Do Language Models discriminate between relatives and pseudorelatives?

Adèle Hénot-Mortier (MIT)

August 18, 2023

4th International Conference on Theoretical East Asian Psycholinguistics (ICTEAP-4)

Introduction

Background: the pseudorelative across languages

- Pseudorelatives (**PR**s), are constructions attested in Romance (except Romanian), Greek, Dutch, Serbo-Croatian, and Inuktitut.[1]

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

(2) Italian:
Ho visto Gianni [**che correva**].
Have seen Gianni that run.IMPF
'I saw Gianni running.'

(3) Inuktitut:
pingasu-nik anguti-nik [**tikit-tu**] -qaq-tuq
three.PL.MOD man.PL.MOD arrive.PTCP have.PTCP.3S.S
'Three men arrived.'

- **In this presentation, we will focus on French data.**

[1]Previous selected works of the PR include Schwarze (1974), Kayne (1975), Radford (1975), Graffi (1980), Guasti (1988), Rizzi (1992), and Casalicchio (2013). For the cross linguistic observations, see Rafel (1999), Desmet et al. (2002), Lovrić (2003), Papadopoulou and Clahsen (2003), Grillo and Costa (2014), and Yuan (2022) a.o.

Background: the pseudorelative across languages

- Pseudorelatives (**PR**s), are constructions attested in Romance (except Romanian), Greek, Dutch, Serbo-Croatian, and Inuktitut.[1]

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

(2) Italian:
Ho visto Gianni [**che correva**].
Have seen Gianni that run.IMPF
'I saw Gianni running.'

(3) Inuktitut:
pingasu-nik anguti-nik [**tikit-tu**] -qaq-tuq
three.PL.MOD man.PL.MOD arrive.PTCP have.PTCP.3S.S
'Three men arrived.'

- **In this presentation, we will focus on French data.**

[1]Previous selected works of the PR include Schwarze (1974), Kayne (1975), Radford (1975), Graffi (1980), Guasti (1988), Rizzi (1992), and Casalicchio (2013). For the cross linguistic observations, see Rafel (1999), Desmet et al. (2002), Lovrić (2003), Papadopoulou and Clahsen (2003), Grillo and Costa (2014), and Yuan (2022) a.o.

Background: the pseudorelative across languages

- Pseudorelatives (**PR**s), are constructions attested in Romance (except Romanian), Greek, Dutch, Serbo-Croatian, and Inuktitut.[1]

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

(2) Italian:
Ho visto Gianni [**che correva**].
Have seen Gianni that run.IMPF
'I saw Gianni running.'

(3) Inuktitut:
pingasu-nik anguti-nik [**tikit-tu**] -qaq-tuq
three.PL.MOD man.PL.MOD arrive.PTCP have.PTCP.3S.S
'Three men arrived.'

- **In this presentation, we will focus on French data.**

[1] Previous selected works of the PR include Schwarze (1974), Kayne (1975), Radford (1975), Graffi (1980), Guasti (1988), Rizzi (1992), and Casalicchio (2013). For the cross linguistic observations, see Rafel (1999), Desmet et al. (2002), Lovrić (2003), Papadopoulou and Clahsen (2003), Grillo and Costa (2014), and Yuan (2022) a.o.

Why "pseudo"-relative?

- PRs resemble relatives clauses (RCs), but (at least in Romance), exhibit a few additional characteristics:
    1. their head noun can be **cliticized**, i.e. raised to a higher position;
    2. they only allow **subject-gap** dependencies;
    3. they mostly involve **perception verbs**;[2]
    4. they require the **matrix and embedded tenses to match**.
- Those properties cluster together to define PRs but it is worth noting that **only Property 1 (cliticization) is truly specific to PRs**, i.e., disallows a RC-parse. Properties 2-4 *can* be true of both PRs and RCs.

[2]Verbs like *find*, *catch* or *meet*, as well as existential/presentative contexts (*There is X who...*/*X is here who...*) also appear compatible with the PR.

Why "pseudo"-relative?

- PRs resemble relatives clauses (RCs), but (at least in Romance), exhibit a few additional characteristics:
  1. their head noun can be **cliticized**, i.e. raised to a higher position;
  2. they only allow **subject-gap** dependencies;
  3. they mostly involve **perception verbs**;[2]
  4. they require the **matrix and embedded tenses to match**.
- Those properties cluster together to define PRs but it is worth noting that **only Property 1 (cliticization) is truly specific to PRs**, i.e., disallows a RC-parse. Properties 2-4 *can* be true of both PRs and RCs.

[2] Verbs like *find*, *catch* or *meet*, as well as existential/presentative contexts (*There is X who...*/*X is here who...*) also appear compatible with the PR.

Why "pseudo"-relative?

- PRs resemble relatives clauses (RCs), but (at least in Romance), exhibit a few additional characteristics:
    1. their head noun can be **cliticized**, i.e. raised to a higher position;
    2. they only allow **subject-gap** dependencies;
    3. they mostly involve **perception verbs**;[2]
    4. they require the **matrix and embedded tenses to match**.
- Those properties cluster together to define PRs but it is worth noting that **only Property 1 (cliticization) is truly specific to PRs**, i.e., disallows a RC-parse. Properties 2-4 *can* be true of both PRs and RCs.

[2]Verbs like *find*, *catch* or *meet*, as well as existential/presentative contexts (*There is X who...*/*X is here who...*) also appear compatible with the PR.

Why "pseudo"-relative?

- PRs resemble relatives clauses (RCs), but (at least in Romance), exhibit a few additional characteristics:
    1. their head noun can be **cliticized**, i.e. raised to a higher position;
    2. they only allow **subject-gap** dependencies;
    3. they mostly involve **perception verbs**;[2]
    4. they require the **matrix and embedded tenses to match**.
- Those properties cluster together to define PRs but it is worth noting that **only Property 1 (cliticization) is truly specific to PRs**, i.e., disallows a RC-parse. Properties 2-4 *can* be true of both PRs and RCs.

[2]Verbs like *find*, *catch* or *meet*, as well as existential/presentative contexts (*There is X who...*/*X is here who...*) also appear compatible with the PR.

Why "pseudo"-relative?

- PRs resemble relatives clauses (RCs), but (at least in Romance), exhibit a few additional characteristics:
    1. their head noun can be **cliticized**, i.e. raised to a higher position;
    2. they only allow **subject-gap** dependencies;
    3. they mostly involve **perception verbs**;[2]
    4. they require the **matrix and embedded tenses to match**.
- Those properties cluster together to define PRs but it is worth noting that **only Property 1 (cliticization) is truly specific to PRs**, i.e., disallows a RC-parse. Properties 2-4 *can* be true of both PRs and RCs.

[2] Verbs like *find*, *catch* or *meet*, as well as existential/presentative contexts (*There is X who...*/*X is here who...*) also appear compatible with the PR.

Why "pseudo"-relative?

- PRs resemble relatives clauses (RCs), but (at least in Romance), exhibit a few additional characteristics:
    1. their head noun can be **cliticized**, i.e. raised to a higher position;
    2. they only allow **subject-gap** dependencies;
    3. they mostly involve **perception verbs**;[2]
    4. they require the **matrix and embedded tenses to match**.
- Those properties cluster together to define PRs but it is worth noting that **only Property 1 (cliticization) is truly specific to PRs**, i.e., disallows a RC-parse. Properties 2-4 *can* be true of both PRs and RCs.

[2]Verbs like *find*, *catch* or *meet*, as well as existential/presentative contexts (*There is X who...*/*X is here who...*) also appear compatible with the PR.

Illustrating Properties 1-4

(4) Baseline (RC/PR):

Je vois Marie [qui danse].
I see Marie who dances.
'I see Marie dancing.'
'I see Marie, who is dancing.'

(6) Cliticization + no perception verb:

* Je la pense [qui danse].
I her.CL think who dances.
Intended: 'I think she is dancing.'

(5) Cliticization (PR only):

Je la vois [qui danse].
I her.CL see who dances.
'I see her dancing.'

(7) Cliticization + tense mismatch:

* Je la voyais [qui danse].
I her.CL saw who dances.
Intended: 'I saw her dancing.'

(8) Cliticization + object gap:

* Je la vois [que Jean appelle].
I her.CL see that Jean calls.
Intended: 'I see Jean calling her.'

Illustrating Properties 1-4

(4) Baseline (RC/PR):

Je vois Marie [qui danse].
I see Marie who dances.
'I see Marie dancing.'
'I see Marie, who is dancing.'

(6) Cliticization + no perception verb:

* Je la pense [qui danse].
I her.CL think who dances.
Intended: 'I think she is dancing.'

(5) Cliticization (PR only):

Je la vois [qui danse].
I her.CL see who dances.
'I see her dancing.'

(7) Cliticization + tense mismatch:

* Je la voyais [qui danse].
I her.CL saw who dances.
Intended: 'I saw her dancing.'

(8) Cliticization + object gap:

* Je la vois [que Jean appelle].
I her.CL see that Jean calls.
Intended: 'I see Jean calling her.'

Illustrating Properties 1-4

(4) Baseline (RC/PR):

Je vois Marie [qui danse].
I see Marie who dances.
'I see Marie dancing.'
'I see Marie, who is dancing.'

(6) Cliticization + no perception verb:

* Je la pense [qui danse].
I her.CL think who dances.
Intended: 'I think she is dancing.'

(5) Cliticization (PR only):

Je la vois [qui danse].
I her.CL see who dances.
'I see her dancing.'

(7) Cliticization + tense mismatch:

* Je la voyais [qui danse].
I her.CL saw who dances.
Intended: 'I saw her dancing.'

(8) Cliticization + object gap:

* Je la vois [que Jean appelle].
I her.CL see that Jean calls.
Intended: 'I see Jean calling her.'

Illustrating Properties 1-4

(4) Baseline (RC/PR):

Je vois Marie [qui danse].
I see Marie who dances.
'I see Marie dancing.'
'I see Marie, who is dancing.'

(5) Cliticization (PR only):

Je la vois [qui danse].
I her.CL see who dances.
'I see her dancing.'

(6) Cliticization + no perception verb:

* Je la pense [qui danse].
I her.CL think who dances.
Intended: 'I think she is dancing.'

(7) Cliticization + tense mismatch:

* Je la voyais [qui danse].
I her.CL saw who dances.
Intended: 'I saw her dancing.'

(8) Cliticization + object gap:

* Je la vois [que Jean appelle].
I her.CL see that Jean calls.
Intended: 'I see Jean calling her.'

Illustrating Properties 1-4

(4) Baseline (RC/PR):

Je vois Marie [qui danse].
I see Marie who dances.
'I see Marie dancing.'
'I see Marie, who is dancing.'

(6) Cliticization + no perception verb:

* Je la pense [qui danse].
I her.CL think who dances.
Intended: 'I think she is dancing.'

(5) Cliticization (PR only):

Je la vois [qui danse].
I her.CL see who dances.
'I see her dancing.'

(7) Cliticization + tense mismatch:

* Je la voyais [qui danse].
I her.CL saw who dances.
Intended: 'I saw her dancing.'

(8) Cliticization + object gap:

* Je la vois [que Jean appelle].
I her.CL see that Jean calls.
Intended: 'I see Jean calling her.'

Research question

- PRs are easily confusable with RCs and recent Large Language Models (LLMs) are not really trained to differentiate them.
- Unambiguous PRs are also comparatively rare in corpora.
- **Do LLMs learn the specificities of the PR anyway?**
- Previous work investigated the capacity of recurrent language models to learn related but less targeted phenomena, such as filler-gap dependencies Wilcox et al., 2018 and relative clauses "LSTMs Can Learn Basic Wh- and Relative Clause Dependencies in Norwegian", 2022.

Research question

- PRs are easily confusable with RCs and recent Large Language Models (LLMs) are not really trained to differentiate them.
- Unambiguous PRs are also comparatively rare in corpora.
- **Do LLMs learn the specificities of the PR anyway?**
- Previous work investigated the capacity of recurrent language models to learn related but less targeted phenomena, such as filler-gap dependencies Wilcox et al., 2018 and relative clauses "LSTMs Can Learn Basic Wh- and Relative Clause Dependencies in Norwegian", 2022.

Research question

- PRs are easily confusable with RCs and recent Large Language Models (LLMs) are not really trained to differentiate them.
- Unambiguous PRs are also comparatively rare in corpora.
- **Do LLMs learn the specificities of the PR anyway?**
- Previous work investigated the capacity of recurrent language models to learn related but less targeted phenomena, such as filler-gap dependencies Wilcox et al., 2018 and relative clauses "LSTMs Can Learn Basic Wh- and Relative Clause Dependencies in Norwegian", 2022.

Research question

- PRs are easily confusable with RCs and recent Large Language Models (LLMs) are not really trained to differentiate them.
- Unambiguous PRs are also comparatively rare in corpora.
- **Do LLMs learn the specificities of the PR anyway?**
- Previous work investigated the capacity of recurrent language models to learn related but less targeted phenomena, such as filler-gap dependencies Wilcox et al., 2018 and relative clauses "LSTMs Can Learn Basic Wh- and Relative Clause Dependencies in Norwegian", 2022.

Experiment 1: verbs of perception & tense anaphoricity

Experiment 1: general goal

- Recall that pseudorelative construction in French requires the embedding verb to be a **verb of perception** (Property 3), and the **matrix and embedded tenses to match** (Property 4).
- Experiment 1 is supposed to test the correlation between those two properties, by replicating the result of a recent psycholinguistic study (Pozniak et al., 2019), but this time with 8 French LLM "subjects".

Model	Lang.	Architecture	Reference
flaubert_base_uncased	fr	Bidirectional	Le et al. (2020)
camembert-base	fr	Bidirectional	Martin et al. (2020)
gpt2-base-french	fr	Autoregressive	ClassCat AI Lab (2022)
gpt2-wechsel-french	fr	Autoregressive	Minixhofer et al. (2022)
bert-base-multi-lingual-cased	multi	Bidirectional	Devlin et al. (2018)
xlm-roberta-base	multi	Bidirectional	Conneau et al. (2019)
xlm-roberta-large	multi	Bidirectional	Conneau et al. (2019)
xlm-mlm-17-1280	multi	Bidirectional	Lample and Conneau (2019)

Table 1: French models used in Experiment 1

Experiment 1: general goal

- Recall that pseudorelative construction in French requires the embedding verb to be a **verb of perception** (Property 3), and the **matrix and embedded tenses to match** (Property 4).
- Experiment 1 is supposed to test the correlation between those two properties, by replicating the result of a recent psycholinguistic study (Pozniak et al., 2019), but this time with 8 French LLM "subjects".

Model	Lang.	Architecture	Reference
flaubert_base_uncased	fr	Bidirectional	Le et al. (2020)
camembert-base	fr	Bidirectional	Martin et al. (2020)
gpt2-base-french	fr	Autoregressive	ClassCat AI Lab (2022)
gpt2-wechsel-french	fr	Autoregressive	Minixhofer et al. (2022)
bert-base-multi-lingual-cased	multi	Bidirectional	Devlin et al. (2018)
xlm-roberta-base	multi	Bidirectional	Conneau et al. (2019)
xlm-roberta-large	multi	Bidirectional	Conneau et al. (2019)
xlm-mlm-17-1280	multi	Bidirectional	Lample and Conneau (2019)

Table 1: French models used in Experiment 1

Experiment 1: stimuli (reused from Pozniak et al. (2019))

(9) a. Perception ✓; tense match ✓ $\Longrightarrow$ RC-parse ✓; PR-parse ✓
Marie a écouté le ministre [qui critiquait le président].
Marie has listened-to the minister who criticized the president.

b. Perception ✓; tense match ✗ $\Longrightarrow$ RC-parse ✓; PR-parse ✗
? Marie écoute le ministre [qui critiquait le président].
Marie listens-to the minister who criticized the president.

c. Perception ✗; tense match ✓ $\Longrightarrow$ RC-parse ✓; PR-parse ✗
Marie a été mariée au ministre [qui critiquait le président].
Marie has been married to-the minister who criticized the president.

d. Perception ✗; tense match ✗ $\Longrightarrow$ RC-parse ✓; PR-parse ✗
Marie est mariée au ministre [qui critiquait le président].
Marie is married to-the minister who criticized the president.

- Ignoring Properties 3 & 4, **the stimuli in (9) are in principle ambiguous between a PR-parse and an RC-parse**.
- Violating Properties 3 or 4 *blocks* the PR parse.
- 18 such examples ("frames") were fed to the LLMs, for a total of $2_{\pm\text{perception}} \times 2_{\pm\text{tense-match}} \times 18_{\text{frame}} = 72$ sentences.

Experiment 1: stimuli (reused from Pozniak et al. (2019))

(9) a. Perception ✓; tense match ✓ $\Longrightarrow$ RC-parse ✓; PR-parse ✓
Marie a écouté le ministre [qui critiquait le président].
Marie has listened-to the minister who criticized the president.

b. Perception ✓; tense match ✗ $\Longrightarrow$ RC-parse ✓; PR-parse ✗
? Marie écoute le ministre [qui critiquait le président].
Marie listens-to the minister who criticized the president.

c. Perception ✗; tense match ✓ $\Longrightarrow$ RC-parse ✓; PR-parse ✗
Marie a été mariée au ministre [qui critiquait le président].
Marie has been married to-the minister who criticized the president.

d. Perception ✗; tense match ✗ $\Longrightarrow$ RC-parse ✓; PR-parse ✗
Marie est mariée au ministre [qui critiquait le président].
Marie is married to-the minister who criticized the president.

- Ignoring Properties 3 & 4, **the stimuli in (9) are in principle ambiguous between a PR-parse and an RC-parse**.
- Violating Properties 3 or 4 *blocks* the PR parse.
- 18 such examples (“frames”) were fed to the LLMs, for a total of $2_{\pm\text{perception}} \times 2_{\pm\text{tense-match}} \times 18_{\text{frame}} = 72$ sentences.

Experiment 1: stimuli (reused from Pozniak et al. (2019))

(9) a. Perception ✓; tense match ✓ $\Longrightarrow$ RC-parse ✓; PR-parse ✓
Marie a écouté le ministre [qui critiquait le président].
Marie has listened-to the minister who criticized the president.

b. Perception ✓; tense match ✗ $\Longrightarrow$ RC-parse ✓; PR-parse ✗
? Marie écoute le ministre [qui critiquait le président].
Marie listens-to the minister who criticized the president.

c. Perception ✗; tense match ✓ $\Longrightarrow$ RC-parse ✓; PR-parse ✗
Marie a été mariée au ministre [qui critiquait le président].
Marie has been married to-the minister who criticized the president.

d. Perception ✗; tense match ✗ $\Longrightarrow$ RC-parse ✓; PR-parse ✗
Marie est mariée au ministre [qui critiquait le président].
Marie is married to-the minister who criticized the president.

- Ignoring Properties 3 & 4, **the stimuli in (9) are in principle ambiguous between a PR-parse and an RC-parse**.
- Violating Properties 3 or 4 *blocks* the PR parse.
- 18 such examples ("frames") were fed to the LLMs, for a total of $2_{\pm\text{perception}} \times 2_{\pm\text{tense-match}} \times 18_{\text{frame}} = 72$ sentences.

Experiment 1: stimuli (reused from Pozniak et al. (2019))

(9) a. Perception ✓; tense match ✓ $\implies$ RC-parse ✓; PR-parse ✓
Marie a écouté le ministre [qui critiquait le président].
Marie has listened-to the minister who criticized the president.

b. Perception ✓; tense match ✗ $\implies$ RC-parse ✓; PR-parse ✗
? Marie écoute le ministre [qui critiquait le président].
Marie listens-to the minister who criticized the president.

c. Perception ✗; tense match ✓ $\implies$ RC-parse ✓; PR-parse ✗
Marie a été mariée au ministre [qui critiquait le président].
Marie has been married to-the minister who criticized the president.

d. Perception ✗; tense match ✗ $\implies$ RC-parse ✓; PR-parse ✗
Marie est mariée au ministre [qui critiquait le président].
Marie is married to-the minister who criticized the president.

- Ignoring Properties 3 & 4, **the stimuli in (9) are in principle ambiguous between a PR-parse and an RC-parse**.
- Violating Properties 3 or 4 *blocks* the PR parse.
- 18 such examples ("frames") were fed to the LLMs, for a total of $2_{\pm\text{perception}} \times 2_{\pm\text{tense-match}} \times 18_{\text{frame}} = 72$ sentences.

Experiment 1: testing

- Building on Surprisal Theory (Hale, 2001; Levy, 2008), our proxy for grammaticality was taken to be the **log-probability assigned to a given sentence** by the LLM. It was computed using the `minicons` library (Misra, 2022).

$$\textsc{Grammaticality}(w_t) \simeq -\textsc{Surprisal}(w_t)$$
$$= \log P(w_t | w_1 \ldots w_{t-1})^3$$
$$\textsc{Grammaticality}(w_1 \ldots w_t) \simeq -\sum_{i=1}^{t} \textsc{Surprisal}(w_i)$$

- Effects assessed using linear mixed-effect modeling (performed with `statsmodels`, Seabold and Perktold (2010)).

[3] In the case of BERT-like bidirectional models, this formula is adapted to masked language modeling: the probability of a word is computed given its left *and* right context.

Experiment 1: testing

- Building on Surprisal Theory (Hale, 2001; Levy, 2008), our proxy for grammaticality was taken to be the **log-probability assigned to a given sentence** by the LLM. It was computed using the `minicons` library (Misra, 2022).

$$\begin{aligned}\textsc{Grammaticality}(w_t) &\simeq -\textsc{Surprisal}(w_t) \\ &= \log P(w_t | w_1 \ldots w_{t-1})^3 \\ \textsc{Grammaticality}(w_1 \ldots w_t) &\simeq -\sum_{i=1}^{t} \textsc{Surprisal}(w_i)\end{aligned}$$

- Effects assessed using linear mixed-effect modeling (performed with `statsmodels`, Seabold and Perktold (2010)).

[3] In the case of BERT-like bidirectional models, this formula is adapted to masked language modeling: the probability of a word is computed given its left *and* right context.

Experiment 1: testing

- Building on Surprisal Theory (Hale, 2001; Levy, 2008), our proxy for grammaticality was taken to be the **log-probability assigned to a given sentence** by the LLM. It was computed using the `minicons` library (Misra, 2022).

$$\begin{aligned}
\textsc{Grammaticality}(w_t) &\simeq -\textsc{Surprisal}(w_t) \\
&= \log P(w_t|w_1 \ldots w_{t-1})^3 \\
\textsc{Grammaticality}(w_1 \ldots w_t) &\simeq -\sum_{i=1}^{t} \textsc{Surprisal}(w_i)
\end{aligned}$$

- Effects assessed using linear mixed-effect modeling (performed with `statsmodels`, Seabold and Perktold (2010)).

[3] In the case of BERT-like bidirectional models, this formula is adapted to masked language modeling: the probability of a word is computed given its left *and* right context.

Experiment 1: predictions and controls

- We overall expect a main effect of verb type and tense anaphoricity, but above all, an **interaction between those two variables**.
- Following the design of Pozniak et al. (2019), English models were also tested on similar stimuli. Because English is not a PR language, no interaction is expected.

Model	Lang.	Architecture	Reference
bert-large-cased	en	Bidirectional	Devlin et al. (2018)
gpt2-large	en	Autoregressive	Radford et al. (2019)
xlnet-large-cased	en	Autoregressive	Yang et al. (2019)
bert-base-multi-lingual-cased	multi	Bidirectional	Devlin et al. (2018)
xlm-roberta-base	multi	Bidirectional	Conneau et al. (2019)
xlm-roberta-large	multi	Bidirectional	Conneau et al. (2019)
xlm-mlm-17-1280	multi	Bidirectional	Lample and Conneau (2019)

Table 2: English models used in Exp. 1 (some overlap with the French models)

Experiment 1: predictions and controls

- We overall expect a main effect of verb type and tense anaphoricity, but above all, an **interaction between those two variables**.
- Following the design of Pozniak et al. (2019), English models were also tested on similar stimuli. Because English is not a PR language, no interaction is expected.

Model	Lang.	Architecture	Reference
bert-large-cased	en	Bidirectional	Devlin et al. (2018)
gpt2-large	en	Autoregressive	Radford et al. (2019)
xlnet-large-cased	en	Autoregressive	Yang et al. (2019)
bert-base-multi-lingual-cased	multi	Bidirectional	Devlin et al. (2018)
xlm-roberta-base	multi	Bidirectional	Conneau et al. (2019)
xlm-roberta-large	multi	Bidirectional	Conneau et al. (2019)
xlm-mlm-17-1280	multi	Bidirectional	Lample and Conneau (2019)

Table 2: English models used in Exp. 1 (some overlap with the French models)

Experiment 1: results for French (target language)

- 6/8 LLMs favor matching tenses (cf. col. 5), and 4/8 more so under perception verbs (`verb*tense` interaction, cf. col. 6).

Model	Lang.	Best AIC?	`verb_type`	`tense`	interaction
flaubert_base_uncased	fr	n	.	n.s.	.
camembert-base	fr	n	.	**	n.s.
gpt2-base-french	fr	y	n.s.	**	*
gpt2-wechsel-french	fr	y	n.s.	**	.
bert-base-multi-lingual-cased	multi	n	n.s.	**	n.s.
xlm-roberta-base	multi	y	n.s.	**	.
xlm-roberta-large	multi	y	n.s.	**	*
xlm-mlm-17-1280	multi	n	**	n.s.	n.s.

Table 3: Significance results of LME modeling for $\texttt{grammaticality} \sim \texttt{verb\_type} + \texttt{tense} + \texttt{verb\_type} * \texttt{tense} + (1|\texttt{frame})$, where `frame` refers to the lexical skeleton shared by all stimuli in e.g. (9).[4]

[4] The 'Best AIC?' column specifies if the formula yielded the lowest Akaike Information Criterion, as opposed to other simpler formulas without interactions or main effects. Other notations: '.' $= p \in].05; .1]$, '*'$= p \in].01; .05]$; '**'$= p \in [0; .01]$; ✓=coefficient validates the hypothesis; ✗=coefficient disproves the hypothesis.

Experiment 1: results for English (control language)

- English models did not exhibit similar effects. This is true even of the bilingual models which identified contrasts in French.
- Strikingly also, all but 1 model did *not* yield the best AIC for the formula involving an interaction term.
- This is consistent with English not allowing pseudorelatives.

Model	Lang.	Best AIC?	verb type	tense	interaction
bert-large-cased	en	n	n.s.	n.s.	.
gpt2-large	en	n	**	*	n.s.
xlnet-large-cased	en	n	n.s.	n.s.	n.s.
bert-base-multi-lingual-cased	multi	y	**	*	*
xlm-roberta-base	multi	n	.	.	.
xlm-roberta-large	multi	n	n.s.	*	n.s.
xlm-mlm-17-1280	multi	n	**	n.s.	n.s.

Table 4: Significance results of LME modeling on the English data ($\texttt{grammaticality} \sim \texttt{verb\_type} + \texttt{tense} + \texttt{verb\_type} * \texttt{tense} + (1|\texttt{frame})$). $2_{\pm\text{perception}} \times 2_{\pm\text{tense-match}} \times 21_{\text{frame}} = 84$ sentences were tested.

Experiment 1: results for English (control language)

- English models did not exhibit similar effects. This is true even of the bilingual models which identified contrasts in French.
- Strikingly also, all but 1 model did *not* yield the best AIC for the formula involving an interaction term.
- This is consistent with English not allowing pseudorelatives.

Model	Lang.	Best AIC?	verb type	tense	interaction
bert-large-cased	en	n	n.s.	n.s.	.
gpt2-large	en	n	**	*	n.s.
xlnet-large-cased	en	n	n.s.	n.s.	n.s.
bert-base-multi-lingual-cased	multi	y	**	*	*
xlm-roberta-base	multi	n	.	.	.
xlm-roberta-large	multi	n	n.s.	*	n.s.
xlm-mlm-17-1280	multi	n	**	n.s.	n.s.

Table 4: Significance results of LME modeling on the English data ($\texttt{grammaticality} \sim \texttt{verb\_type} + \texttt{tense} + \texttt{verb\_type} * \texttt{tense} + (1|\texttt{frame})$). $2_{\pm\text{perception}} \times 2_{\pm\text{tense-match}} \times 21_{\text{frame}} = 84$ sentences were tested.

Experiment 1: results for English (control language)

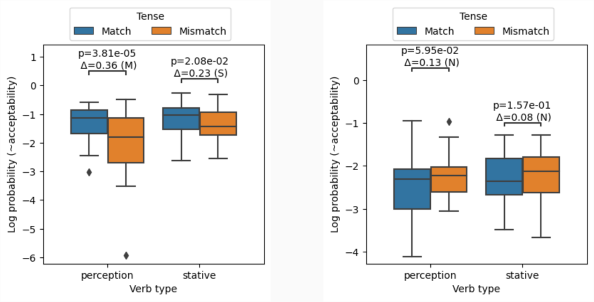
- English models did not exhibit similar effects. This is true even of the bilingual models which identified contrasts in French.
- Strikingly also, all but 1 model did *not* yield the best AIC for the formula involving an interaction term.
- This is consistent with English not allowing pseudorelatives.

Model	Lang.	Best AIC?	verb type	tense	interaction
bert-large-cased	en	n	n.s.	n.s.	.
gpt2-large	en	n	**	*	n.s.
xlnet-large-cased	en	n	n.s.	n.s.	n.s.
bert-base-multi-lingual-cased	multi	y	**	*	*
xlm-roberta-base	multi	n	.	.	.
xlm-roberta-large	multi	n	n.s.	*	n.s.
xlm-mlm-17-1280	multi	n	**	n.s.	n.s.

Table 4: Significance results of LME modeling on the English data ($\texttt{grammaticality} \sim \texttt{verb\_type} + \texttt{tense} + \texttt{verb\_type} * \texttt{tense} + (1|\texttt{frame})$). $2_{\pm\text{perception}} \times 2_{\pm\text{tense-match}} \times 21_{\text{frame}} = 84$ sentences were tested.

Experiment 1: a well-performing bilingual model

- Visible disadvantage of tense mismatching structures under perception verbs in French; no such contrast in English.



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

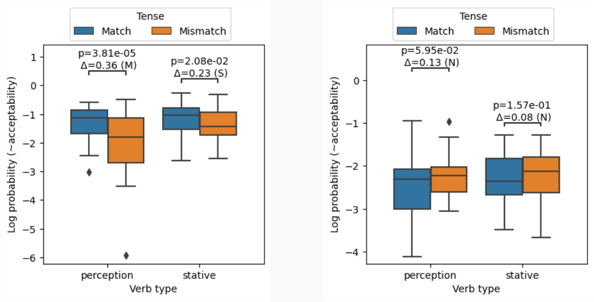
Figure 1: Grammaticality scores obtained with `xlm-roberta-large`.[5]

- **Limitation: the result could be incorrectly driven by the RC-parse... we want a design that isolates the PR-parse!!**

[5]The scores are overall negative because they correspond to negative log probabilities. Δ=Cliff's Delta. N, S, M resp. mean 'negligible', 'small', 'medium'.

Experiment 1: a well-performing bilingual model

- Visible disadvantage of tense mismatching structures under perception verbs in French; no such contrast in English.



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

Figure 1: Grammaticality scores obtained with `xlm-roberta-large`.[5]

- **Limitation: the result could be incorrectly driven by the RC-parse... we want a design that isolates the PR-parse!!**

[5]The scores are overall negative because they correspond to negative log probabilities. Δ=Cliff's Delta. N, S, M resp. mean 'negligible', 'small', 'medium'.

Experiment 2: cliticization, gap, & embedding verb type

Experiment 2: goal and methodology

- Recall that cliticizing the head noun of a relative allows to block a standard RC-parse, to only retain the PR-parse (Property 1).
- **This allows to test Property 2 (subject-gap requirement) and Property 3 (matrix perception verb requirement) without an interference of the standard RC-parse.**
- We do so using the same 8 French LLMs as before, feeding them with $2_{\pm\text{subject-gap}} \times 2_{\pm\text{clitic}} \times (10_{\text{perception}} + 10_{\text{psych}} + 10_{\text{action}})_{\text{matrix-verb}} \times 10_{\text{embedded-verb}} \times 4^5 = 4800$ semi-automatically generated sentences.

(10) Glossed template of the stimuli:

$$\underset{\text{Subject}}{\begin{Bmatrix} \text{He} \\ \text{She} \end{Bmatrix}} \underset{\text{(CL)}}{\begin{Bmatrix} \text{him.CL} \\ \text{her.CL} \\ \emptyset \end{Bmatrix}} \underset{\text{V}}{\begin{Bmatrix} \text{sees/...} \\ \text{thinks/...} \\ \text{greets/...} \end{Bmatrix}} \underset{\text{(Object)}}{\begin{Bmatrix} \emptyset \\ \text{Marie} \\ \text{Jean} \end{Bmatrix}} \underset{\text{Relative}}{\begin{Bmatrix} \text{subject-gap relative} \\ \text{object-gap relative} \end{Bmatrix}}$$

[5] This last multiplicative factor comes from the fact that the matrix and embedded subject were balanced for gender: $4 = 2_{\pm\text{fem-matrix-subj}} \times 2_{\pm\text{fem-emb-subj}}$.

Experiment 2: goal and methodology

- Recall that cliticizing the head noun of a relative allows to block a standard RC-parse, to only retain the PR-parse (Property 1).
- **This allows to test Property 2 (subject-gap requirement) and Property 3 (matrix perception verb requirement) without an interference of the standard RC-parse.**
- We do so using the same 8 French LLMs as before, feeding them with $2_{\pm\text{subject-gap}} \times 2_{\pm\text{clitic}} \times (10_{\text{perception}} + 10_{\text{psych}} + 10_{\text{action}})_{\text{matrix-verb}} \times 10_{\text{embedded-verb}} \times 4^5 = 4800$ semi-automatically generated sentences.

(10) Glossed template of the stimuli:

$$\begin{array}{ccccc} \text{Subject} & \text{(CL)} & \text{V} & \text{(Object)} & \text{Relative} \\ \left\{\begin{array}{c}\text{He}\\ \text{She}\end{array}\right\} & \left\{\begin{array}{c}\text{him.CL}\\ \text{her.CL}\\ \emptyset\end{array}\right\} & \left\{\begin{array}{c}\text{sees/...}\\ \text{thinks/...}\\ \text{greets/...}\end{array}\right\} & \left\{\begin{array}{c}\emptyset\\ \text{Marie}\\ \text{Jean}\end{array}\right\} & \left\{\begin{array}{c}\text{subject-gap relative}\\ \text{object-gap relative}\end{array}\right\} \end{array}$$

[5] This last multiplicative factor comes from the fact that the matrix and embedded subject were balanced for gender: $4 = 2_{\pm\text{fem-matrix-subj}} \times 2_{\pm\text{fem-emb-subj}}$.

Experiment 2: goal and methodology

- Recall that cliticizing the head noun of a relative allows to block a standard RC-parse, to only retain the PR-parse (Property 1).
- **This allows to test Property 2 (subject-gap requirement) and Property 3 (matrix perception verb requirement) without an interference of the standard RC-parse.**
- We do so using the same 8 French LLMs as before, feeding them with $2_{\pm\text{subject-gap}} \times 2_{\pm\text{clitic}} \times (10_{\text{perception}} + 10_{\text{psych}} + 10_{\text{action}})_{\text{matrix-verb}} \times 10_{\text{embedded-verb}} \times 4$[5] $= 4800$ semi-automatically generated sentences.

(10) Glossed template of the stimuli:

$$\begin{array}{ccccc} \text{Subject} & \text{(CL)} & \text{V} & \text{(Object)} & \text{Relative} \\ \left\{\begin{array}{c} \text{He} \\ \text{She} \end{array}\right\} & \left\{\begin{array}{c} \text{him.CL} \\ \text{her.CL} \\ \emptyset \end{array}\right\} & \left\{\begin{array}{c} \text{sees/...} \\ \text{thinks/...} \\ \text{greets/...} \end{array}\right\} & \left\{\begin{array}{c} \emptyset \\ \text{Marie} \\ \text{Jean} \end{array}\right\} & \left\{\begin{array}{c} \text{subject-gap relative} \\ \text{object-gap relative} \end{array}\right\} \end{array}$$

[5] This last multiplicative factor comes from the fact that the matrix and embedded subject were balanced for gender: $4 = 2_{\pm\text{fem-matrix-subj}} \times 2_{\pm\text{fem-emb-subj}}$.

Experiment 2: goal and methodology

- Recall that cliticizing the head noun of a relative allows to block a standard RC-parse, to only retain the PR-parse (Property 1).
- **This allows to test Property 2 (subject-gap requirement) and Property 3 (matrix perception verb requirement) without an interference of the standard RC-parse.**
- We do so using the same 8 French LLMs as before, feeding them with $2_{\pm\text{subject-gap}} \times 2_{\pm\text{clitic}} \times (10_{\text{perception}} + 10_{\text{psych}} + 10_{\text{action}})_{\text{matrix-verb}} \times 10_{\text{embedded-verb}} \times 4^5 = 4800$ semi-automatically generated sentences.

(10) Glossed template of the stimuli:

$$\begin{array}{ccccc} \text{Subject} & \text{(CL)} & \text{V} & \text{(Object)} & \text{Relative} \\ \left\{\begin{array}{c}\text{He}\\\text{She}\end{array}\right\} & \left\{\begin{array}{c}\text{him.CL}\\\text{her.CL}\\\emptyset\end{array}\right\} & \left\{\begin{array}{c}\text{sees/...}\\\text{thinks/...}\\\text{greets/...}\end{array}\right\} & \left\{\begin{array}{c}\emptyset\\\text{Marie}\\\text{Jean}\end{array}\right\} & \left\{\begin{array}{c}\text{subject-gap relative}\\\text{object-gap relative}\end{array}\right\} \end{array}$$

[5] This last multiplicative factor comes from the fact that the matrix and embedded subject were balanced for gender: $4 = 2_{\pm\text{fem-matrix-subj}} \times 2_{\pm\text{fem-emb-subj}}$.

Experiment 2: results

- Robust preference for subject-gaps (8/8 models, cf. col. 3) and more so under perception verbs (5/8 models, cf. col. 6).
- The desired `clitic*gap*verb_type` interaction however, was only captured by 1/8 models (cf. col. 8).
- Additionally, the interaction between cliticization and subject-gaps (cf. col. 7) is predicted by most models to have a *negative* effect on grammaticality scores (!!)

Model	verb type	gap	clitic	v*c	v*g	c*g	v*c*g
flaubert_base_uncased	.	**	**	**	.	**	n.s.
camembert-base	.	**	**	**	**	**	n.s.
gpt2-base-french	n.s.	**	**	**	**	**	.
gpt2-wechsel-french	n.s.	**	**	**	**	**	**
bert-base-multi-lingual-cased	n.s.	**	**	**	n.s.	**	n.s.
xlm-roberta-base	n.s.	**	**	**	**	**	.
xlm-roberta-large	n.s.	**	**	**	**	**	**
xlm-mlm-17-1280	n.s.	**	*	**	**	**	n.s.

Table 5: Significance results of LME modeling for
`grammaticality` $\sim$ `verb_type` + `gap` + `clitic` + `verb_type` $*$ `clitic` $*$ `gap`.

Experiment 2: results

- Robust preference for subject-gaps (8/8 models, cf. col. 3) and more so under perception verbs (5/8 models, cf. col. 6).
- The desired `clitic*gap*verb_type` interaction however, was only captured by 1/8 models (cf. col. 8).
- Additionally, the interaction between cliticization and subject-gaps (cf. col. 7) is predicted by most models to have a *negative* effect on grammaticality scores (!!)

Model	verb type	gap	clitic	v*c	v*g	c*g	v*c*g
flaubert_base_uncased	.	**	**	**	.	**	n.s.
camembert-base	.	**	**	**	**	**	n.s.
gpt2-base-french	n.s.	**	**	**	**	**	.
gpt2-wechsel-french	n.s.	**	**	**	**	**	**
bert-base-multi-lingual-cased	n.s.	**	**	**	n.s.	**	n.s.
xlm-roberta-base	n.s.	**	**	**	**	**	.
xlm-roberta-large	n.s.	**	**	**	**	**	**
xlm-mlm-17-1280	n.s.	**	*	**	**	**	n.s.

Table 5: Significance results of LME modeling for
$\texttt{grammaticality} \sim \texttt{verb\_type} + \texttt{gap} + \texttt{clitic} + \texttt{verb\_type} * \texttt{clitic} * \texttt{gap}$.

Experiment 2: results

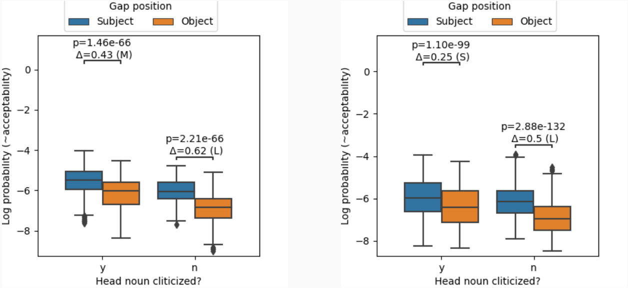
- Robust preference for subject-gaps (8/8 models, cf. col. 3) and more so under perception verbs (5/8 models, cf. col. 6).
- The desired `clitic*gap*verb_type` interaction however, was only captured by 1/8 models (cf. col. 8).
- Additionally, the interaction between cliticization and subject-gaps (cf. col. 7) is predicted by most models to have a *negative* effect on grammaticality scores (!!)

Model	verb type	gap	clitic	v*c	v*g	c*g	v*c*g
flaubert_base_uncased	.	**	**	**	.	**	n.s.
camembert-base	.	**	**	**	**	**	n.s.
gpt2-base-french	n.s.	**	**	**	**	**	.
gpt2-wechsel-french	n.s.	**	**	**	**	**	**
bert-base-multi-lingual-cased	n.s.	**	**	**	n.s.	**	n.s.
xlm-roberta-base	n.s.	**	**	**	**	**	.
xlm-roberta-large	n.s.	**	**	**	**	**	**
xlm-mlm-17-1280	n.s.	**	*	**	**	**	n.s.

Table 5: Significance results of LME modeling for
$\texttt{grammaticality} \sim \texttt{verb\_type} + \texttt{gap} + \texttt{clitic} + \texttt{verb\_type} * \texttt{clitic} * \texttt{gap}$.

Experiment 2: a not-so-badly-performing French-only model

- Very visible main effect of subject-gap.
- The contrast between object- and subject-gap is:
    - stronger under perception verbs, yay!
    - weaker when the structure is cliticized :(((
- Lastly, the contrast between cliticized and non-cliticized structures stronger under perception verbs :))

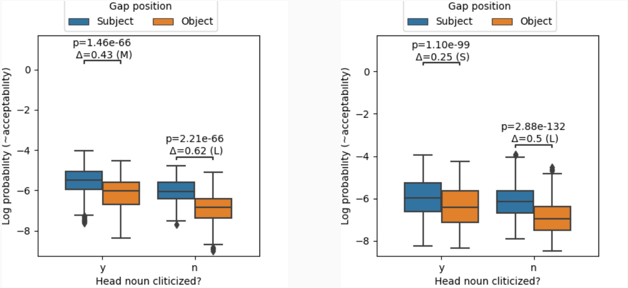


(a) Perception verbs **(b)** Attitude/action verbs

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

Experiment 2: a not-so-badly-performing French-only model

- Very visible main effect of subject-gap.
- The contrast between object- and subject-gap is:
  - stronger under perception verbs, yay!
  - weaker when the structure is cliticized :(((
- Lastly, the contrast between cliticized and non-cliticized structures stronger under perception verbs :))

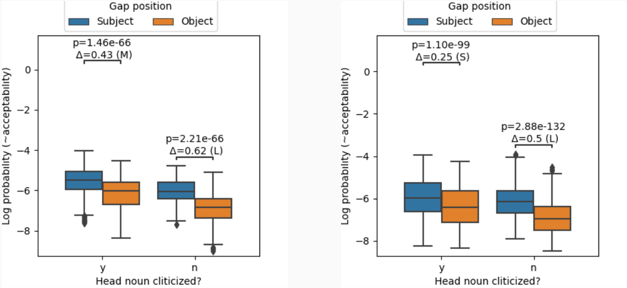


(a) Perception verbs **(b)** Attitude/action verbs

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

Experiment 2: a not-so-badly-performing French-only model

- Very visible main effect of subject-gap.
- The contrast between object- and subject-gap is:
    - stronger under perception verbs, yay!
    - weaker when the structure is cliticized :(((
- Lastly, the contrast between cliticized and non-cliticized structures stronger under perception verbs :))

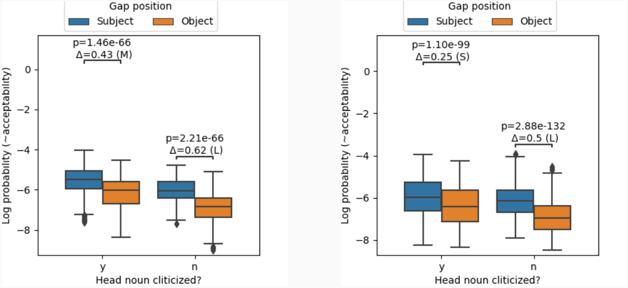


(a) Perception verbs **(b)** Attitude/action verbs

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

Experiment 2: a not-so-badly-performing French-only model

- Very visible main effect of subject-gap.
- The contrast between object- and subject-gap is:
  - stronger under perception verbs, yay!
  - weaker when the structure is cliticized :(((
- Lastly, the contrast between cliticized and non-cliticized structures stronger under perception verbs :))

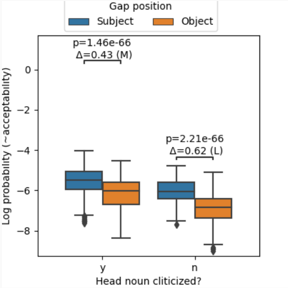


(a) Perception verbs **(b)** Attitude/action verbs

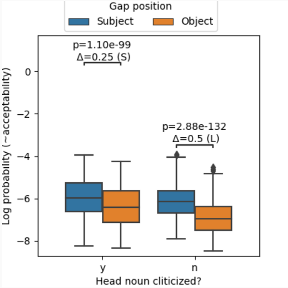
Figure 2: Grammaticality scores obtained with `gpt2-base-french`.

Experiment 2: a not-so-badly-performing French-only model

- Very visible main effect of subject-gap.
- The contrast between object- and subject-gap is:
  - stronger under perception verbs, yay!
  - weaker when the structure is cliticized :(((
- Lastly, the contrast between cliticized and non-cliticized structures stronger under perception verbs :))



(a) Perception verbs



(b) Attitude/action verbs

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

Experiment 3: the “existence commitment” inference

A last, semantic property of the PR

- PRs, unlike **infinitival or gerundive complements** (but, roughly, like RCs), imply the existence/truth of the embedded event *even under matrix negation*.
- Following Moulton and Grillo (2015), we call this the **existence commitment** (**EC**).

(11) a. Negation + relative (PR or RC parse):

Je ne vois pas Marie [qui danse].
I NEG see NEG Marie who dances.

'I don't see Marie dancing (**EC: yet she does!**).'

b. Negation + relative + cliticization (PR parse only):

Je ne la vois pas [qui danse].
I NEG her.CL see NEG who dances.

'I don't see her dancing (**EC: yet she does!**).'

c. Negation + infinitive (+ cliticization):

Je ne (la) vois pas (Marie) [danser].
I NEG her.CL see NEG Marie to-dance.

'I don't see her/Marie dance (**she may or may not**).'

A last, semantic property of the PR

- PRs, unlike **infinitival or gerundive complements** (but, roughly, like RCs), imply the existence/truth of the embedded event *even under matrix negation*.
- Following Moulton and Grillo (2015), we call this the **existence commitment** (**EC**).

(11) a. Negation + relative (PR or RC parse):
Je ne vois pas Marie [qui danse].
I NEG see NEG Marie who dances.
'I don't see Marie dancing (**EC: yet she does!**).'

b. Negation + relative + cliticization (PR parse only):
Je ne la vois pas [qui danse].
I NEG her.CL see NEG who dances.
'I don't see her dancing (**EC: yet she does!**).'

c. Negation + infinitive (+ cliticization):
Je ne (la) vois pas (Marie) [danser].
I NEG her.CL see NEG Marie to-dance.
'I don't see her/Marie dance (**she may or may not**).'

A last, semantic property of the PR

- PRs, unlike **infinitival or gerundive complements** (but, roughly, like RCs), imply the existence/truth of the embedded event *even under matrix negation*.
- Following Moulton and Grillo (2015), we call this the **existence commitment** (**EC**).

(11) a. Negation + relative (PR or RC parse):

Je	ne	vois	pas	Marie	[qui	danse].	
I	NEG	see	NEG	Marie	who	dances.	

‘I don’t see Marie dancing (**EC: yet she does!**).’

b. Negation + relative + cliticization (PR parse only):

Je	ne	la	vois	pas	[qui	danse].
I	NEG	her.CL	see	NEG	who	dances.

‘I don’t see her dancing (**EC: yet she does!**).’

c. Negation + infinitive (+ cliticization):

Je	ne	(la)	vois	pas	(Marie)	[danser].
I	NEG	her.CL	see	NEG	Marie	to-dance.

‘I don’t see her/Marie dance (**she may or may not**).’

Experiment 3: design

- Given a negated matrix perception verb embedding a subject-gap clause $\mathscr{C}$, either as an infinitive or as a (pseudo)relative, with or without cliticization, we measure the likelihood of the EC.

(12) Glossed template for the stimuli

$$\begin{array}{cccccccc}
\text{Subject} & \text{not} & \text{(CL)} & \text{V} & \text{not} & \text{(Object)} & \text{Embedded clause} \\
\left\{\begin{array}{c}\text{He}\\\text{She}\end{array}\right\} & \text{NEG} & \left\{\begin{array}{c}\text{him.CL}\\\text{her.CL}\\\emptyset\end{array}\right\} & \left\{\text{voit/...}\right\} & \text{NEG} & \left\{\begin{array}{c}\emptyset\\\text{Marie}\\\text{Jean}\end{array}\right\} & \left\{\begin{array}{c}\text{relative}\\\text{infinitive}\end{array}\right\}
\end{array}$$

- $2_{\pm\text{clitic}} \times 2_{\pm\text{relative}} \times 6_{\text{matrix-perception}} \times 10_{\text{emb-action}} \times 4$[6] $= 960$ sentences following the above template were fed to 4 BERT-like LLMs fine-tuned to perform natural language inference (NLI).

[6] The last factor again comes from gender swaps on the matrix and embedded subjects.

Experiment 3: design

- Given a negated matrix perception verb embedding a subject-gap clause $\mathscr{C}$, either as an infinitive or as a (pseudo)relative, with or without cliticization, we measure the likelihood of the EC.

(12) Glossed template for the stimuli

$$\begin{array}{ccccccc} \text{Subject} & \text{not} & \text{(CL)} & \text{V} & \text{not} & \text{(Object)} & \text{Embedded clause} \\ \left\{\begin{array}{c}\text{He}\\ \text{She}\end{array}\right\} & \text{NEG} & \left\{\begin{array}{c}\text{him.CL}\\ \text{her.CL}\\ \emptyset\end{array}\right\} & \left\{\text{ voit/... }\right\} & \text{NEG} & \left\{\begin{array}{c}\emptyset\\ \text{Marie}\\ \text{Jean}\end{array}\right\} & \left\{\begin{array}{c}\text{relative}\\ \text{infinitive}\end{array}\right\} \end{array}$$

- $2_{\pm\text{clitic}} \times 2_{\pm\text{relative}} \times 6_{\text{matrix-perception}} \times 10_{\text{emb-action}} \times 4^{6} = 960$ sentences following the above template were fed to 4 BERT-like LLMs fine-tuned to perform natural language inference (NLI).

[6] The last factor again comes from gender swaps on the matrix and embedded subjects.

Experiment 3: design

- Given a negated matrix perception verb embedding a subject-gap clause $\mathscr{C}$, either as an infinitive or as a (pseudo)relative, with or without cliticization, we measure the likelihood of the EC.

(12) Glossed template for the stimuli

$$\begin{array}{cccccccc} \text{Subject} & \text{not} & \text{(CL)} & \text{V} & \text{not} & \text{(Object)} & \text{Embedded clause} \\ \left\{\begin{array}{c}\text{He}\\ \text{She}\end{array}\right\} & \text{NEG} & \left\{\begin{array}{c}\text{him.CL}\\ \text{her.CL}\\ \emptyset\end{array}\right\} & \left\{\text{voit/...}\right\} & \text{NEG} & \left\{\begin{array}{c}\emptyset\\ \text{Marie}\\ \text{Jean}\end{array}\right\} & \left\{\begin{array}{c}\text{relative}\\ \text{infinitive}\end{array}\right\} \end{array}$$

- $2_{\pm\text{clitic}} \times 2_{\pm\text{relative}} \times 6_{\text{matrix-perception}} \times 10_{\text{emb-action}} \times 4^6 = 960$ sentences following the above template were fed to 4 BERT-like LLMs fine-tuned to perform natural language inference (NLI).

[6] The last factor again comes from gender swaps on the matrix and embedded subjects.

Experiment 3: predictions

Model	Lang.	Architecture	Reference
camembert-base-xnli	fr	Bidirectional	Doyen, 2023
xlm-roberta-large-xnli-finetuned-mnli	multi	Bidirectional	Özsoy, 2022
mDeBERTa-v3-base-mnli-xnli	multi	Bidirectional	Laurer et al., 2022
mDeBERTa-v3-base-xnli-multilingual-nli-2mil7	multi	Bidirectional	Laurer et al., 2022

Table 6: Models used in Experiment 3

- **We expect the EC to be overall stronger when the embedded clause is a relative as opposed to an infinitive**, whether of not the head noun is cliticized.

Experiment 3: results

- Embedded relatives systematically lead to a stronger EC as opposed to infinitives (cf. col. 3).
- However, non-cliticized subjects also lead to a stronger EC *across the board* (col. 4)!!
- **This all suggests that LLMs associate the EC with the occurrence of RCs, but not really PRs...**

Model	Best AIC?	embedded clause (RC)	clitic	RC/clitic interaction
camembert-base-xnli	y	**	**	**
xlm-roberta-large-xnli-finetuned-mnli	y	**	**	**
mDeBERTa-v3-base-mnli-xnli	y	**	**	**
mDeBERTa-v3-base-xnli-multilingual-nli-2mil7	y	**	**	**

Table 7: Significance results of LME modeling for
$\texttt{EC\_strength} \sim \texttt{emb\_clause} + \texttt{clitic} + \texttt{emb\_clause} * \texttt{clitic}$.

Experiment 3: results

- Embedded relatives systematically lead to a stronger EC as opposed to infinitives (cf. col. 3).
- However, non-cliticized subjects also lead to a stronger EC *across the board* (col. 4)!!
- **This all suggests that LLMs associate the EC with the occurrence of RCs, but not really PRs...**

Model	Best AIC?	embedded clause (RC)	clitic	RC/clitic interaction
camembert-base-xnli	y	**	**	**
xlm-roberta-large-xnli-finetuned-mnli	y	**	**	**
mDeBERTa-v3-base-mnli-xnli	y	**	**	**
mDeBERTa-v3-base-xnli-multilingual-nli-2mil7	y	**	**	**

Table 7: Significance results of LME modeling for $\texttt{EC\_strength} \sim \texttt{emb\_clause} + \texttt{clitic} + \texttt{emb\_clause} * \texttt{clitic}$.

Experiment 3: results

- Embedded relatives systematically lead to a stronger EC as opposed to infinitives (cf. col. 3).
- However, non-cliticized subjects also lead to a stronger EC *across the board* (col. 4)!!
- **This all suggests that LLMs associate the EC with the occurrence of RCs, but not really PRs...**

Model	Best AIC?	embedded clause (RC)	clitic	RC/clitic interaction
camembert-base-xnli	y	**	**	**
xlm-roberta-large-xnli-finetuned-mnli	y	**	**	**
mDeBERTa-v3-base-mnli-xnli	y	**	**	**
mDeBERTa-v3-base-xnli-multilingual-nli-2mil7	y	**	**	**

Table 7: Significance results of LME modeling for $\texttt{EC\_strength} \sim \texttt{emb\_clause} + \texttt{clitic} + \texttt{emb\_clause} * \texttt{clitic}$.

Experiment 3: results

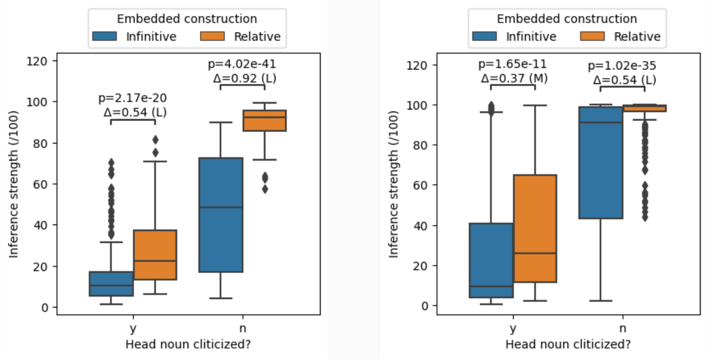
- Embedded relatives systematically lead to a stronger EC as opposed to infinitives (cf. col. 3).
- However, non-cliticized subjects also lead to a stronger EC *across the board* (col. 4)!!
- **This all suggests that LLMs associate the EC with the occurrence of RCs, but not really PRs...**

Model	Best AIC?	embedded clause (RC)	clitic	RC/clitic interaction
camembert-base-xnli	y	**	**	**
xlm-roberta-large-xnli-finetuned-mnli	y	**	**	**
mDeBERTa-v3-base-mnli-xnli	y	**	**	**
mDeBERTa-v3-base-xnli-multilingual-nli-2mil7	y	**	**	**

Table 7: Significance results of LME modeling for $\texttt{EC\_strength} \sim \texttt{emb\_clause} + \texttt{clitic} + \texttt{emb\_clause} * \texttt{clitic}$.

Experiment 3: two best-but-not-so-well performing NLI-LLMs

- Clear contrast between infinitival and relative complementation.
- But also, main effect of cliticization...
- In particular, cliticized constructions featuring an embedded relative (unambiguously PRs), do not lead at all to a strong EC :((

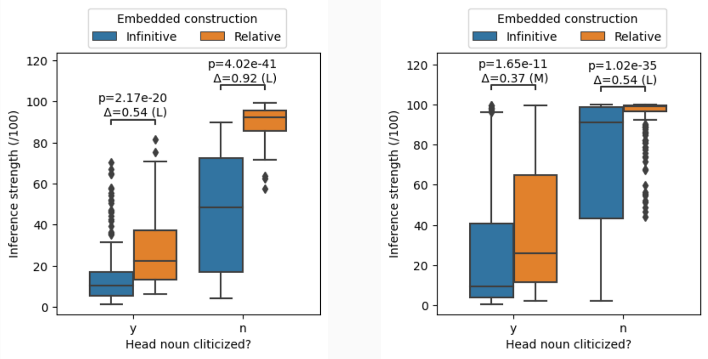


(a) CamemBERT **(b)** mDeBERTa

Figure 3: Distributions of the EC's strength scores (/100)

Experiment 3: two best-but-not-so-well performing NLI-LLMs

- Clear contrast between infinitival and relative complementation.
- But also, main effect of cliticization...
- In particular, cliticized constructions featuring an embedded relative (unambiguously PRs), do not lead at all to a strong EC :((

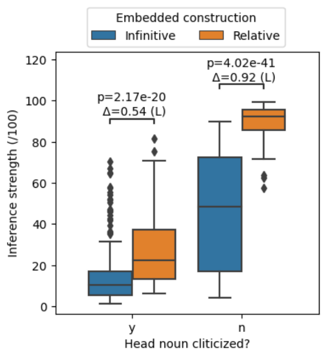


(a) CamemBERT **(b)** mDeBERTa

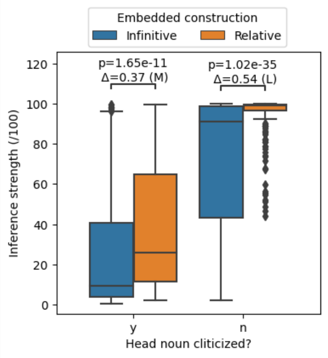
Figure 3: Distributions of the EC's strength scores (/100)

Experiment 3: two best-but-not-so-well performing NLI-LLMs

- Clear contrast between infinitival and relative complementation.
- But also, main effect of cliticization...
- In particular, cliticized constructions featuring an embedded relative (unambiguously PRs), do not lead at all to a strong EC :((



(a) CamemBERT



(b) mDeBERTa

Figure 3: Distributions of the EC's strength scores (/100)

Conclusion

- In this work, we investigated a structure (the pseudorelative) which is relatively **rare** in corpora and often **string-ambiguous** with a standard relative clause, making it challenging for LLMs to learn.
- 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, or inference patterns in Experiment 3.
- **This still raises the question whether LLMs really get the specificity of the pseudorelative as a syntactic construction (Experiment 2) with a specific semantics (Experiment 3); or whether they simply recycle general processing heuristics or biases applicable to other structures (e.g. standard RCs)...**

Conclusion

- In this work, we investigated a structure (the pseudorelative) which is relatively **rare** in corpora and often **string-ambiguous** with a standard relative clause, making it challenging for LLMs to learn.
- 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, or inference patterns in Experiment 3.
- **This still raises the question whether LLMs really get the specificity of the pseudorelative as a syntactic construction (Experiment 2) with a specific semantics (Experiment 3); or whether they simply recycle general processing heuristics or biases applicable to other structures (e.g. standard RCs)...**

Conclusion

- In this work, we investigated a structure (the pseudorelative) which is relatively **rare** in corpora and often **string-ambiguous** with a standard relative clause, making it challenging for LLMs to learn.
- 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, or inference patterns in Experiment 3.
- **This still raises the question whether LLMs really get the specificity of the pseudorelative as a syntactic construction (Experiment 2) with a specific semantics (Experiment 3); or whether they simply recycle general processing heuristics or biases applicable to other structures (e.g. standard RCs)...**

Conclusion

- In this work, we investigated a structure (the pseudorelative) which is relatively **rare** in corpora and often **string-ambiguous** with a standard relative clause, making it challenging for LLMs to learn.
- 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, or inference patterns in Experiment 3.
- **This still raises the question whether LLMs really get the specificity of the pseudorelative as a syntactic construction (Experiment 2) with a specific semantics (Experiment 3); or whether they simply recycle general processing heuristics or biases applicable to other structures (e.g. standard RCs)...**

Thank you !

Selected references i

Schwarze, C. (1974). Les constructions du type "je le vois qui arrive". In C. Rohrer & N. Ruwet (Eds.), *Actes du colloque franco-allemand de grammaire transformationnelle* (pp. 18–30). Niemeyer.

Kayne, R. S. (1975). *French syntax. the transformational cycle.*. MIT Press.

Radford, A. (1975). Pseudo-relatives and the unity of subject raising. *Archivum Linguisticum*.

Graffi, G. (1980). Su alcune costruzioni "pseudorelative". *Rivista di Grammatica Generativa*, *5*, 117–139.

Guasti, M. T. (1988). La pseudorelative et les phénomènes d'accord. *Rivista di Grammatica Generativa*, *13*, 35–80.

Rizzi, L. (1992). Direct perception, government and thematic sharing. *Geneva Generative Papers*, *1*, 39–52.

Rafel, J. (1999). *Complex small clauses* (Doctoral dissertation). Universitàt Autonoma de Barcelona.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. *Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001 - NAACL '01*. https://doi.org/10.3115/1073336.1073357

Selected references ii

Desmet, T., de Baecke, C., & Brysbaert, M. (2002). The influence of referential discourse context on modifier attachment in dutch. *Memory & Cognition*, *30*(1), 150–157. https://doi.org/10.3758/bf03195274

Lovrić, N. (2003). *Implicit prosody in silent reading: Relative clause attachment in croatian*.

Papadopoulou, D., & Clahsen, H. (2003). Parsing strategies in l1 and l2 sentence processing. *Studies in Second Language Acquisition*, *25*(4), 501–528. https://doi.org/10.1017/s0272263103000214

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. *9th Python in Science Conference*.

Casalicchio, J. (2013). *Pseudorelative, gerundi e infiniti nelle varietà romanze. affinità (solo) superficiali e corrispondenze strutturali* (Doctoral dissertation). Università di Padova.

Grillo, N., & Costa, J. (2014). A novel argument for the universality of parsing principles. *Cognition*, *133*(1), 156–187. https://doi.org/10.1016/j.cognition.2014.05.019

Moulton, K., & Grillo, N. (2015). Pseudo relatives: Big and direct. In T. Bui & D. Özyıldız (Eds.), *Proceedings of 45 north eastern linguistic society* (pp. 193–202). MIT.

Selected references iii

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR, abs/1810.04805*. http://arxiv.org/abs/1810.04805

Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018). What do RNN language models learn about filler–gap dependencies? *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 211–221. https://doi.org/10.18653/v1/W18-5423

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *CoRR, abs/1911.02116*. http://arxiv.org/abs/1911.02116

Lample, G., & Conneau, A. (2019). Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Pozniak, C., Hemforth, B., Haendler, Y., Santi, A., & Grillo, N. (2019). Seeing events vs. entities: The processing advantage of pseudo relatives over relative clauses. *Journal of Memory and Language*, *107*, 128–151. https://doi.org/10.1016/j.jml.2019.04.001

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.

Selected references iv

Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR, abs/1906.08237*. http://arxiv.org/abs/1906.08237

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., & Schwab, D. (2020). Flaubert: Unsupervised language model pre-training for french. *Proceedings of The 12th Language Resources and Evaluation Conference*, 2479–2490. https://www.aclweb.org/anthology/2020.lrec-1.302

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, É. V., Seddah, D., & Sagot, B. (2020). Camembert: A tasty french language model. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

ClassCat AI Lab. (2022). Gpt2-base-french model card on huggingface [Accessed: 2023-05-09].

Laurer, M., Atteveldt, W. v., Casas, A. S., & Welbers, K. (2022). Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT - NLI [Publisher: Open Science Framework]. *Preprint*. Retrieved July 28, 2022, from https://osf.io/74b8k

Lstms can learn basic wh- and relative clause dependencies in norwegian. (2022). *44th Annual Meeting of the Cognitive Science Society: Cognitive Diversity, CogSci 2022*.

Selected references v

Minixhofer, B., Paischer, F., & Rekabsaz, N. (2022). WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3992–4006. https://aclanthology.org/2022.naacl-main.293

Misra, K. (2022). Minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

Özsoy, T. (2022). Xlm-roberta-large-xnli-finetuned-mnli model card on huggingface [Accessed: 2023-05-09].

Yuan, M. (2022). A pseudo-relative in inuit. *Proceedings of The 40th meeting of the West Coast Conference on Formal Linguistics (WCCFL 40)*.

Doyen, B. (2023). Camembert-base-xnli model card on huggingface [Accessed: 2023-05-09].

Appendices

Illustrating the comparative rareness of unambiguous PRs in corpora

(13) a. "Il V * qui"
He V wildcard that
b. "Il {le, la, l'} V qui"
He CL V that

regular expression → V↓	(13a	(13b)	$\frac{\#(13b)}{\#(13a)+\#(13b)}$
voir ('see')	15157	168	1.1e-2
apercevoir ('notice')	725	1	1.4e-3
regarder ('look at')	2442	28	1.1e-2
observer ('watch')	813	0	0.0
épier ('spy on')	13	0	0.0
surprendre ('catch')	99	0	0.0
entendre ('hear')	1975	27	1.3e-2
écouter ('listen to')	632	1	1.6e-3

Table 8: Number of matches for non-cliticized (ambiguous) and cliticized (unambiguous) regular expressions on 10,160,000 documents from the OSCAR corpus (containing a total of 52,037,098 documents).

Example stimuli for Experiment 2

(14)
a. Il voit Marie qui embrasse Jean.
He sees Marie who kisses Jean.

b. Il voit Jean que Marie embrasse.
He sees Jean who Marie kisses.

c. Il la voit qui embrasse Jean.
He her.CL sees who kisses Jean.

d. * Il le voit que Marie embrasse.
He him.CL sees that Marie kisses.

e. * Il pense Marie qui embrasse Jean.
He thinks Marie who kisses Jean.

f. * Il pense Jean que Marie embrasse.
He thinks Jean that Marie kisses.

g. * Il la pense qui embrasse Jean.
He her.CL thinks who kisses Jean.

h. * Il le pense que Marie embrasse.
He him.CL thinks that Marie kisses.

Sentence	clitic?	gap	verb_type
(14a)	n	S	perception
(14b)	n	O	perception
(14c)	y	S	perception
(14d)	y	O	perception
(14e)	n	S	attitude
(14f)	n	O	attitude
(14g)	y	S	attitude
(14h)	y	O	attitude

Table 9: Summary of the $2{\times}2{\times}2$ design of Experiment 2

Example stimuli for Experiment 3

(15) a. Il ne voit pas Marie qui danse.
He NEG sees NEG Marie that dances.
$\Longrightarrow$ Marie is dancing. EC ✓

b. Il ne la voit pas qui danse.
He NEG CL sees NEG that dances.
$\Longrightarrow$ She is dancing. EC ✓

c. Il ne voit pas Marie danser.
He NEG sees NEG Marie dancing.
$\not\Longrightarrow$ Marie is dancing. EC ✗

d. Il ne la voit pas danser.
He NEG CL sees NEG dancing.
$\not\Longrightarrow$ She is dancing. EC ✗

Sentence	clitic?	emb_clause
(15a)	n	relative
(15b)	y	relative
(15c)	n	infinitive
(15d)	y	infinitive

Table 10: Summary of the 2×2 design of Experiment 3