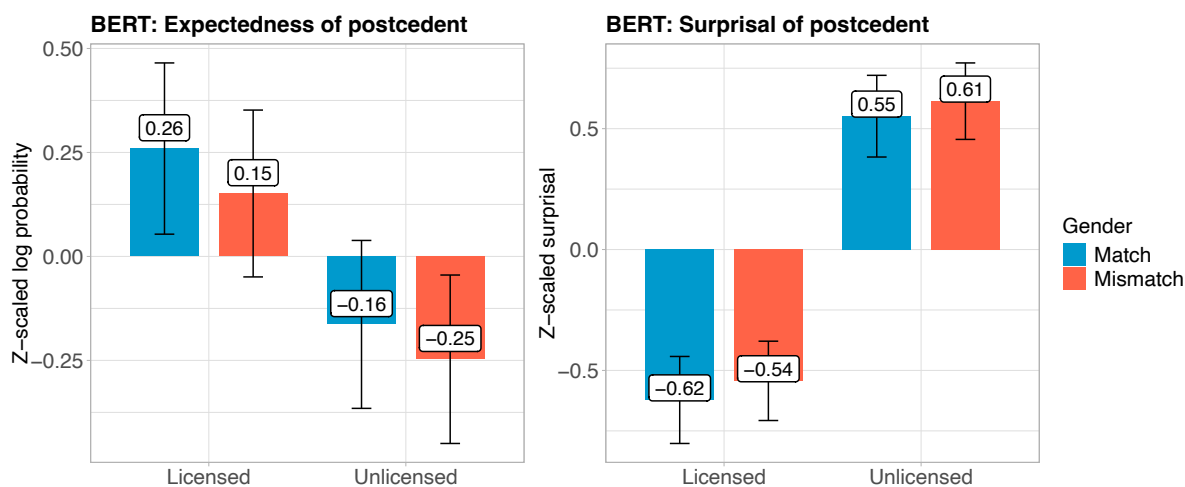


Fig.2: Simulated log probabilities (left) and surprisal by condition based on 24 target items.

References [1] Huang, James, C.-T. 1982. Logical Relations in Chinese and the Theory of Grammar. Ph.D. dissertation. MIT. [2] Huang, James C.-T. & J.-W. Lin. 2021. Quantificational binding without surface c-command in Mandarin Chinese. *Current Issues in Syntactic Cartography: A Crosslinguistic Perspective*. [3] Kazanina, N. et al. 2007. The effect of syntactic constraints on the processing of backward anaphora. *JML*. [4] Kush, D. & B. Dillon. 2021. Principle B constrains the processing of cataphora: Evidence for syntactic and discourse predictions. *JML*.