Assessing the Capacity of Transformer to Abstract Syntactic Representations: A Contrastive Analysis Based on Long-distance Agreement

Bingzhi Li and **Guillaume Wisniewski** and **Benoît Crabbé**
Université Paris Cité, LLF, CNRS, 75 013 Paris, France
bingzhi.li@etu.u-paris.fr
{guillaume.wisniewski,benoit.crabbe}@u-paris.fr

Abstract

Many studies have shown that transformers are able to predict subject-verb agreement, demonstrating their ability to uncover an abstract representation of the sentence in an unsupervised way. Recently, Li et al. (2021) found that transformers were also able to predict the object-past participle agreement in French, the modeling of which in formal grammar is fundamentally different from that of subject-verb agreement and relies on a movement and an anaphora resolution.

To better understand transformers' internal working, we propose to contrast how they handle these two kinds of agreement. Using probing and counterfactual analysis methods, our experiments on French agreements show that *(i)* the agreement task suffers from several confounders that partially question the conclusions drawn so far and *(ii)* transformers handle subject-verb and object-past participle agreements in a way that is consistent with their modeling in theoretical linguistics.

1 Introduction

Since Linzen et al. (2016), the long-distance agreement task has been a paradigmatic test for assessing the ability of Neural Network Language Models (NNLMs) to uncover syntactic information from raw texts: A model able to predict the correct verb form (especially when the verb does not agree with the word just before it) has to, to some extent, acquire a representation of the syntactic structure and encode it in its internal representations.

In this work, we seek to identify to what extent the NNLM abstracts its representations from surface pattern recognition to structurally motivated representations. To do this, we focus on two kinds of number agreement in French (both morphologically marked):

(1) Les **chat·s** [que Noûr aime bien]$_{RC}$ **jou·ent** dans le jardin.
The_Pl cats_**Pl** [that Noûr likes_Sg a_lot]$_{RC}$ play_**Pl** in the garden.

(2) Il aime les **chat·s** [que Noûr a **adopté·s**]$_{RC}$.
He_Sg loves_Sg the_Pl cats_**Pl** [that Noûr has_Sg adopted_**Pl**]$_{RC}$

These sentences involve two agreements: The first one between a noun *chats* and the main verb *jouent* and the second one between the same noun and the past participle *adoptés*.[1] A naive look at (1) and (2) may suggest that they are two identical agreements between a noun and a verbal form separated by a few words. Yet from a linguistic perspective these two agreements receive a substantially different analysis. As in English, example (1) involves a number agreement between the main clause verb and its subject where an embedded relative clause occurs between the two. To predict this kind of agreement, the model has to learn an abstract representation that is not solely based on the linear sequence of words but also on the long syntactic dependency between the verb and its subject, in order to ignore the linear proximity of the noun *Noûr*.

[1]The two agreements could occur in the same sentence, as in the example of Figure 1.

Transactions of the Association for Computational Linguistics, vol. 11, pp. 18–33, 2023. https://doi.org/10.1162/tacl_a_00531
Action Editor: Xavier Carreras. Submission batch: 4/2022; Revision batch: 8/2022; Published 1/2023.

On the other hand, the agreement between *chats* and *adoptés* in (2) is an object relative clause agreement. Overall, it involves an agreement between a noun in the main clause and a past participle in the relative clause. To predict this kind of agreement, the model must be able to detect a complex set of patterns across different clauses that brings into play an anaphora resolution and a movement, a set of operations whose nature is fundamentally different from phrase structure embedding involved in the subject-verb agreement in (1).

While these two kinds of agreement show very similar surface patterns, their modelings in formal grammar result in completely different representations of the sentence structure, and it is not clear whether and how a sequential language model can identify these abstract representations based merely on the words sequence. Even though a tremendous number of studies have brought to light the ability of neural networks to capture and abstract the information needed to predict the subject-verb agreement (see Section 6 for an overview), it is only very recently that Li et al. (2021) showed that they were also able to predict the much rarer past participle agreement across clause boundaries like the agreement between *chats* and *adoptés* in (2).

Building on Li et al. (2021, 2022), the goal of the present work is to contrast how transformers handle these two kinds of agreement: We aim to determine whether they encode the *same* abstract structure in their internal representations to capture the information required to choose the correct verb form or, on the contrary, if the abstract structure they encode reflects the *distinction* made in the theoretical modeling of these two agreements. This contrast will shed a new light on our understanding of the internal working of transformers.

The contributions of this paper are twofold. First, we assess incremental transformers' syntactic ability to process two different syntax-sensitive dependencies with similar surface forms in French. Our results show that transformers perform consistently well on both agreement phenomena, and crucially, that they are able to abstract away from potential confounds such as lexical co-occurences or superficial heuristics. Second, we use linguistic probes as well as targeted masking intervention on self-attention to test **where** transformer-based models encode syntactic information in their internal representations and **how** they use it. We find that, for both constructions, even though the long-distance agreement information is mainly encoded locally across the tokens between the two elements involved in the agreement, transformers are able to leverage distinct linguistically motivated strategies to process these two phenomena.[2]

The rest of this paper is organized as follows: In Section 2, we define more precisely the two kinds of agreement considered in this work. Then, in Section 3, we contrast the capacity of transformers to capture subject-verb and object-past participle agreement. Following this, in Section 4, by using probing and counterfactual analysis, we examine the way transformers encode and use the syntactic information to process these two linguistic phenomena. We then discuss, in Section 5, the impact of some potential confounders, in Section 6 the related work, and finally we conclude in Section 7.

2 Long-distance Agreement

Overall, this work aims to assess to what extent transformers rely on shallow surface heuristics to predict agreement or to what extent they infer more abstract knowledge. To study this question, we rely on a contrast between two agreement tasks as exemplified in (1), (2), and in Figure 1. Both involve a long-distance agreement but they receive substantially different linguistic analyses.

Syntactic Phenomena In this work, we focus on corpus-extracted sentences involving *object relatives* such as the one analyzed in Figure 1. These kinds of sentences could involve two types of agreement, which receive substantially different linguistic analyses. The first one between the ''cue'' and ''target-v'' is an agreement between a verb and its subject separated by an object relative clause. As exemplified in Figure 1, the number of the main verb *resteront* (will_stay Plural) is determined by the head of the subject *moments*. In this first agreement, there is no relevant relationship between the relative pronoun *que* and the main verb. To correctly predict the number of the verb, a model must infer an abstract structural dependency across the relative clause to distinguish

[2]Datasets and code are available at `https://github.com/bingzhilee/contrastive_analysis`.

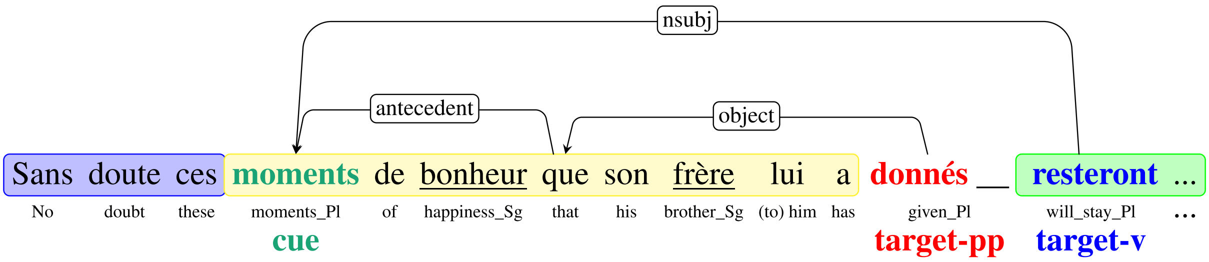

Figure 1: Example of object-past participle agreement and long-distance subject-verb agreement. To predict the target past participle (in red) number, a human is expected to get the feature from the object relative pronoun *(que)* that gets it from its antecedent (*moments* in bold green). The latter is also the grammatical subject of the main verb (*resteront* in bold blue) and determines its number. For the object-pp agreement, the prefix is highlighted in blue, the context in yellow, and the suffix in green.

the embedded subject (*frère*) from the main clause subject (*moments*). The model has to resist the lure of the linearly closer but irrelevant attractors *frère* and *bonheur* (attractors, underlined in Figure 1, are intervening nouns with misleading agreement features).

We also consider the agreement of the past participle in the relative clause: the past-participle (denoted ''target-pp'' in Figure 1) agrees with its complement (the ''cue'') if the latter moves before it. This agreement relies on an abstract set of relations between tokens occurring in different clauses. It involves an anaphora (designated by the `antecedent` arc) and a filler-gap dependency: The filler is *que* and the gap, indicated with a line in Figure 1, is an empty syntactic position licensed by the filler. The relative pronoun *que* is the pre-verbal direct object of the past participle *donnés* and triggers the agreement of the past participle. To obtain its agreement features, the relative pronoun has to be linked by anaphora to its nominal antecedent *moments*. In other words, to correctly agree the past participle in theory, it is necessary to identify the object relative pronoun and its antecedent. The model has also to ignore the effect of attractors occurring between the antecedent of the relative pronoun and the past participle.

Terminology In this paper, for both agreement phenomena, we refer to the noun item as the **cue**, and the verbal item as the **target** (see Figure 1). *pp* is short for past participle. We call the tokens before the cue the *prefix*, the tokens between the cue and the target the *context*, and the tokens after the target the *suffix*. We only consider number agreement as *(i)* number agreement is the only feature shared by the two agreements we consider,[3] *(ii)* the main purpose is to design reasonably simple patterns allowing us to extract a sufficiently large number of representative examples.

Datasets For object-past participle agreement, we consider the number agreement evaluation set introduced by Li et al. (2021): To study the ability of transformers to predict the object-past participle agreement, they have parsed automatically the French Gutenberg Corpus, and have extracted, with simple heuristics, a set of 68,497 French sentences (65% singular and 35% plural) involving an object relative. For the subject-verb agreement considered in this paper, we extracted from the same parsed corpus 27,582 sentences (70% singular 30% plural), in which the sequence between the subject and the verb contains at least one object relative clause. There are fewer items in subject-verb agreement evaluation set because noun phrases modified by relative clause(s) occur more frequently in the object position of the main clause as in example (2) than in the subject position like in example (1). In these two evaluation sets, an arbitrary number of words can occur between the *cue* and the *target*: On average, there are 5 tokens between the antecedent and the past participle, and 11 tokens between the subject and the verb. The intervening tokens include varied constructions such as prepositional phrases, participials or nested relative clauses, which would make the agreement tasks more

[3]In French, the verb has to agree in number with its subject, and the past participle conjugated with the auxiliary *avoir* agrees in number *and* in gender with its direct object if the latter appears before it.

Number of heuristics	Difficulty of agreement	Examples
5	−−−	Si les **idées** que ces mots représentent ne **sont** pas ... (5) *If the ideas*(4)(1) that these words(2) represent(3) are not...
4	−−	Les **choses** que nous avions vues cent fois avec indifférence nous **touchent**... (5)*The things*(4)(1) *that we had seen a hundred times with indifference us*(3) *touch*
3	−	Un philosophe est curieux de savoir si les **idées** qu' il a semées **auront**... *A philosopher is curious to know if the ideas*(2)(4) *that he has sown*(3) *have. . .*
2	+	Les **emblèmes** qu' on y rencontre à chaque pas **disent** ... *The emblems*(4)(1) *that we meet at each step say ...*
1	++	Les **qualités** qui t'ont fait arriver si jeune au grade que tu as **doivent** te porter ... *The qualities*(1) *that made you arrive so young at the rank you have must bring you ...*
0	+++	Ce soir les **hommes** que j'ai postés sur la route que doit suivre le roi **prendront** ... *Tonight the men that I have posted on the road that the king must follow will_take ...*

Table 1: Examples from our evaluation set of subject-verb agreement, stratified by the count of surface heuristics predicting the *target*'s number, a proxy to the task difficulty. The target verbs and their subjects are in bold. Orange numbers in parenthesis indicate the presence of different types of heuristics.

challenging. We also ensure that the two evaluation sets are completely separate from the training data of LMs.

3 Number Agreement Prediction

3.1 Experimental Setting

Following Linzen et al. (2016) and Gulordava et al. (2018), we use the number agreement task to evaluate incremental transformers' ability to capture syntactic information: Given a sentence, the model is fed with all tokens preceding the target verb (either the verb of the main clause or the past participle in the relative clause) and we compare the probabilities it assigned to the singular and plural variants of the target. The model's syntactic ability is measured by the percentage of sentences for which the verb form with the higher probability is the one that respects the agreement rules of the language (i.e., matches the number of the *cue*).

To measure the prediction difficulty of the evaluation sentences in the two agreement tasks, we follow Li et al. (2021) and define a scale that quantifies the agreement difficulty for a given sentence by counting the number of surface heuristics that predict the correct form of the target verb. We target five superficial heuristics, namely, (1) the *first noun* of the sentence, (2) the *closest noun*, (3) the *closest token* with a mark of number, (4) the *noun* before the closest *que*, and (5) the *majority number* expressed in the sequence preceding the *target*: As illustrated in Table 1, the correct form of the target verb in the *5 heuristics* subset (the ''easiest'' subset of our evaluation sets) can be easily predicted by simply memorizing any of the five surface heuristics (e.g., the target form should match the first noun in the sentence). On the contrary, the *0 heuristic* subset gathers sentences for which the target verb form can only be predicted by constructing an abstract representation of the sentence. We will mainly focus on the hardest cases (i.e., *0* and *1 heuristic* subsets) in our evaluation.

Models The experiments were carried out with two incremental language models designed by Li et al. (2021): an LSTM and an incremental transformer language model. Both models perform predictions incrementally using the conditional probability $P(w_i|w_1 \ldots w_{i-1}; \boldsymbol{\theta})$. The LSTM model has 2 layers and the transformer model has 16 layers and 16 heads. Word embeddings are of size 768. Note that the LSTM model has fewer parameters than the transformers and, consequently, a direct comparison of their performance is not completely fair. As training an LSTM with a comparable number of parameters is hardly computationally tractable and the LSTM language model has been shown to perform consistently well across varied agreement dependencies (Linzen et al., 2016; Gulordava et al., 2018; Marvin and Linzen, 2018), we only use it in this work as a strong baseline model and focus, in our analyses, on transformers. Hyperparameters were chosen by minimizing the

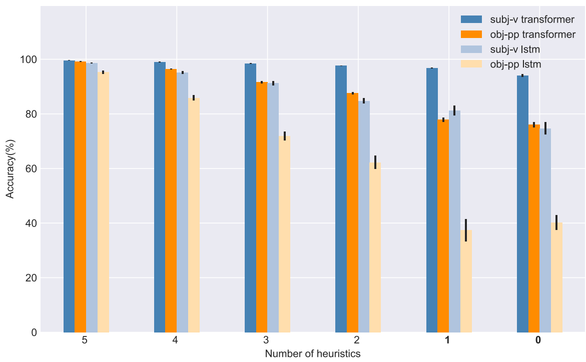

Figure 2: Accuracies achieved by transformers and LSTMs (indicated by lighter color bars) as a function of the agreement prediction difficulty. Blue bars represent the subject-verb agreement and orange bars for the object-pp agreement. The more heuristics are present, the easier the task is.

perplexity on the validation set and the best hyper-parameters were used to train for each architecture five models on the French Wikipedia corpus (90 million tokens). All the results reported in this paper are averaged across five models. More details about the models are given in Section B in the appendix.

3.2 Number Prediction Results

Overall Accuracy: The two architectures are able to predict both long-distance agreements with a very good accuracy:[4] transformers made correct number prediction in 94.6% of the object past participle agreement (LSTM 82.1%) and in 98.9% of the subject-verb agreement (LSTM 94.3%), a result consistent with the conclusions drawn by previous studies (Linzen and Baroni, 2021).

Surface Heuristics: However, detailed scores according to the difficulty of the task in Figure 2 show more nuanced results. With respect to the type of agreement, we observe that both LSTMs and transformers achieve much better performance for subject-verb agreement than for the object-past participle agreement, especially in the hardest cases (i.e., *0 & 1 heuristic* subsets), even though the linear distance between the *cue* and the *target* in the subject-verb dependency is twice as long as that in the object-past participle agreement (11 tokens vs. 5 tokens on average). This performance difference could result from the agreement frequency in the training data: The subject-verb agreement exists nearly in each sentence of the training data, while only 0.35% of the training sentences involve an object-past participle agreement.

However, we do find a similar pattern across these two agreement tasks: The models' performance always decreases with the task difficulty. This observation generalizes the conclusions of Li et al. (2021): It shows that the impact of surface heuristics is not limited to a relatively infrequent and complex kind of agreement (i.e., the object-past participle agreement) but also concerns the subject-verb agreement, further confirming that results on long-distance agreement tasks should be interpreted with great care.

It also appears that transformers outperform LSTM across the board even though, as explained above, the performance of these two models are not directly comparable: For the two types of agreements, transformers are able to predict the correct verb form most of the time, even in the hardest cases for which the performance of LSTM is below chance level. Above all, this comparison highlights the very strong capacity of transformers to capture syntactic information that even LSTM, a very good baseline on which are based most of the conclusions drawn so far on the syntactic capacity of neural networks, is unable to capture.

4 Do Transformers Process the Two Agreements in the Same Way?

In the last section, transformer language model has demonstrated a strong ability for capturing syntactic information that generalizes beyond surface heuristics when processing long-range subject-verb and object-past participle agreements. In this section, we investigate whether transformers acquire and use a meaningful, abstract representation of the sentence to predict the two theoretically distinct agreements we considered in this work or, on the contrary, whether they use one unified agreement mechanism to resolve the two agreement tasks. To do this, we use in Section 4.1 linguistic probes (Veldhoen et al., 2016; Conneau et al., 2018) to compare the distribution of the number information across the sentence token representations. Then, in Section 4.2, we compare the way transformers use the encoded syntactic information to process these two different long-range dependencies.

[4] Detailed scores are provided in table 7 in the appendix.

	Mean probing Accuracy	
	Object-pp	Subject-verb
prefix	$58.6\%_{\pm 0.1}$	$59.5\%_{\pm 0.2}$
context	$92.3\%_{\pm 0.2}$	$93.0\%_{\pm 0.1}$
suffix	$73.6\%_{\pm 0.2}$	$78.1\%_{\pm 0.2}$

Table 2: Probing task results across different sentence parts (see Figure 1). Reported scores are the average of all the PoS-based classifiers' accuracies for each part of the sentence.

4.1 Probing Contextualized Representations

We now investigate how transformers represent the syntactic information required to predict the correct verb form: Is it represented across all the tokens following the *cue* in the sentence as made theoretically possible by the self-attention mechanism and observed by Klafka and Ettinger (2020), or is it encoded mainly locally around the *cue* and the *target* tokens? To this end, we use linguistic probes. A probe is a classifier trained to predict linguistic properties (in our case: the number of the target verb) from the model representations (the representation of a token uncovered by the model). Intuitively, a probe that achieves high accuracy implies that these properties were encoded in the representation. To test the two hypotheses and find out whether number information is encoded only locally or globally in the whole sentence, we simply have to train probes for all words in different parts of the sentence.

More precisely, for each sentence of our evaluation set, we extract the token representations from the last layer of the transformer and associate them with a label describing the number of the target verb. We then train one logistic regression classifier[5] for each PoS tag of each part of the sentence (*prefix*, *context*, or *suffix* as defined in the ''Terminology'' paragraph of Section 2). We consider 80% of the examples as training data and use the remaining 20% to evaluate the probe's performance.

Results Table 2 reports the average accuracy achieved by our probes on different parts of the sentence. We observe a similar pattern for the object-pp and subj-verb agreement: The *target* number information is essentially encoded locally within the tokens of the *context* and is not represented uniformly across all the sentence tokens.

As expected, in the two cases, the performance of the probe on *prefix* is very low: Since we consider an incremental model, the tokens of the *prefix* cannot attend to the *cue* and their representation cannot encode information about its number. The accuracy mainly reflects the imbalance between singular and plural forms in our evaluation set.

In contrast, the accuracy becomes consistently high when the tokens of the *context* are considered as input features, while accuracy on the *suffix* tokens drops sharply even though it remains better than that observed in the *prefix*, suggesting that the information required to predict the correct *target* form in both dependencies is spread over all tokens between the *cue* (where the number of the *target* verb is specified) and the *target* (where the information is ''used'').

Representations Within the Context To get a more accurate picture of how the number information is distributed within the *context*, we focus on a specific sentence pattern: We only consider sentences for which the *cue* is separated from the relative pronoun only by a prepositional phrase. For the object-past participle agreement, the pattern thus corresponds to the sequence `Antecedent ADP N que PRON AUX`, as in:

(3) ... **bureau·x** en métal qu' il a **trouvé·s**
... desks Prep. metal that he has found_Pl...

For subject-verb agreement, it corresponds to the sequence `Subject ADP N que PRON V` as in:

(4) ... **bureaux** en métal qu' il aime
... desks Prep. metal that he loves
coût·ent ...
cost...

We trained one probing classifier for each position on 800 randomly sampled examples.[6]

[5]All classifiers were implemented with the `scikit-Learn` library (Pedregosa et al., 2011), setting the `max_iter` parameter to 1,000 and the `class_weight` to `balanced`.

[6]We used three sampling seeds and, for each sampling, three train/test splits. The sample is constrained to ensure balance between singulars and plurals in training and test sets.

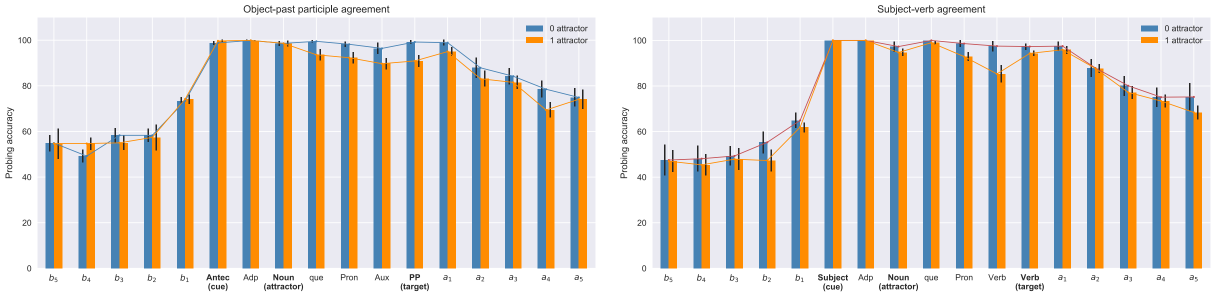

Figure 3: Probing accuracy at each position based on the number of the *target*. The `bI` (resp., `aI`) position denotes the I-th token before (resp., after) the pattern.

We then evaluated the probe on a balanced set of 200 examples, differentiating sentences in which the embedded noun is an attractor from sentences in which this noun has the same number as the *cue*.

Figure 3 reports the probing accuracy at each position. We observe that in the *prefix* (i.e., b-positions) the probe accuracy is low for the two kinds of agreement. On the contrary, in the *context*, the predictions of the probe are almost perfect, even when there is an attractor. It is quite remarkable that, in the *suffix*, once the *target* is encountered, the accuracy on the subsequent tokens drops quickly, especially in the presence of an attractor. These observations are consistent across the two agreement tasks we consider.

The results of the probing experiments show us that transformer language models encode the syntactic information in a very similar way across the two long-range agreements. In terms of acquired abstractions, nothing from the probing methodology allows to conclude that the model acquires substantially different representations for each agreement phenomenon.

4.2 Causal Intervention on Attention

However, probing has a well-known limitation (Belinkov and Glass, 2019): It only brings out a correlation between the representations and the syntactic information measured by the probe, and does not tell us whether and how this information is actually involved in processing the dependency. In this section, we aim to identify which tokens are actually responsible for providing the number information. To do so, we design a causal experiment in which we mask certain words preceding the *target* verb in the sentence to better determine their role in transformers predictions.

Masking Tokens in Self-attention Computation Transformers rely on a self-attention mechanism to build a contextualized representation for each token by iteratively defining (as a first approximation) the token representation as a linear combination of the representations of the other tokens in the sentence. We propose to neutralize the contribution of one or more specific tokens in the construction of the target verb's representation by forcing the weight of this or these tokens in this linear combination to be zero. This intervention can be implemented in a straightforward manner by extending the masking mechanism used in incremental transformers to prevent a token representation from taking into account future words.

More precisely, we consider the same NA tasks as in Section 3 but this time when predicting the target verb (and only at this moment!), we also mask either the *cue* (and its dependents),[7] the relative pronoun *que*, both of these tokens or, all tokens in the *context* except these two tokens. Table 3 provides an example of sentence processing before and after an intervention of masking *que*. As the intervention occurs only when the target verb is being predicted, there is no effect on the tokens preceding it. As can be seen in this example, transformers originally assigned a higher probability to the correct plural form *accepté·s* than to the incorrect singular form *accepté*. After the intervention, the situation is reversed and the model predicts the (incorrect) singular form.

This intervention allows us to build counterfactual representations for both the past participle

[7]Masking all the *cue* dependents (predicted by an automatic dependency analysis of the sentence) allows us to ''hide'' all tokens with a morphological indication of the cue's number such as the determiner and adjectives qualifying it.

	<bos>	Les	cadeau**x**	que	le	directeur	a	accepté·**s** / accepté*
		The_Pl	gifts_**Pl**	that	the	director	has	accepted_**Pl** / accepted_Sg*
Original		−2.8	−9.5	−7.3	−1.8	−6.1	−3.9	−**5.9** / −8.3
Mask 'que'		−2.8	−9.5	−7.3	−1.8	−6.1	−3.9	−13.7 / −**11.9**

Table 3: Example sentence processed by our transformer LM, without intervention and with masking 'que' intervention. We report the log-probabilities for each token of the sentence prefixes containing either the plural form of the target verb *acceptés*, or its singular form *accepté*.

Subsets	Size (in sentences)	Original	Mask *context* tokens except `cue` `que`	Mask `cue`	Mask `que`	Mask `cue+que`
Object-past participle						
Overall	68,497	$94.6_{\pm 0.4}$	$87.1_{\pm 1.0}$	$71.4_{\pm 0.8}$	$78.8_{\pm 0.5}$	$66.9_{\pm 0.3}$
5 heuristics	32,149	$99.3_{\pm 0.05}$	$98.1_{\pm 0.1}$	$93.4_{\pm 0.3}$	$95.9_{\pm 0.2}$	$92.5_{\pm 0.5}$
4 heuristics	12,711	$96.3_{\pm 0.3}$	$86.1_{\pm 1.1}$	$78.3_{\pm 0.7}$	$85.3_{\pm 0.4}$	$75.7_{\pm 0.4}$
3 heuristics	9,159	$91.7_{\pm 0.5}$	$77.9_{\pm 1.7}$	$51.1_{\pm 1.4}$	$63.9_{\pm 1.0}$	$42.8_{\pm 0.4}$
2 heuristics	10,621	$87.4_{\pm 0.8}$	$70.8_{\pm 3.3}$	$30.1_{\pm 1.6}$	$49.4_{\pm 1.4}$	$19.6_{\pm 0.5}$
1 heuristic	2,870	$76.9_{\pm 2.3}$	$67.9_{\pm 3.3}$	$25.0_{\pm 1.4}$	$32.1_{\pm 1.0}$	$12.6_{\pm 0.4}$
0 heuristic	987	$73.8_{\pm 2.3}$	$63.0_{\pm 3.4}$	$33.1_{\pm 3.1}$	$30.8_{\pm 1.1}$	$10.0_{\pm 1.3}$
Subject-verb across object relative						
Overall	27,582	$98.9_{\pm 0.04}$	$82.0_{\pm 0.7}$	$89.1_{\pm 1.3}$	$96.7_{\pm 0.3}$	$86.0_{\pm 0.6}$
5 heuristics	14,708	$99.6_{\pm 0.05}$	$94.4_{\pm 0.5}$	$97.5_{\pm 0.4}$	$99.1_{\pm 0.1}$	$95.8_{\pm 0.3}$
4 heuristics	3,799	$99.0_{\pm 0.1}$	$83.1_{\pm 1.0}$	$89.8_{\pm 1.6}$	$96.5_{\pm 0.4}$	$87.3_{\pm 0.6}$
3 heuristics	4,189	$98.4_{\pm 0.1}$	$70.8_{\pm 0.8}$	$80.1_{\pm 3.0}$	$93.4_{\pm 0.4}$	$74.9_{\pm 1.1}$
2 heuristics	3,166	$97.7_{\pm 0.1}$	$54.9_{\pm 1.4}$	$70.1_{\pm 3.2}$	$91.5_{\pm 0.8}$	$64.2_{\pm 1.3}$
1 heuristics	1,451	$96.8_{\pm 0.1}$	$51.4_{\pm 1.9}$	$73.0_{\pm 2.3}$	$94.1_{\pm 0.6}$	$67.7_{\pm 1.9}$
0 heuristics	269	$94.1_{\pm 0.5}$	$51.0_{\pm 2.0}$	$67.9_{\pm 1.9}$	$89.3_{\pm 1.0}$	$63.8_{\pm 1.9}$

Table 4: Number prediction accuracies before and after the different masking interventions, based on prediction difficulty measured by the number of heuristics. *cue* here refers to the antecedent and its modifiers (determiners and adjectives) in object-past participle agreement and to the subject and its modifiers in subject-verb agreement.

and the main verb that do not take into account some tokens in the sentence when computing their representation, thereby removing any direct access to the information encoded in the representations of the masked tokens (e.g., the number information encoded in the *cue*).

However, information encoded in these masked tokens can still be taken into account indirectly: The target verb's representation indeed relies on the representations of all the preceding tokens in the sentence, for which the masking mechanism is kept unchanged—therefore these tokens can still encode relevant information. By comparing performances on the agreement tasks with and without intervention, we can evaluate whether the representation of one or several specific token(s) has a direct impact on the choice of the target verb form.

Results Table 4 reports the results of our causal interventions on the object-past participle and subject-verb agreement tasks. Accuracies are broken down by agreement difficulty. It appears that, for these two kinds of agreement, the *cue* (i.e., the antecedent group and subject group) is critical for predicting the corresponding agreement: Masking these tokens strongly degrades transformers prediction performance on the harder cases (i.e., 0, 1, and 2 heuristics subsets). This suggests that transformers learn representations that are consistent with the French grammar: The model relies on the same tokens as humans to choose the correct form of the target past participle and the main verb.

Quite remarkably, the relative pronoun *que* plays a very different role in determining the form of the target verbs in these two agreement phenomena: Masking the relative pronoun on object-past participle agreement results in lower than chance prediction accuracies (e.g., the accuracy on the *1 heuristic* subset drops from 76.9% to 32.1%), while it hardly impacts the prediction of the subject-verb agreement: Accuracy drops by no more than 6.2 points (for the *2 heuristics* subset). This suggests that even though the two agreement phenomena have almost identical surface forms and, as reported in §4.1, the syntactic number information is encoded in a similar way in both phenomena, transformers use two distinct agreement mechanisms to resolve the object-past participle and the subject-verb agreement. This distinction is consistent with the analysis of theoretical linguistics.

Results reported in Table 4 also show that, in the case of subject-verb agreement across object relatives, the *context* tokens (i.e., the tokens in the relative clause between the subject and the verb) have a bigger contribution to the model's decision than the subject group tokens (i.e., the subject and its dependents) with which the verb agrees. This counter intuitive observation seems to confirm the findings of Ravfogel et al. (2021) that, to predict the subject-verb agreement, the model uses information about relative boundaries, encoded in its word representations. To explain this intriguing observation, we hypothesize that even though the grammatical number is distributed across all tokens in the *context* segment, the information about relative boundaries is crucial for the model to determine how to use it to inflect the main verb, which would explain why the *context* tokens control the agreement in such an important way. This hypothesis needs to be confirmed by further experiments.

5 Discussion

Our experiments clearly show that transformers are capable of predicting quite well the object-past participle and subject-verb agreements in the case of French object relative clauses. Moreover, even though several confounders exist, the results consistently indicates that transformers base their predictions mainly on the tokens involved in agreement rules and they apply different strategies despite the very similar superficial forms of these two linguistic phenomena.

Lexical Cues Confounder To evaluate the impact of lexical information on model predictions, we converted our original evaluation sets into nonsensical but grammatically well-formed evaluation sets: Following Gulordava et al. (2018), we have created a *nonce* dataset for each agreement phenomenon by substituting each content word of the original evaluation sentence with a random word having the same syntactic category.[8] We then evaluated our model's predictions on number agreement tasks in this *nonce* setting. The results in Table 5 show only a mild degradation relative to the *original* setting for the two architectures across the two agreement tasks: A drop of 3.4 percentage points in global performance for transformers (7.3 for LSTM) in subject-verb agreement, and a drop of 1.5 percentage points for transformers (5.0 for LSTM) in object-past participle agreement.

Evaluation sets	Size in sentences	LSTMs	Transformers
Object-pp Original			
overall	68,497	$82.1_{\pm 1.1}$	$94.6_{\pm 0.2}$
singular	44.599	$95.4_{\pm 0.7}$	$99.2_{\pm 0.1}$
plural	23,898	$57.2_{\pm 2.9}$	$86.2_{\pm 0.4}$
Object-pp Nonce			
overall	68,497*3	$77.1_{\pm 2.3}$	$93.9_{\pm 0.2}$
singular	44.599*3	$90.3_{\pm 1.3}$	$97.5_{\pm 0.1}$
plural	23,898*3	$52.4_{\pm 5.5}$	$87.2_{\pm 0.6}$
Subject-v Original			
overall	27,582	$94.3_{\pm 0.3}$	$98.9_{\pm 0.04}$
singular	19,224	$98.0_{\pm 0.3}$	$99.4_{\pm 0.05}$
plural	8,358	$85.9_{\pm 1.5}$	$97.8_{\pm 0.1}$
Subject-v Nonce			
overall	27,582*3	$87.0_{\pm 0.4}$	$95.5_{\pm 0.2}$
singular	19,224*3	$93.7_{\pm 0.7}$	$97.1_{\pm 0.1}$
plural	8,358*3	$71.6_{\pm 2.5}$	$91.9_{\pm 0.4}$

Table 5: Accuracy achieved by LSTMs and transformers on nonce experimental setting, compared to the original setting, by agreement dependency and by target number.

This drop in performance is of the same order of magnitude as that reported by Gulordava et al. (2018), suggesting that the two models are able to abstract away from the potential lexical confounds.

Simplicity Confounders As reported in Table 6, the simple heuristics of Li et al. (2021),

[8] We generated three nonce variants for each original sentence.

Heuristics	object-pp	subject-verb
(1) First noun	69.5	83.7
(2) Most recent noun	88.6	77.5
(3) Most recent token	60.3	66.9
(4) Majority number	70.0	75.9
(5) Noun before ‘‘que’’	95.7	91.6

Table 6: Accuracy (%) achieved by the 5 surface heuristics considered in this work on the long-distance agreement task.

which only rely on surface information, are able to predict the correct verb form with a very high accuracy (they even outperform a state-of-the-art LSTM). This observation, already reported by Li et al. (2021), questions the principle of using the agreement task as evidence of syntactic information being encoded in neural representations. To mitigate this ‘‘simplicity’’ confounding effect, we have systematically reported our results according to the ‘‘difficulty’’ of the agreement task and focus, in our analyses, on the hardest cases that require an abstract representation of the sentence.

It must also be noted that humans also make agreement errors (Bock and Miller, 1991); these sequential statistical heuristics and the method of sampling evaluation sets based on heuristics may therefore serve as a source of hypotheses for experiments assessing human syntactic abilities.

Frequency Bias and Imbalanced Distribution Confounders Given the fact that in written French, singular verbs (3rd person) occur five to ten times as often as their plural counterparts (Ågren and Van de Weijer, 2013), do our models' performance depend on the number of the *target* verb? As Table 5 shows, in the *original* setting, across the two agreement dependencies, both models perform consistently better on the singular condition than on the plural condition, suggesting a bias towards singular verbs in both models. This observation is in line with the conclusions of Wei et al. (2021), who found that more frequent forms are more likely to be better predicted in number agreement tasks. At the same time, the results of our intervention experiments reveal a similar pattern: The masking interventions lead to greater degradation for the plural condition across the board. The remarkable different contribution of ‘‘que’’ persists, even though the plural condition is mainly responsible for the lower than chance accuracy after masking ‘‘que’’ intervention in object-past participle agreement.

A further analysis of our evaluation datasets reveals that the relation between the class distribution and the task difficulty (measured by the count of heuristics in §3) shows a very similar pattern across both agreement dependencies, with the singular class dominating in the easiest cases—*5 heuristic* subset (Obj-pp: 94%, Subj-v: 91%) and the plural class dominating in the most difficult cases—*0 heuristic* subset (Obj-pp: 99%, Subj-v: 96%).

Therefore, this asymmetry in singular and plural condition, in terms of prediction accuracy on number agreement tasks, may be either due to the higher frequency of singular verbs in the language, or to the frequency imbalances in verb number across syntactic constructions, or to the different ways the model encodes singularity and plurality as suggested by Jumelet et al. (2019). In future work, these hypotheses could be tested by artificially manipulating the relative frequency of singular and plural nouns in various constructions in the training corpus.

6 Related Work

6.1 Understanding the Inner Working of Neural Networks

Three kinds of approaches have been used in the literature to analyze the inner working of neural networks (Rogers et al., 2020; Belinkov et al., 2020). Our analyses, based on a combination of these existing methods, bridge the gap between the representations and behavioral approaches, and provide a framework for systematically measuring the causal factors underlying the model's behavior in order to evaluate a precise hypothesis about the abstract structure.

Linguistic Probes Probing (Alain and Bengio, 2017; Hupkes et al., 2018) consists of training a supervised classifier to predict linguistic properties from models' internal representations; achieving high accuracy in this task implies that these properties were encoded in the representation. Substantial prior work (see Belinkov [2022] for a review) has used this approach to assess neural networks linguistic capacity, finding that NLMs encode a variety of information about, among others, grammatical number, part of speech and syntactic role (Giulianelli et al., 2018; Tenney et al., 2019; Jawahar et al., 2019).

Many recent studies, such as Hewitt and Liang (2019) or Pimentel et al. (2020b), have criticized the theoretical foundations of probing, in particular by raising issues related to the capacity of the classifiers used for probing representations. There is an ongoing debate as to whether the choice of a linear classifier (as we did in our experiments) is the right decision, in particular when comparing the capacity of two neural networks to capture a given linguistic phenomenon (Pimentel et al., 2020a). As we are only interested in detecting whether one linguistic property is encoded and consider a single model in our experiments, the choice of a linear classifier is fully justified.

Causal Approaches These approaches rely on interventions that modify parts of the neural networks and analyze the impact/consequences on model output (Vig et al., 2020; Ravfogel et al., 2021). They allow us to avoid one of the main pitfalls of linguistic probes, namely, that probes can only detect correlations: As shown by Vanmassenhove et al. (2017), information encoded in the representation (revealed by the probe) is not necessarily used by the model. Giulianelli et al. (2018), one of the first studies to use this approach, combined the probing and causal intervention by using a trained probe to modify NN's representations, which allows them to improve the model's agreement prediction. Finlayson et al. (2021) performed controlled interventions on input sentences to analyze NLM's syntactic agreement mechanism, showing that LMs rely on similar sets of neurons when processing similar syntactic structure. Our intervention, described in §4.2, is most closely related to Elazar et al. (2021), who proposed to erase the ''part-of-speech property'' encoded in a model representation to measure how important this information is for word prediction. Our approach differs from much prior work in causal analysis in two aspects: (i) We perform intervention on the attention mechanism rather than on the input sentences. (ii) We test linguistically motivated hypotheses through two phenomena with very similar surface forms, which provides more fine-grained insights and can be easily extended to other linguistic hypotheses.

Challenge Sets This approach seeks to assess the linguistic competence of a model on examples carefully selected to exhibit a particular language characteristic (Isabelle et al., 2017). For instance, Linzen et al. (2016), an iconic example of the approach on which our work is based, study the ability of a neural network to learn syntax-sensitive dependencies and in particular the agreement between the subject and the verb. Warstadt et al. (2019) and Warstadt et al. (2020) introduce challenges sets covering a wide range of grammatical phenomena in English. To the best of our knowledge, most of the existing challenge sets are for English; our work perfectly illustrates the benefit of extending this approach to non-English languages: Considering French, a morpho-syntactically richer language, allows us to evaluate neural networks in a more fine-grained way and to gain additional insights into their inner working.

6.2 Number Agreement Task

The prediction of subject-verb agreement has long been identified as a way to assess the capacity of neural networks to track syntactic dependencies: it was already used by Elman (1989), one of the first studies on the analysis of the inner working of neural networks. Revitalized by the seminal work of Linzen et al. (2016), this task has been used in a tremendous number of studies to bring to light the ability of neural networks to capture abstract information.

Building on these results, a first line of research has tried to provide further evidence regarding neural networks capacity to build an abstract representation of sentences either by generalizing them to other languages (Lakretz et al., 2021), or to other models (Goldberg, 2019), or by identifying possible confounds such as lexical co-occurrences (Gulordava et al., 2018) or unbalanced datasets (Li et al., 2021).

Another interesting line of research aims at generalizing the conclusions of Linzen et al. (2016) to other constructions. For instance, Marvin and Linzen (2018) evaluated the grammaticality judgement of a neural network on a wide array of linguistic phenomena, using carefully designed templates to generate sentences, and Li et al. (2021) investigated the object-verb agreement in French.

A few studies have questioned the conclusions drawn by all these studies. For instance, Newman et al. (2021) questioned the use of the 0/1 loss to evaluate the number agreement task, arguing that this evaluation does not take into account the probability distributions over the vocabulary and, when not restricted to choose between two verb forms, the model might actually give more weight to verbs with the wrong number. Lasri et al. (2022) observed that the capacity of a model to predict the correct form of the verb decreased with the sentence complexity, an observation confirmed by the results reported in Section 3.2 or reported by Li et al. (2021). Despite these potential confounds, our results align with these authors, confirming that the drop in performance is not sufficient to invalidate the conclusions of Linzen et al. (2016).

7 Conclusion

This work addresses the question of the capacity of transformers to uncover abstract representation of sentences rather than to merely capture surface patterns. To this end, we investigate the mechanisms implemented by this model to predict two kinds of agreement in French: the subject-verb and the object-past participle agreements. Even though these two agreements involve very similar word sequences, their linguistic analyses are fundamentally different.

The first set of experiments we reported show that transformers are indeed able to predict these two kinds of agreement and, by comparing their predictions with those of simple surface heuristics, we highlight their ability to capture syntactic information. In a second set of experiments, we used probing and counterfactual analysis methods to provide evidence that transformers are actually succeeding for good reasons: our analyses reveal that *(i)* the incremental transformer bases its predictions on cues that are linguistically motivated and consistent with French grammar, *(ii)* the abstract structure uncovered by transformers reflects the distinction made in the theoretical modeling of these two agreements.

Our work is a first step towards a better understanding of the inner representations of NNLMs. Designing new probes, supported by causal analysis and involving a wider range of languages, could improve our understanding of such models. In particular, our observation about the linguistically motivated distribution of syntactic information in transformers' representations could be extended to other linguistic phenomenon and languages.

Acknowledgments

We sincerely thank the reviewers and action editors for their careful reviews and insightful comments, which are of great help in improving the manuscript. This work was granted access to the HPC resources of French Institute for Development and Resources in Intensive Scientific Computing (IDRIS) under the allocation 2020-AD011012282 and 2021-AD011012408 made by GENCI. We would also like to gratefully acknowledge support from the Labex EFL (ANR-10-LABX-0083).

A Evaluation Datasets

Although in standard French, normative grammars indicate object-past participle agreement under wh-movement as obligatory, it in fact appears to be optional in colloquial French where the past participle is often produced in its default singular, masculine form (Belletti, 2017). Therefore, to determine to what proportion the sentences in the training data of language models respect the normative past participle agreement rule, we analyzed the training data with the same automatic extraction procedure used for constructing evaluation datasets (Section 2). We found that in 10,377 sentences (0.35% of training data) containing the target filler-gap dependency, 10% of them do not respect the agreement rule.

B Language Models

Hyperparameters and Perplexities The results reported in the paper are averaged over five best models in terms of the validation perplexity.

The total parameters of the LSTM models are 47,900,241. As described in the appendix of Li et al. (2021), the model of batch size 64, with dropout 0.1 and learning rate 0.0001, achieved the lowest perplexity scores: 37.1. We then trained four LSTM models with the same combination of hyperparameters, the resulting perplexities are: 36.8, 36.8, 36.9, and 37.0.

The total parameters of the transformer models are 126,674,513. We explored the initial learning rate of 0.01 and 0.02, with a dropout rate of 0.1, 0.2, 0.3, 0.4, resulting in a total of 8 combinations. The model with learning rate 0.02 and dropout rate 0.2 achieved the lowest perplexity—27.0; we trained another four transformer models with

Constructions	Size (in sentences)	LSTMs	Transformers
Object past participle			
overall	68,497	$82.1_{\pm 1.1}$	$94.6_{\pm 0.2}$
5 heuristics	32,149	$95.3_{\pm 0.6}$	$99.2_{\pm 0.1}$
4 heuristics	12,711	$85.9_{\pm 1.0}$	$96.5_{\pm 0.1}$
3 heuristics	9,159	$71.9_{\pm 1.6}$	$91.6_{\pm 0.4}$
2 heuristics	10,621	$62.2_{\pm 2.4}$	$87.6_{\pm 0.4}$
1 heuristic	2,870	$37.4_{\pm 4.1}$	$77.9_{\pm 0.8}$
0 heuristic	987	$40.2_{\pm 2.7}$	$76.1_{\pm 1.0}$
Subject-verb across object relative clause			
overall	27,582	$94.3_{\pm 0.3}$	$98.9_{\pm 0.04}$
5 heuristics	14,708	$98.6_{\pm 0.1}$	$99.6_{\pm 0.05}$
4 heuristics	3,799	$95.2_{\pm 0.5}$	$99.0_{\pm 0.1}$
3 heuristics	4,189	$91.3_{\pm 0.8}$	$98.4_{\pm 0.1}$
2 heuristics	3,166	$84.8_{\pm 1.0}$	$97.7_{\pm 0.1}$
1 heuristic	1,451	$81.2_{\pm 1.8}$	$96.8_{\pm 0.1}$
0 heuristic	269	$74.7_{\pm 2.2}$	$94.1_{\pm 0.5}$

Table 7: Accuracies (%) achieved by LSTMs and transformers as a function of the agreement prediction difficulty, on the object-pp and the subject-verb agreement tasks.

these hyperparameters and the resulting perplexities are: 26.8, 27.0, 27.1, and 27.2. Training was performed with stochastic gradient descent and used the same hyperparameters described in appendix of Li et al. (2021): The initial learning rate was fixed to 0.02 and we used a cosine scheduling on 90 epochs without annealing. The first epoch was dedicated to warmup with a linear incremental schedule for the learning rate. Batches of size 64 ran in parallel on 8 GPUs except for warm-up, where the size was fixed to 8.

C Detailed Number Predictions Results

We also report the detailed scores of Figure 2 in Table 7.

References

Malin Ågren and Joost Van de Weijer. 2013. Input frequency and the acquisition of subject-verb agreement in number in spoken and written french. *Journal of French Language Studies*, 23(3):311–333. https://doi.org/10.1017/S0959269512000312

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219. https://doi.org/10.1162/coli_a_00422

Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick. 2020. Interpretability and analysis in neural NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–5, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-tutorials.1

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72. https://doi.org/10.1162/tacl_a_00254

Adriana Belletti. 2017. (Past) participle agreement. *The Wiley Blackwell Companion to Syntax, Second Edition*, pages 1–29. https://doi.org/10.1002/9781118358733.wbsyncom081

Kathryn Bock and Carol A. Miller. 1991. Broken agreement. *Cognitive Psychology*, 23(1):45–93. https://doi.org/10.1016/0010-0285(91)90003-7

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1198

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175. https://doi.org/10.1162/tacl_a_00359

Jeffrey L. Elman. 1989. Representation and structure in connectionist models. Techical report, California University San Diego, La Jolla Center for Research in Language.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. 2021. Causal analysis of syntactic agreement mechanisms in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1828–1843, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.144

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5426

Yoav Goldberg. 2019. Assessing BERT's syntactic abilities. *CoRR*, abs/1901.05287.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. 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 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1108

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1275

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926. https://doi.org/10.1613/jair.1.11196

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1263

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1356

Jaap Jumelet, Willem Zuidema, and Dieuwke Hupkes. 2019. Analysing neural language models: Contextual decomposition reveals default reasoning in number and gender assignment. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 1–11, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/K19-1001

Josef Klafka and Allyson Ettinger. 2020. Spying on your neighbors: Fine-grained probing of contextual embeddings for information about surrounding words. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4801–4811, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.434

Yair Lakretz, Dieuwke Hupkes, Alessandra Vergallito, Marco Marelli, Marco Baroni, and Stanislas Dehaene. 2021. Mechanisms for handling nested dependencies in neural-network language models and humans. *Cognition*, page 104699. https://doi.org/10.1016/j.cognition.2021.104699, PubMed: 33941375

Karim Lasri, Alessandro Lenci, and Thierry Poibeau. 2022. Does BERT really agree? Fine-grained analysis of lexical dependence on a syntactic task. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2309–2315, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-acl.181

Bingzhi Li, Guillaume Wisniewski, and Benoit Crabbé. 2021. Are Transformers a modern version of ELIZA? Observations on French object verb agreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4599–4610, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bingzhi Li, Guillaume Wisniewski, and Benoit Crabbé. 2022. How distributed are distributed representations? An observation on the locality of syntactic information in verb agreement tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 501–507, Dublin, Ireland. Association for Computational Linguistics.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7:195–212. https://doi.org/10.1146/annurev-linguistics-032020-051035

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535. https://doi.org/10.1162/tacl_a_00115

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1151

Benjamin Newman, Kai-Siang Ang, Julia Gong, and John Hewitt. 2021. Refining targeted syntactic evaluation of language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3710–3723, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.290

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Tiago Pimentel, Naomi Saphra, Adina Williams, and Ryan Cotterell. 2020a. Pareto probing: Trading off accuracy for complexity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3138–3153, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.254

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020b. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.420

Shauli Ravfogel, Grusha Prasad, Tal Linzen, and Yoav Goldberg. 2021. Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 194–209, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.conll-1.15

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866. https://doi.org/10.1162/tacl_a_00349

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung

Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Eva Vanmassenhove, Jinhua Du, and Andy Way. 2017. Investigating ''aspect'' in NMT and SMT: Translating the English simple past and present perfect. *Computational Linguistics in the Netherlands Journal*, 7:109–128.

Sara Veldhoen, Dieuwke Hupkes, and Willem H. Zuidema. 2016. Diagnostic classifiers revealing how neural networks process hierarchical structure. In *CoCo@ NIPS*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: A benchmark of linguistic minimal pairs for English. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 409–410, New York, New York. Association for Computational Linguistics. https://doi.org/10.1162/tacl_a_00321

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641. https://doi.org/10.1162/tacl_a_00290

Jason Wei, Dan Garrette, Tal Linzen, and Ellie Pavlick. 2021. Frequency effects on syntactic rule learning in transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 932–948, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.72