Do Language Models discriminate between relatives and pseudorelatives?

Syntactic background

- Pseudorelatives (**PR**s, [1, 2, 3, 4, 11] a.o.), are attested in Romance, Greek, Dutch, Serbo-Croatian, and Inuktitut [5, 12, 9, 7, 8, 17]. **We focus on French here**.

(1) Je vois Marie [**qui danse**].
I see Marie who dances.
'I see Marie dancing.'

- PRs resemble relatives clauses (RCs), but:
  1. their head noun can be **cliticized**;
  2. they only allow **subject-gap** dependencies;
  3. they mostly involve **perception verbs**;
  4. they require the matrix and embedded **tenses to match**.

(2) Cliticization (PR-parse only):
Je **la** vois [qui danse].
I **her.CL** see who dances.
'I see her dancing.'

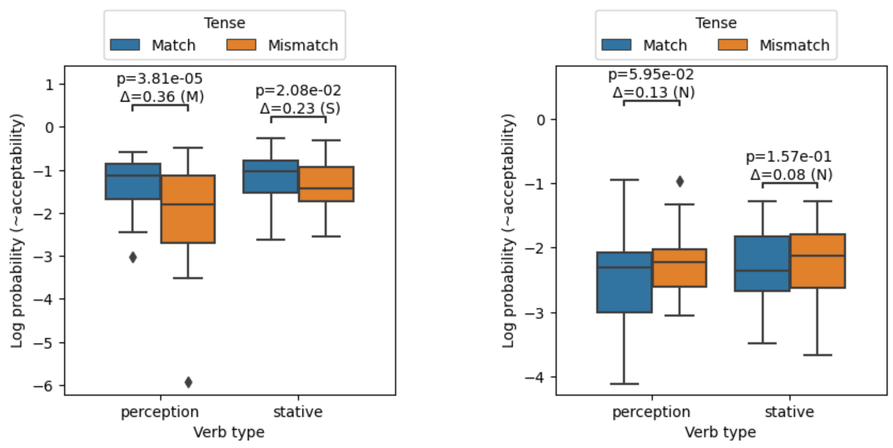
(3) Cliticization + no perception verb:
* Je la **pense** [qui danse].
I her.CL **think** who dances.

(4) Cliticization + tense mismatch:
* Je la voy**ais** [qui danse].
I her.CL see.**PST** who dances.

(5) Cliticization + object gap:
* Je la vois [que Jean appelle __].
I her.CL see that Jean calls __.

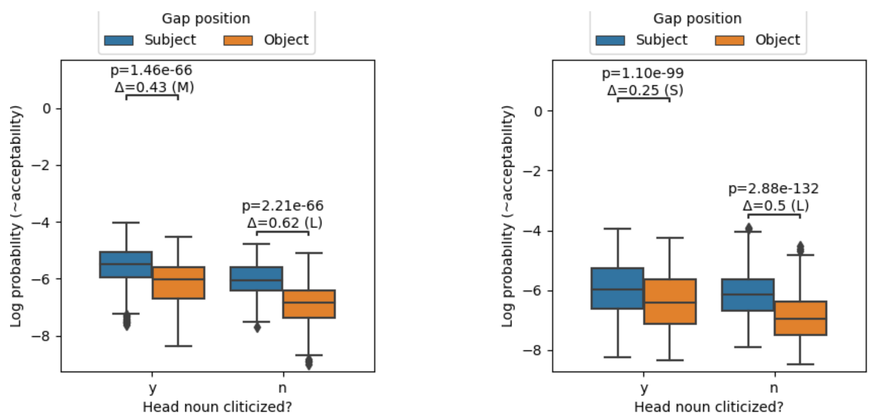
Motivation & previous work

- PRs are easily confusable with RCs and recent Large Language Models (LLMs) are not directly trained to differentiate them. **Do LLMs learn the specificities of the PR anyway?**
- Previous work investigated the capacity of RNNs to learn phenomena such as filler-gap dependencies and island effects [13, 18], various garden-path effects [14], and relative clauses [16]. **This is the investigation of the pseudorelative through that lens.**



(a) French **(b)** English

Figure 1: Grammaticality scores for Exp. 1 obtained with `xlm-roberta-large`.



(a) Perception verbs **(b)** Attitude/action verbs

Figure 2: Grammaticality scores for Exp. 2 obtained with `gpt2-base-french`.

Experiment 1: verb type, tense anaphoricity

Design

- Experiment 1 tests Properties 3 & 4 by replicating the result of [15] with 8 French LLM "subjects".
- 18 frames of the form S V O [Wh V' O'] where V is ±perception, and the tenses of V and V' are ±matching were fed to the LLMs.
- Our proxy for grammaticality was the **log-probability assigned to a given sentence** by the LLM [6, 10]. Effects were assessed with LMER (same in Experiment 2).

Predictions & Results

- We expect a main effect of verb type and tense anaphoricity, plus an interaction.
- 6/8 LLMs favored matching tenses, and 4/8 more so under perception verbs. No such effects with (control) English LLMs tested on comparable stimuli (expected!).
- **Limitation: the result could be incorrectly driven by an RC-parse...**

Experiment 2: cliticization, gap, verb type

Design

- We test Properties 1, 2 & 3 by feeding the same 8 French LLMs with 4800 sentences following the (glossed) pattern below. Same scores and models as before.

{He / She} {him.CL / her.CL / ∅} {sees/... / thinks/... / greets/...} {∅ / Marie / Jean} {subject-gap relative / object-gap relative}

Predictions & Results

- We expect a positive association between cliticization, matrix perception verbs, and subject gaps.
- Robust preference for subject-gaps (8/8 models) and more so under perception verbs (5/8 models)... but the desired 3-way interaction was only captured by 1/8 models.
- **Additionally, the interaction between cliticization and subject-gaps is predicted by 7/8 models to have a *negative* effect on grammaticality scores (‼)**

Conclusion & outlook

- The experiments we run show that LLMs capture certain properties of PRs, pertaining to acceptable filler-gap dependencies, matrix verbs, and tense combinations.
- Yet, the property that is perhaps the most specific to pseudorelatives, **cliticization**, does not seem to influence sentence probability scores in Experiment 2, nor specific semantic inference patterns in a third Experiment omitted for reasons of space (but ask me about it!).
- **This still raises the question whether LLMs really get the specificity of the PR as a *syntactic construction* (Experiment 2) with a specific *semantics* (Experiment 3)**; or whether they simply recycle general processing heuristics applicable to other structures (e.g. standard RCs)...
- By extension, this might constitute (weak) evidence in favor of an innate ±high-attachment parameter in humans, controlling the acquisition and mastery of PRs.

Adèle Hénot-Mortier
mortier@mit.edu
Massachusetts Institute of Technology

Linguistics