Do Language Models discriminate between relatives and pseudorelatives?

Adèle Hénot-Mortier
Department of Linguistics & Philosophy, Massachusetts Institute of Technology
32 Vassar Street, Cambridge, MA 02139, USA
mortier@mit.edu

Abstract

Large Language Models (LLMs) are often evaluated against massive benchmarks based on general-purpose tasks, which, despite being useful for concrete applications, tell us very little about the capacity of LLMs to learn *specific* and *challenging* aspects of the grammar. Here, we evaluate whether LLMs learn to identify a particular structure attested in Romance (and French in particular), called the pseudorelative. This structure, which is often surface-similar to a relative clause, is linked to robust syntactic and semantic restrictions. We present a series of experiments to test if LLMs pretrained on massive yet general corpora, manage to learn those various restrictions. Our results suggest that LLMs learn some but not all of these properties, but crucially fail at recognizing the most specific of them: cliticization.

1 Background on pseudorelatives

Pseudorelatives (**PR**s) (Schwarze (1974); Radford (1975); Kayne (1975); Guasti (1988) a.o.) resemble relative clauses (**RC**s) but exhibit a specific cluster of properties: (1) their head noun can be cliticized; (2) they only feature subject gaps; (3) they only appear below perception verbs; (4) they require the matrix and embedded tenses to match; (5) they imply the existence/truth of the embedded event even under matrix negation (Moulton and Grillo, 2015). Those various properties are illustrated below.

(1) Head noun cliticization
Jean **la** voit qui sourit.
Jean **3.SG.CL** sees that smiles.
'Jean sees her smiling.'

(2) Object gap (+cliticization)
* Jean la voit que Marc salue _.
Jean 3.SG.CL sees that Marc greets.
Intended: 'Jean sees Marc greeting her.'

(3) Non-perception verb (+cliticization)
* Jean la **pense** qui sourit.
Jean 3.SG.CL **thinks** that smiles.
Intended: 'Jean thinks she is smiling.'

(4) Tense mismatch (+cliticization)
* Jean la voit qui souri**ait**.
Jean 3.SG.CL sees.PRS that smiled.**PST**.
Intended: 'Jean sees her while she smiled.'

(5) Event presupposition
Jean ne la voit pas qui sourit.
Jean NEG 3.SG.CL sees NEG that smiles.
'Jean doesn't see her smiling (she **does**).'

Cliticization is perhaps the most robust diagnostic used to disambiguate PRs from RCs; without a cliticized head, and assuming conditions (2)-(5) are met, a PR will usually remain ambiguous with a (string-identical) RC. This is shown in (6).

(6) No cliticization: ambiguous parse
Jean voit Marie qui sourit.
Jean sees Marie that smiles.
'Jean sees Marie, who smiles.' (**relative**).
'Jean sees Marie smiling.' (**pseudorelative**)

Because of is rareness in corpora, its ambiguity with relative clauses, and the inability of LLMs to access external disambiguating cues (comma, intonation), the pseudorelative remains relatively opaque to current NLP benchmarks (Wang et al., 2018, 2019; Bowman et al., 2015; Williams et al., 2018; Rajpurkar et al., 2016, 2018; Zellers et al., 2018) which are used to assess LLMs' performances. Do LLMs pretrained on massive (but also very general, non-targeted, and therefore impoverished) corpora learn something about pseudorelatives anyway?

2 Preliminary corpus study

We run a corpus study to verify the claim that LLMs are mostly exposed to structurally ambiguous sentences such as (6). We start with simple exact Google queries following the patterns in (7), where V denotes one of the verbs listed in Table 1, and CL is a clitic pronoun (*le* or *la* if V starts with a consonant, *l'* if V starts with a vowel).

(7) a. "Il V * qui"
He V wildcard that

b. "Il {le, la, l'} V qui"
He CL V that

The number of hits for these queries are gathered in Table 1. If some perception verbs are clearly more frequent than others (compare *voir*, 'see' vs. *épier*, 'spy on'), the tendency regarding cliticized constructions is clear: they are between 10,000 and 100,000 times less frequent than the string-ambiguous structures similar to (6).

exact query → V↓	(7a)	(7b)	$\frac{\#(7b)}{\#(7a)+\#(7b)}$
voit (see)	262,000,000	22,440	8.5e-5
apercoit (spot)	11,700,000	1,230	1.1e-4
regarde (look at)	192,000,000	6,370	3.3e-5
observe (watch)	51,100,000	759	1.5e-5
épie (spy on)	237,000	1	4.2e-6
surprend (catch)	21,900,000	247	1.1e-5
entend (hear)	70,200,000	7,820	1.1e-4
écoute (listen to)	121,000,000	18,200	1.5e-4

Table 1: Number of results for non-cliticized (ambiguous) and cliticized (unambiguous) PR structures returned by Google Search for different perception verbs.

To confirm this intuition, we matched a series of regular expressions[1] against a subset of the French OSCAR corpus (Ortiz Suárez et al., 2019; Caswell et al., 2021; Abadji et al., 2021), used to train models such as CamemBERT (Martin et al., 2020). The results shown in Table 2 confirm that a typical French LLM is mostly trained on ambiguous PR structures. Learning properties (1)-(5) would therefore require the models to exploit weak signals in the data to draw syntactic and semantic generalizations. The experiments that follow test whether LLMs achieve this goal – or not.

3 Experiment 1

Adapting a recent psycholinguistic experiment (Pozniak et al., 2019), we test if 8 LLMs trained on general French corpora (see Tab. 3 rows 1-8), learned the association between properties (3)-(4), pertaining to the type of the embedding verb and tense anaphoricity. The expected effects are: a

[1]The regular expressions were refined from the templates in (7) to include all possible subject pronouns and allowed up to 3 unspecified words in the wildcard (*). This restricts the search space for ambiguous relative constructions of the from of (6) but ensures that other constructions (such as an unambiguous relative clause located "far away" from the perception verb) are not matched by accident. It also allows to speed-up the search. Consequently, the matches and the proportions gathered respectively in the second and last columns of Table 2 should be respectively read as lower- and upper-bounds.

regular expression → V↓	(7a)'	(7b)'	$\frac{\#(7b)'}{\#(7a)'+\#(7b)'}$
voir	15157	168	1.1e-2
apercevoir	725	1	1.4e-3
regarder	2442	28	1.1e-2
observer	813	0	0.0
épier	13	0	0.0
surprendre	99	0	0.0
entendre	1975	27	1.3e-2
écouter	632	1	1.6e-3

Table 2: 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).

preference for embedding of (pseudo)relatives under perception verbs (as opposed to e.g. stative verbs); a preference for matching tenses between the matrix clause and the embedded clause; an interaction between those two factors, favoring tense matching specifically under perception verbs. We take the interaction to be the most critical effect. These predictions were assessed reusing the 2×2 design (`verb_type`×`tense_match`) introduced by the original study. Example stimuli illustrating this design are given in (8) and their parameters are summarized in Table 4.

ID	Model	Lang.	Reference
1	flaubert_base_uncased	fr	Le et al. (2020)
2	camembert-base	fr	Martin et al. (2020)
3	gpt2-base-french	fr	(Cla)
4	gpt2-wechsel-french	fr	Minixhofer et al. (2022)
5	bert-base-multi-lingual-cased	multi	Devlin et al. (2018)
6	xlm-roberta-base	multi	Conneau et al. (2019)
7	xlm-roberta-large	multi	Conneau et al. (2019)
8	xlm-mlm-17-1280	multi	Lample and Conneau (2019)
9	bert-large-cased	en	Devlin et al. (2018)
10	gpt2-large	en	Radford et al. (2019)
11	xlnet-large-cased	en	Yang et al. (2019)

Table 3: Models used in Exp. 1 and 2

(8) Example stimuli reused from (Pozniak et al., 2019).

a. Marie a écouté le ministre qui critiquait le président.
b. ?Marie écoute le ministre qui critiquait le président.
c. Marie a été mariée au ministre qui critiquait le président.
d. Marie est mariée au ministre qui critiquait le président.

Sentence	`verb_type`	`tense_match`
(a.)	perception	y
(b.)	perception	n
(c.)	stative	y
(d.)	stative	n

Table 4: Summary of the 2×2 design of (Pozniak et al., 2019) reused in Exp. 1

Building on (Hale, 2001; Levy, 2008), our proxy for grammaticality was taken to be the log-probability assigned to a given sentence by the

LLM (see equations below). It was computed using the `minicons` library (Misra, 2022).

$$\begin{aligned}\textsc{Grammaticality}(w_t) &\simeq -\textsc{Surprisal}(w_t)\\ &= \log P(w_t|w_1 \dots w_{t-1})\end{aligned}$$

[2]

$$\textsc{Grammaticality}(w_1 \dots w_t) \simeq -\sum_{i=1}^{t} \textsc{Surprisal}(w_i)$$

Linear mixed-effect modeling (performed with `statsmodels`, (Seabold and Perktold, 2010)) reveals that 6/8 LMs favor matching tenses, and 4/8 more so under perception verbs (`verb*tense` interaction) – supporting the expected interaction between (3) and (4) in French. Among the best performing models are a French-only (autoregressive) GPT-2 model (model 3) and a (bidirectional) multilingual RoBERTa model (model 7).

ID	best AIC?	verb_type	tense	interaction
1	n	. ✗	n.s.	. ✓
2	n	. ✓	** ✓	n.s.
3	y	n.s.	** ✓	* ✓
4	y	n.s.	** ✓	. ✓
5	n	n.s.	** ✓	n.s.
6	y	n.s.	** ✓	. ✓
7	y	n.s.	** ✓	* ✓
8	n	** ✓	n.s.	n.s.

Table 5: Significance results of LME modeling for `grammaticality` $\sim$ `verb_type+tense+verb_type*tense` $+ (1|$`frame`$)$, where `frame` refers to the lexical skeleton shared by all stimuli in e.g. (8).[3]

English models (cf Tab. 3, rows 9-11) tested on English equivalents of the stimuli exemplified in (8), did not exhibit similar effects – consistent with English not allowing pseudorelatives. Plots of the distributions of grammaticality scores obtained with `xlm-roberta-large` (model 7) in both languages are given in Figure 1.

4 Experiment 2

We test the same LLMs on 4800 semi-automatically generated sentences following the template in (9) and differing in (1) head noun cliticization; (2) the gap's position (subject/object) and (3) the matrix verb's type (perception vs. attitude/action).

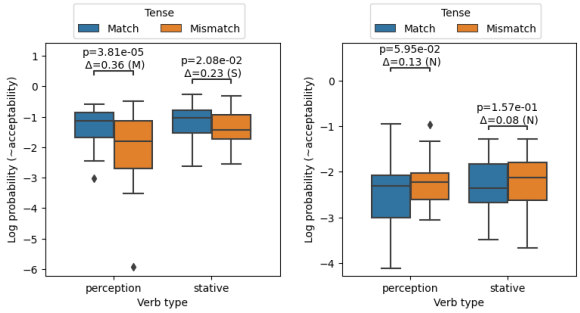
[2]In the case of BERT-like bidirectional models, this formula is adapted to masked language modeling: the probability of a word is computed given its left *and* right context.

[3]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: '.' = $p \in]0.05; .1]$, '*'= $p \in]0.01; .05]$; '**'= $p \in [0; .01]$; ✓=coefficient validates the hypothesis; ✗=coefficient disproves the hypothesis.

[4]The scores are overall negative because they correspond to negative log probabilities (cf. equations above).

ID	best AIC?	verb type	tense	interaction
5	y	** ✓	* ✗	* ✗
6	n	. ✓	. ✗	. ✗
7	n	n.s.	* ✗	n.s.
8	n	** ✓	n.s.	n.s.
9	n	n.s.	n.s.	. ✓
10	n	** ✓	* ✓	n.s.
11	n	n.s.	n.s.	n.s.

Table 6: Significance results of LME modeling with English data. Same notations and parameters as Table 5. Strikingly, all but 1 model did not yield the best AIC for the formula involving an interaction term.



(a) French (b) English

Figure 1: Distributions of the grammaticality scores[4] for Exp. 1 with `xlm-roberta-large`. Δ refers to Cliff's Delta (non-parametric measure of effect size). N, S, M resp. mean 'negligible', 'small', 'medium'.

(9) Template for the stimuli of Exp. 2

$$\begin{array}{ccccccc}\left\{\begin{array}{c}\text{Il/Elle}\end{array}\right\} & - & \left\{\begin{array}{c}\text{le/la/l'}\\ \emptyset\end{array}\right\} & - & \left\{\begin{array}{c}\text{voit/...}\\ \text{pense/...}\end{array}\right\} & - & \\ \text{PRO} & - & \text{(CL)} & - & \text{V} & - & \\ \left\{\begin{array}{c}\emptyset\\ \text{Marie/Jean}\end{array}\right\} & - & \left\{\begin{array}{c}\text{subject-gap relative}\\ \text{object-gap relative}\end{array}\right\} & & & & \\ \text{(N)} & - & \text{CP} & & & & \end{array}$$

(10) Example stimuli for Exp. 2

a. Il voit Marie qui embrasse Jean.
b. Il voit Jean que Marie embrasse.
c. Il la voit qui embrasse Jean.
d. * Il le voit que Marie embrasse.
e. * Il pense Marie qui embrasse Jean.
f. * Il pense Jean que Marie embrasse.
g. * Il la pense qui embrasse Jean.
h. * Il le pense que Marie embrasse.

Given this design, we expect an overall preference for matrix perception verbs, subject gaps and non-cliticized constructions, but also a positive interaction between perception verbs and clitics, perception verbs and subject gaps, clitics and subject gaps, and all three variables together. As Tab.

Sentence	clitic?	gap	verb_type
(10a)	n	S	perception
(10b)	n	O	perception
(10c)	y	S	perception
(10d)	y	O	perception
(10e)	n	S	attitude
(10f)	n	O	attitude
(10g)	y	S	attitude
(10h)	y	O	attitude

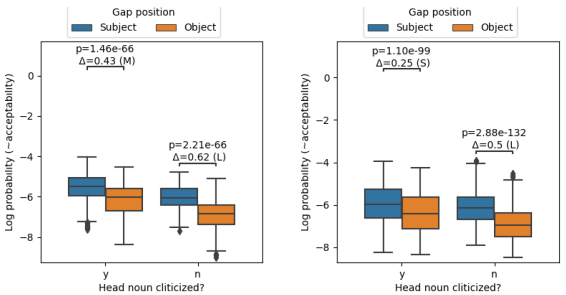
Table 7: Summary of the 2×2×2 design of Exp. 2

8 shows, linear mixed-effect modeling reveals a robust preference for subject-gaps (8/8 models, cf. col. 3) and more so under perception verbs (5/8 models, cf. col. 6), supporting (2)+(3). The desired clitic*gap*verb_type interaction however, was only captured by 1/8 models (cf. col. 8). Strikingly also, the interaction between cliticization and subject gaps is predicted by most models to have a negative effect on grammaticality, *contra* (1)+(2).

ID	v	g	c	v*c	v*g	c*g	v*c*g
1	. ✓	** ✓	** ✓	** ✗	. ✗	** ✓	n.s.
2	. ✓	** ✓	** ✓	** ✗	** ✓	** ✗	n.s.
3	n.s.	** ✓	** ✗	** ✓	** ✓	** ✗	. ✓
4	n.s.	** ✓	** ✗	** ✗	** ✗	** ✗	** ✓
5	n.s.	** ✓	** ✗	** ✓	n.s.	** ✗	n.s.
6	n.s.	** ✓	** ✗	** ✓	** ✓	** ✗	. ✗
7	n.s.	** ✓	** ✗	** ✓	** ✓	** ✗	** ✗
8	n.s.	** ✓	* ✗	** ✗	** ✓	**	n.s.

Table 8: Significance results of LME modeling for grammaticality ∼ verb_type + gap + clitic + verb_type ∗ clitic ∗ gap. Same notations as before.

The best performing model for this experiment appears to be a French-only GPT-2 model (model 3) – which was also among the best models for Exp. 1. Grammaticality scores corresponding to this model are plotted in Fig. 4.



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

Figure 2: Distributions of the grammaticality scores for Exp. 2 with gpt2-base-french. Same notations as before.

5 Experiment 3

We finally test property (5) on 4 BERT-like LMs fine-tuned to perform natural language inference (see Table 9).

ID	Model	Lang.	Reference
12	camembert-base-xnli	fr	(Doy)
13	xlm-roberta-large-xnli-finetuned-mnli	multi	(Ozs)
14	mDeBERTa-v3-base-mnli-xnli	multi	(Laurer et al., 2022)
15	mDeBERTa-v3-base-xnli-multilingual-nli-2mil7	multi	(Laurer et al., 2022)

Table 9: Models used in Exp. 3

Given a negated matrix perception verb embedding a clause $\mathcal{C}$ either as an infinitive or as a (pseudo)relative, with or without cliticization of its subject (2×2 design, see (11)), we measure how likely LLMs are to infer the truth of $\mathcal{C}$ ("target inference", **TI**).

(11) Template for the stimuli of Exp. 3

$$\left\{\begin{array}{c}\text{Il/Elle}\\ \end{array}\right\}\ \text{ne}\ \left\{\begin{array}{c}\text{le/la/l'}\\ \emptyset\end{array}\right\}\left\{\begin{array}{c}\text{voit/...}\\ \end{array}\right\}\text{pas}\ -$$

PRO NEG (CL) V NEG–

$$\left\{\begin{array}{c}\emptyset\\ \text{Marie/Jean}\end{array}\right\}\ -\ \left\{\begin{array}{c}\text{subject-gap relative}\\ \text{subject-gap infinitive}\end{array}\right\}$$

(N) – CP

(12) Example stimuli for Exp. 2

a. Il ne voit pas Marie qui danse.
He NEG sees NEG Marie that dances.
⟹ Marie is dancing. TI ✓

b. Il ne la voit pas qui danse.
He NEG CL sees NEG that dances.
⟹ She is dancing. TI ✓

c. Il ne voit pas Marie danser.
He NEG sees NEG Marie dancing.
⟹̸ Marie is dancing. TI ✗

d. Il ne la voit pas danser.
He NEG CL sees NEG dancing.
⟹̸ She is dancing. TI ✗

Sentence	clitic?	emb_clause
(12a)	n	relative
(12b)	y	relative
(12c)	n	infinitive
(12d)	y	infinitive

Table 10: Summary of the 2×2 design of Exp. 3

We expect the TI to be overall stronger when the embedded clause is a relative as opposed to an infinitive, whether of not the head noun is cliticized. The effect of cliticization in the case of a structure embedding a relative is a little bit

less clear: in the absence of cliticization the clause is ambiguous between a PR and a RC, and it is reasonable to think that both parses encourage a TI. Assuming that the RC parse imposes a somewhat stronger TI than a PR parse, then we might expect sentences like (12b), which are unambiguously PRs due to cliticization, to lead to a slightly weaker TI than sentences like (12a) which allow a RC parse. In other words, we expect non-cliticized sentences embedding an RC to yield the strongest TI.

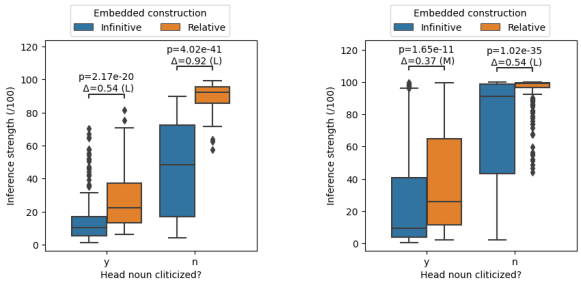
Linear mixed-effect modeling reveals that embedded relative constructions systematically lead to a stronger target inference as opposed to infinitives (cf. Table 11 col. 3), which is *consistent* with property (5), might be driven by the RC-parse only. Non-cliticized subjects also lead to a stronger target inference across the board (col. 4). This is made particularly clear in Figure 3. This pattern cannot be fully explained by the theory but makes sense if we consider that non-cliticized constructions are way more frequent in the data (so that LLMs may be more confident about the inferences related to such constructions, as opposed to cliticized ones). Finally, 2⁄4 models associate non-cliticized RC-embedding constructions to a stronger TI, which corresponds to the stipulation discussed in the previous paragraph. This all suggests that LLMs associate the target inference with the occurrence of RCs, but not really PRs: otherwise, *cliticized* relative constructions (unambiguously PRs) would have lead to stronger target inferences. Figure 3 in particular, shows that cliticized constructions featuring an embedded relative (unambiguously PRs), do not lead at all to a strong TI, suggesting the RC-parse (and not the PR-parse), is driving this inference.

ID	best AIC?	embedded clause (RC)	clitic	RC/clitic interaction
12	y	** (+)	** (-)	** (-)
13	y	** (+)	** (-)	** (+)
14	y	** (+)	** (-)	** (+)
15	y	** (+)	** (-)	** (-)

Table 11: Significance results of LME modeling for `target_inference_strength` $\sim$ `emb_clause` + `clitic` + `emb_clause` $*$ `clitic`.

6 Discussion and outlook

In this work, we investigated a structure (the pseudorelative) with two interesting distributional properties: (1) it can be ambiguous with a relative clause when the head noun is *not* cliticized; (2) the disambiguating (cliticized) structure is less frequent in corpora by several orders of magnitude. We think that the conjunction of these two properties makes learning the specific syntactic and semantic properties of PRs particularly challenging, even for models trained on large amount of data.



(a) CamemBERT (b) mDeBERTa

Figure 3: Distributions of the TI strength scores (/100) for Exp. 3 and models 12 and 15.

The experiments we run show that LLMs capture certain properties of PRs, pertaining to acceptable filler-gap dependencies, matrix verbs, and tense combinations. Interestingly, 3⁄4 multilingual models exposed to both French (a PR-language) and English (devoid of PRs) in Exp. 1 managed to contrast the two languages. Yet, the property that is perhaps the most specific to pseudorelatives, cliticization, does not seem to influence sentence probability scores in Exp. 2, and inference patterns in Exp. 3. This raises the question whether LLMs really get the specificity of the pseudorelative as a *syntactic construction* (Exp. 2) with a specific *semantics* (Exp. 3); or whether they simply recycle general processing heuristics applicable to other structures (e.g. standard RCs). Such heuristics may involve a preference for shorter dependencies (subject-gaps) across the board; or learning a statistical correlation between the use of perception verbs and the *agentive* structure of the perceived event.
Future work may involve investigating other languages allowing the pseudorelative, but also refining the current design by looking at the influence of the different perception verbs. We think this might be particularly relevant given the rather large frequency differences between these verbs in actual corpora (cf. Tables 1 and 2), and the potential imbalance between ambiguous vs. unambiguous PR-structures for each of those verbs.

References

camembert-base-xnli model card on huggingface. https://huggingface.co/BaptisteDoyen/camembert-base-xnli. Accessed: 2023-05-09.

gpt2-base-french model card on huggingface. https://huggingface.co/ClassCat/gpt2-base-french. Accessed: 2023-05-09.

xlm-roberta-large-xnli-finetuned-mnli model card on huggingface. https://huggingface.co/tuni/xlm-roberta-large-xnli-finetuned-mnli. Accessed: 2023-05-09.

Julien Abadji, Pedro Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2022. Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. *arXiv e-prints*, page arXiv:2201.06642.

Julien Abadji, Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2021. Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9) 2021. Limerick, 12 July 2021 (Online-Event), pages 1 – 9, Mannheim. Leibniz-Institut für Deutsche Sprache.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Isaac Caswell, Julia Kreutzer, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Javier Ortiz Suárez, Iroro Orife, Kelechi Ogueji, Rubungo Andre Niyongabo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2021. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *arXiv e-prints*, page arXiv:2103.12028.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Aniello De Santo and So Young Lee. 2022. Evaluating structural economy claims in relative clause attachment. In *Proceedings of the Society for Computation in Linguistics 2022*, pages 65–75, online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001 - NAACL '01*. Association for Computational Linguistics.

Richard S. Kayne. 1975. *French syntax. The transformational cycle.* MIT Press, Cambridge (MA).

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Moritz Laurer, Wouter van Atteveldt, Andreu Salleras Casas, and Kasper Welbers. 2022. Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT - NLI. *Preprint*. Publisher: Open Science Framework.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. Flaubert: Unsupervised language model pre-training for french. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.

Kanishka Misra. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

Keir Moulton and Nino Grillo. 2015. Pseudo relatives: Big and direct. In *Proceedings of 45 North Eastern Linguistic Society*, pages 193–202. MIT.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoit Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Benoit Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut f"ur Deutsche Sprache.

Céline Pozniak, Barbara Hemforth, Yair Haendler, Andrea Santi, and Nino Grillo. 2019. Seeing events vs. entities: The processing advantage of pseudo relatives over relative clauses. *Journal of Memory and Language*, 107:128–151.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Andrew Radford. 1975. Pseudo-relatives and the unity of subject raising. *Archivum Linguisticum*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250.

C Schwarze. 1974. 'les constructions du type "je le vois qui arrive". In *Actes du Colloque Franco-Allemand de Grammaire Transformationnelle*, pages 18–30, Tübingen. Niemeyer.

Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *CoRR*, abs/1905.00537.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *CoRR*, abs/1804.07461.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.