Are Transformers a Modern Version of ELIZA? Observations on French Object Verb Agreement

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

Abstract

Many recent works have demonstrated that unsupervised sentence representations of neural networks encode syntactic information by observing that neural language models are able to predict the agreement between a verb and its subject. We take a critical look at this line of research by showing that it is possible to achieve high accuracy on this agreement task with simple surface heuristics, indicating a possible flaw in our assessment of neural networks' syntactic ability. Our fine-grained analyses of results on the long-range French object-verb agreement show that contrary to LSTMs, Transformers are able to capture a non-trivial amount of grammatical structure.

1 Introduction

The long distance agreement task is one of the most popular method to assess neural networks (NN) ability to encode syntactic information: Linzen et al. (2016) showed that LSTMs are able to predict the subject-verb agreement in English and has initiated a very active line of research. Since then, many studies have generalized this observation to other languages (Gulordava et al., 2018), other models such as Transformers (Goldberg, 2019; Jawahar et al., 2019) or have identified possible confounding factors that could distort the stated conclusions (Gulordava et al., 2018; Marvin and Linzen, 2018). All of these studies show that NN are able to learn a 'substantial amount' of syntactic information (Belinkov and Glass, 2019).

In this work, we propose to take an alternative look at these results by studying whether neural networks are able to predict the correct form of a verb because they are able to build an abstract, high-level (maybe hierarchical) sentence representation (Giulianelli et al., 2018; Lakretz et al., 2019) or solely because they capture surface statistical regularities, as suggested by several recent work (Sennhauser and Berwick, 2018; Chaves, 2020; Li and Wisniewski, 2021). Overall, this set of results questions one of the most fundamental assumption in linguistics (Lakretz et al., 2021), namely that a sentence has a recursive structure (Everaert et al., 2015): while LSTMs with proper parametrization can model context-free patterns (Suzgun et al., 2019), Transformers are essentially feed forward models relying on a large number of attention heads. Consequently, they are, in theory, not adapted to model hierarchical syntactic patterns (Hahn, 2020) and explaining their capacity to predict accurately syntactic agreement patterns remains an open issue.

We bring new light on this problematic by identifying simple heuristics (§4) that can be used to correctly predict verbal agreement, pushing further the observation of Kuncoro et al. (2018) that a simple rule can provide highly accurate results on the task. Using our extended set of heuristics, we identify sentences for which predicting the correct verb form requires a more abstract representation of the sentence. By comparing models' performance on these examples, we show that contrary to LSTMs, Transformers perform consistently well in these critical cases.

2 Test Set for French Object Past-Participle Agreement[1]

We focus on the object-verb agreement (i.e. object past-participle agreement) in French: agreement in number and gender occurs between the object and the past participle when the latter is used with the auxiliary *avoir* (to have) and the object is located before the verb. As shown in Figure 1, this is, for instance, the case for past participles in object relatives. When agreement is required, a -s suffix (resp. -e) has to be added to past participles for

[1] The code of all our experiments as well as the corpora we used in this work can be downloaded from https://gitlab.huma-num.fr/bli/syntactic-ability-nlm.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 4599–4610
November 7–11, 2021.

plural object (resp. feminine).

To predict the past participle agreement in object relatives, a model has to identify the object relative pronoun, its antecedent and the auxiliary. It has also to ignore the effect of attractors (nouns with misleading agreement features) occurring between the object and the past participle. Compared to the subject-verb agreement, the French object past participle agreement is more difficult as the target verb form depends on a noun that is never adjacent to the verb. The auxiliary *avoir* before the target verb could also be an attractor. [2]

We restrict ourselves to the number agreement between object and past participle in the case of object relatives to (1) design reasonably simple patterns that can be easily extracted automatically from raw texts, (2) extract a sufficiently large number of representative examples and (3) reduce the importance of the anaphoric resolution problem. These restrictions allow us to carry out a fine-grained analysis of NN ability to extract syntactic generalizations from non-annotated corpora (§4).

Building a Test Set Sentences used in the number agreement task are extracted automatically from the 8,638,145 sentences of the French Gutenberg corpus.[3] We use the FLAUBERT parser (Le et al., 2019) and the pretrained French model of `spaCy` (Honnibal et al., 2020) to automatically parse the sentences of the Gutenberg project. We also consider the gold annotations of the 45,088 sentences of the French Universal Dependency treebanks (Zeman et al., 2020) to evaluate the impact of parsing errors on the corpus quality.

We extract examples of object-verb agreement from sentences' syntactic and morphological annotations using simple rules,[4] resulting in a corpus of 104 sentences (68% singular and 32% plural) extracted from the UD treebank and of 68,794 sentences (65% singular and 35% plural) extracted from the Gutenberg project. In French, the singular is identical to the unmarked form of the past participle verbs, making the frequency statistics unbalanced in favor of singular.

We evaluate the quality of our automatic extraction procedure by comparing the examples extracted using the gold annotations of UD treebank to those extracted from predicted annotations of UD treebank sentences (generated by FLAUBERT and `spaCy`). Our automatic procedure correctly picked up 98%[5] of the object past participle agreement examples.

3 Language Models

We contrast two types of incremental language models in our experiments: LSTM models and incremental Transformer models. Both models are modeling the probability of a sentence $\mathbf{x}$ as:

$$P(\mathbf{x}) = \prod_{i=1}^{n} P(x_i|x_1 \dots x_{i-1}) \qquad (1)$$

All neural models are trained to compute $P(x_i|x_1 \dots x_{i-1})$ and they all use the same generic template:

$$P(x_i|x_1 \dots x_{i-1}) = \text{SOFTMAX}(\mathbf{W}_{dec}\mathbf{c}_{i-1} + \mathbf{b}) \qquad (2)$$

$$\mathbf{c}_{i-1} = \text{CONTEXT}(\mathbf{e}_1 \dots \mathbf{e}_{i-1}) \qquad (3)$$

$$\mathbf{e}_i = \mathbf{W}_{enc}\mathbf{x}_i \qquad (4)$$

where $\mathbf{x}_i$ are one-hot word vectors; $\mathbf{W}_{enc}$ and $\mathbf{W}_{dec}$ are tied parameter matrices, the latter being the transpose of the former, encoding respectively the word embeddings and the output layer of the language model.

A context model (CONTEXT) is either an incremental LSTM or a Transformer decoder where the sequence of embeddings $\mathbf{e}_i \dots \mathbf{e}_n$ is masked (i.e. the probability of the i-th word is estimated knowing only the first (i-1) words of the sentences, contrary to the 'standard' Transformer models which assume that the whole sentence is known). The context vector $\mathbf{c}$ returned by the context model is either the hidden vector of the LSTM at step $i-1$ or the vector returned by the upper layer of the Transformer at step $i-1$.

Our LSTM models use 2 layers while our Transformer language model use 16 layers and 16 heads. Both models are using embeddings of size 768 and are trained on the same data. For Transformers we add positional embeddings to the word embeddings $\mathbf{e}_i$ using the cosine scheme and weighting described by Vaswani et al. (2017). Since all the models use a word-based tokenization and not a subword tokenizer, we bound the vocabulary to the 50,000 most frequent tokens found in the training data and use an `<unk>` token to encode the least frequent tokens.

[2] See example(1) in Figure 1, *a* (has_3Sg) could be a number attractor for target verb *acceptées*(accepted_Pl)

[3] `https://www.gutenberg.org/`

[4] See appendix C for a full description

[5] Qualitative analysis is in Section C of the appendix.

(1)	Le	nombre	d'	offres	**que**	le	directeur	a	**acceptées**
	The-DET-Sg	number-N-M-Sg	of-ADP	offers-N-**F-Pl**	that-PRON	the-DET-3Sg	director-N-M-Sg	has-AUX-3Sg	accepted-PP-**F-Pl**

The number of offers that the director has accepted...

(2)	Les	offres$_{h1}$	**que**	les	directeurs$_{h2}$	ont$_{h3}$	**acceptées**	...
	The-DET-Pl	offers-N-**F-Pl**	that-PRON	the-DET-Pl	directors-N-M-**Pl**	have-AUX-3**Pl**	accepted-PP-**F-Pl**	...

The offers that the directors have accepted...

Figure 1: Examples of object-verb agreement in French. The past participle in the relative clause (in blue) has to agree in gender and in number with its object (also in blue) when the latter is placed before the verb. To predict the agreement the model has to identify the antecedent of the relative pronoun (dashed arrow)

corpus	size in sentences	LSTMs	Transformers
Original Test Set			
overall	68,497	80,8±1.2	93.5 ±1.4
4 heuristics	32,311	96.4 ±0.6	99.0 ±0.4
3 heuristics	13,222	84.0 ±1.7	95.1 ±1.5
2 heuristics	8,869	66.5 ±2.7	89.5 ±2.3
1 heuristic	10,946	55.7 ±3.5	84.2 ±3.0
0 heuristic	3,149	34.9 ±6.8	74.1 ±4.1
Permuted Test Set			
overall	68,497	69.0 ±0.6	70.4 ±1.0
4 heuristics	32,311	87.0 ±1.2	88.0 ±0.8
3 heuristics	13,222	73.6 ±0.6	73.9 ±0.9
2 heuristics	8,869	52.0 ±0.3	54.8 ±1.8
1 heuristic	10,946	35.3 ±0.3	37.4 ±1.5
0 heuristic	3,149	30.2 ±0.4	32.6 ±1.2

Table 1: Accuracy achieved by LSTMs and Transformers on the object-verb agreement task for the *Original* and *Permuted* test sets. Results are averaged over the three best models in terms of the validation perplexity for each architecture

This setting aims to get reasonably fair comparisons between LSTM and Transformers. To train the models, we extracted raw text from a recent French Wikipedia dump using `WikiExtractor` (Attardi, 2015) and then segmented and tokenized it with the `Moses` tokenizer (Koehn et al., 2007). We filtered out sentences with more than 5% unknown words based on the lemma annotations generated by `TreeTagger` (Schmid, 1999). Finally, we sampled a subset containing 100M tokens and split it into training, validation and test sets with a standard 8:1:1 proportion.

4 Experimental Results

In our experiments, following Linzen et al. (2016) and Gulordava et al. (2018), we compare the probabilities a language model assigns to the singular form of the target participle and its plural form given a *prefix*.[6] We consider the model has predicted the agreement correctly if the form with the correct number has a higher probability than the form with the incorrect number.

Table 1 reports the accuracy of two types of models evaluated in this framework. Even for this difficult task, the models perform, overall, very well: LSTMs achieve an accuracy of 80.8%, a performance similar to the one reported in the literature.[7] With an accuracy of 93.5%, Transformers perform even better. These preliminary results support the conclusion, drawn by many works, that neural networks encode syntactic information.

However, we believe that this conclusion must be taken with great care: because of confounding factors, a language model could predict the correct form without actually capturing syntactic information. For instance, as our test set is unbalanced (section 2) a naive model always choosing the singular form of the participle achieves an accuracy of 65%, a score that puts into perspective the performance of LSTMs. More importantly, Gulordava et al. (2018) and Kuncoro et al. (2018) observed that the agreement task can be partially solved by collocational information or a simple heuristic, namely the number of the first noun of the prefix. In the following, we propose several experiments to strengthen these first results.

4.1 Agreement with Surface Heuristics

Extending the observations of Kuncoro et al. (2018), we identify four heuristics that a model could adopt to predict the verb's number only from

[6] The prefix is made of words from the beginning of a sentence up to and excluding the target past participle

[7] For instance, for the subject-verb agreement task, Gulordava et al. (2018) reported an overall accuracy of 81% for English and Mueller et al. (2020) of 83% for a wide array of constructions in French.

Heuristics	Accuracy
h1: First noun	69.5%
h2: Last noun	88.6%
h3: Last token	60.3%
h4: Majority number	70.0%

Table 2: Heuristics' accuracy on French object past participle agreement task

corpus	size in sentences	LSTMs	Transformers
Original Test Set			
overall	68,497	80,8 ±1.2	93.5 ±1.4
singular	44,599	96.4 ±1.1	98.9 ±0.4
plural	23,898	51.6 ±4.7	83.5 ±3.3
Nonce Test Set			
overall	68,497*3	78,1 ±1.2	92.6 ±1.9
singular	44,599*3	93 ±2.3	96.8 ±0.9
plural	23,898*3	50.3 ±6.8	84.7 ±3.6
Mirror Test Set			
overall	68,497	59,8 ±2.5	81.3 ±2.7
singular	23,898	90.6 ±1.8	91.8 ±0.7
plural	44,599	43.5 ±4.5	75.8 ±3.8

Table 3: Accuracy achieved by LSTMs and Transformers on different experimental settings, by target verb number and averaged

surface information. Each of these heuristics assumes that the target past participle agrees systematically in number with:

h1. the *first noun* in the prefix;
h2. the *last noun* in the prefix;
h3. the *last token* in the prefix with a mark of number;
h4. the majority number expressed in the prefix.

The example (2) in Figure 1 illustrates the tokens that each heuristic relies on to make its decision. These heuristics are not tailored to the prediction of the object-past participle agreement in French: they could easily be used to other agreement tasks in other language. More complicated, task-specific heuristics could have been designed. We could, for instance, consider the first noun on the left of the relative pronoun.

Surprisingly enough, as reported in Table 2, on our test set, these heuristics achieve an accuracy between 60.3% (for h3) and 88.6% (for h2). These results challenge our previous conclusion: they show that the ability to predict the correct number of the verb cannot be used to prove that a model captures abstract syntactic relations, since a simple surface heuristic outperforms LSTMs and achieves an accuracy only slightly worse than that of Transformers. On the contrary, it suggests that NN, like Eliza, only extract and combine surface patterns to make their decisions.

To shed further light on this new perspective, we use these heuristics to quantify the 'difficulty' of the task: for each example of our test set, we count the number of heuristics that predict the correct form and consider that the higher this number, the easier the prediction. We then divide our test set into five different subsets according to the number of heuristics that a model could rely on to predict the verb form: the *4 heuristics* group gathers the 'easiest' examples, while examples in the *0 heuristic* group are the most difficult, for which the choice of the verb number cannot rely on simple surface heuristics and requires building a more abstract representation of the sentence. [8]

Table 1 reports the results achieved by our models according to the prediction difficulty. The two architectures have a very different behavior: while they both show high agreement prediction accuracy in the simplest case (the *4 heuristics* group), LSTMs' performance drops sharply with increasing task difficulty: with an accuracy of only 34.9% on the most difficult examples (the *0 heuristic* group), they perform worse than random. On the contrary, even if Transformers' performance also degrades with increasing task difficulty, they perform consistently much better on all groups: they are still predicting the correct verb number for 74.1% of the most difficult examples, suggesting that Transformers are able to extract certain abstract generalizations.

4.2 Control Experiments

To corroborate these results and avoid some known pitfalls of the agreement task, we have performed four control experiments.

Lexical Cues Following Gulordava et al. (2018), we convert the original test set into a nonsensical but grammatically correct test set to ensure that the model is not using collocational information to choose the correct form of the verb. [9] Results in Table 3 show that for LSTMs (resp. Transformers), the global accuracy drops from 80.8% (resp.

[8] Table 5 in the appendix describes examples of sentences in each of these groups.

[9] Generation procedure is detailed in Section C of the appendix.

93.5%) for the original set to 78.1% (resp. 92.6%) for the so-called *nonce* test set. This drop is of the same order of magnitude as that reported by Gulordava et al. (2018), showing that the lexical or collocational confounds have only a moderate impact on models' performance in our agreement prediction task.

Frequency Bias and Imbalanced Data Another possible confound identified in this work (§2) results from the imbalance between classes: most of the past participles in French are singular and 65% of the target past participle in our test set are singular. That is why, as expected, models perform better in predicting singular form than plural form(*Original Test Set* of Table 3): both LSTMs and Transformers predict almost perfectly singular forms (accuracy: 96.4% and 98.9%), but accuracy on plural verbs drops sharply: LSTMs correctly predict 51.6% of the plural forms and Transformers appear to be more robust with an accuracy of 83.5%.

To ensure, a model is not simply memorizing the most frequent form of a verb, we have generated a *mirror test set* in which each plural verb is automatically transformed into singular (and vice-versa) as well as the corresponding object and all its adjective and pronoun modifiers to make sure that the modified sentence is grammatically correct.

The accuracy of LSTMs and Transformers on the *mirror set* is of 59.8% and 81.3% (Table 3). This drop suggests that more frequent forms are more likely to be better predicted, even though Transformers are more robust to the low frequency bias. Compared to the *nonce* setting, models' performance is impacted to a much larger degree in *mirror* setting. We don't have a clear explanation to this surprising observation, which need to be explored through new experiments.

Distance Following Linzen et al. (2016) we have examined how models' performance on this agreement task is affected by distance between the object and the target verb. Results, reported in Table 8 in the appendix show that models' performance decreases slightly as the distance increases, except for the shortest distance, thus replicating the results of Linzen et al. (2016).

Word Order We now test to which extent a model relies on word order to predict the verb number. We convert each original example into a scrambled example by randomly permuting its prefix. As reported in Table 1, despite the fact that the syntax has been destroyed in shuffled prefixes setting, both models still achieve high accuracy for the easy examples but achieve worse than chance accuracy for the *0* and *1 heuristic* groups, confirming that syntactic information is critical for models to solve the most difficult cases. For Transformers, the difference in accuracy between the original and permuted setting on the *0 heuristic* group extends up to 41.5 percentage points! These results suggest that Transformers perform significantly better than surface heuristics and capture a non trivial amount of word order information.

5 Conclusions

We ran a fine-grained analysis of NN's syntactic generalization capabilities in processing French object-verb agreement for grammatical number, a phenomenon crucially depending on hierarchical syntactic structure. We designed a new evaluation protocol based on four shallow heuristics that the models could adopt to perform number agreement task. Our experiments show that, contrary to LSTMs, Transformers extract a non trivial amount of syntactic information.

In future work, we will investigate the kind of syntactic information Transformers are encoding and the relationship between the superficial heuristics and hierarchical syntactic structure processing in Transformer models. In particular, our results intriguingly suggest that Transformers rely on word order information to predict verb agreement, despite the fact that they don't model word order explicitly beyond marking each word with its absolute-position embedding. We plan to study this question in future work.

Acknowledgments

We sincerely thank the reviewers and Program Chairs 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.

References

Giusepppe Attardi. 2015. Wikiextractor. https://github.com/attardi/wikiextractor.

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

Rui Chaves. 2020. What don't RNN language models learn about filler-gap dependencies? In *Proceedings of the Society for Computation in Linguistics 2020*, pages 1–11, New York, New York. Association for Computational Linguistics.

Noam Chomsky et al. 1957. Mouton. *Syntactic Structures", The Hague*.

Martin B.H. Everaert, Marinus A.C. Huybregts, Noam Chomsky, Robert C. Berwick, and Johan J. Bolhuis. 2015. Structures, not strings: Linguistics as part of the cognitive sciences. *Trends in Cognitive Sciences*, 19(12):729–743.

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. *arXiv preprint arXiv:1808.08079*.

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.

Michael Hahn. 2020. Theoretical limitations of self-attention in neural sequence models. *Trans. Assoc. Comput. Linguistics*, 8:156–171.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

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.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Adhiguna Kuncoro, Chris Dyer, John Hale, and Phil Blunsom. 2018. The perils of natural behaviour tests for unnatural models: the case of number agreement. *Poster presented at Learning Language in Humans and in Machines, Paris, Fr., July*, pages 5–6.

Yair Lakretz, Theo Desbordes, Jean-Rémi King, Benoît Crabbé, Maxime Oquab, and Stanislas Dehaene. 2021. Can rnns learn recursive nested subject-verb agreements? *CoRR*, abs/2101.02258.

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in lstm language models. *arXiv preprint arXiv:1903.07435*.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2019. Flaubert: Unsupervised language model pre-training for french. *arXiv preprint arXiv:1912.05372*.

Bingzhi Li and Guillaume Wisniewski. 2021. Are neural networks extracting linguistic properties or memorizing training data? an observation with a multilingual probe for predicting tense. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3080–3089, Online. Association for Computational Linguistics.

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.

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.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-linguistic syntactic evaluation of word prediction models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online. Association for Computational Linguistics.

Helmut Schmid. 1999. Improvements in part-of-speech tagging with an application to german. In *Natural language processing using very large corpora*, pages 13–25. Springer.

Luzi Sennhauser and Robert Berwick. 2018. Evaluating the ability of LSTMs to learn context-free grammars. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 115–124, Brussels, Belgium. Association for Computational Linguistics.

Mirac Suzgun, Yonatan Belinkov, and Stuart M. Shieber. 2019. On evaluating the generalization of LSTM models in formal languages. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielė Aleksandravičiūtė, Ika Alfina, Lene Antonsen, Katya Aplonova, Angelina Aquino, Carolina Aragon, Maria Jesus Aranzabe, H͡órunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Kepa Bengoetxea, Gözde Berk, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Kristín Bjarnadóttir, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Ethan Chi, Yongseok Cho, Jinho Choi, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Arawinda Dinakaramani, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Marília Fernanda, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Fabrício Ferraz Gerardi, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Sam Hardwick, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Eva Huber, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Sarveswaran K, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Sookyoung Kwak, Veronika Laippala, Lucia Lam, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phng Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Cheuk Ying Li, Josie Li, Keying Li, Yuan Li, KyungTae Lim, Krister Lindén, Nikola Ljubešić, Olga Loginova, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Karina Mischenkova, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lng Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Alireza Nourian, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Arzucan Özgür, Balkız Öztürk Başaran, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Eiríkur Rögnvaldsson, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Olga Rudina, Jack Rueter, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Benoît Sagot, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Kevin Scannell, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Einar Freyr Sigurðsson, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Steinh͡ór Steingrímsson, Antonio Stella, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Samson Tella, Isabelle Tellier, Guillaume Thomas, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, Hanzhi Zhu, and Anna Zhuravleva. 2020. Universal dependencies 2.7. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

A Language Models

Hyperparameters and perplexities The results reported in the paper are averaged over three best models in terms of the validation perplexity after 40 training epochs for LSTMs and 50 training epochs for Transformer. The detailed information of the top 3 LSTM and Transformer models is described in table 4.

For the LSTM models, we used embeddings of size 768, with 2 layers. The total parameters are 47,900,241 and we explored the following hyperparameters, for a total of 12 combinations:

1. batch size: 32, 64(only for learning rate 0.0001)
2. dropout rate: 0.0, 0.1, 0.2, 0.3
3. learning rate: 0.001, 0.0001

For the Transformer models we used embeddings of size 768, with 16 layers, each with 16 heads. The total parameters are 126,674,513. Training was performed with stochastic gradient descent. The initial learning rate was fixed to 0.02 and we used a cosine scheduling on 50 epochs without annealing. The first epoch was dedicated to warmup with a linear incremental schedule for the learning rate. Batches are of size 64 run in parallel on 8 GPUs except for warmup where the size was fixed to 8. We explored the initial learning rate of 0.01 and 0.02, the dropout rate of 0.0, 0.1, and 0.2, resulting in a total of 6 combinations.

B Surface heuristics

We defined four heuristics that a model could adopt to predict the verb's number only from surface information. And then we divided the test set into five subsets based on the number of heuristics, Table 5 describes the examples for each subgroup.

C Construction of test sets

Extraction procedure Extraction of the object-verb agreement examples is based on the dependency structure and morphological information of sentences. Concretely, a valid example has to include a NOUN and VERB connected by an `acl:relcl` dependency arc as well as a direct object *que* (that); the auxiliary connected to the target verb has to be *avoir* (to have). Using the morphological information, we filtered out sentences in which the noun and the verb do not agree in number and gender as well as sentences in which not all words from the antecedent to the target(enclosed) occur in the language model's vocabulary. To reduce the importance of anaphoric resolution problems, we have ruled out the complex and ambiguous cases: long distance dependencies (First example in Figure 2) and coordinated object noun phrase as antecedent case (Second example in Figure 2). But we didn't exclude the propositional phrase as antecedent case, because there is no ambiguity in determining the antecedent of the relative pronoun, illustrated by the third example in Figure 2.

Qualitative evaluation of extraction procedure Our automatic extraction procedure correctly identified 102 examples from automatically parsed UD treebanks sentences among 104 examples using the gold annotation of French UD treebanks. Our procedure excluded the first missed example by annotating the intervening relative pronoun *que* (that) as conjunction: **formule qu**'avec un sens de la nuance plus marseillais que britannique, le président de l'académie a **appliquée** (***formula***$_{\textit{Fem-Sg}}$ *with a sense of nuance that was more Marseillais than British,* ***that*** *the president of the academy* ***applied***$_{\textit{Fem-Sg}}$). And for the second one: une manière de **révolution** sur lui-même, qu'il a **opérée**... (A way of **revolution**$_{\text{Fem-Sg}}$ on himself, that he **operated**$_{\text{Fem-Sg}}$...), the automatically parsed annotation erroneously identified the antecdent as 'way' instead of 'revolution'. The two missed examples reflect also the difficulty of this task for a model.

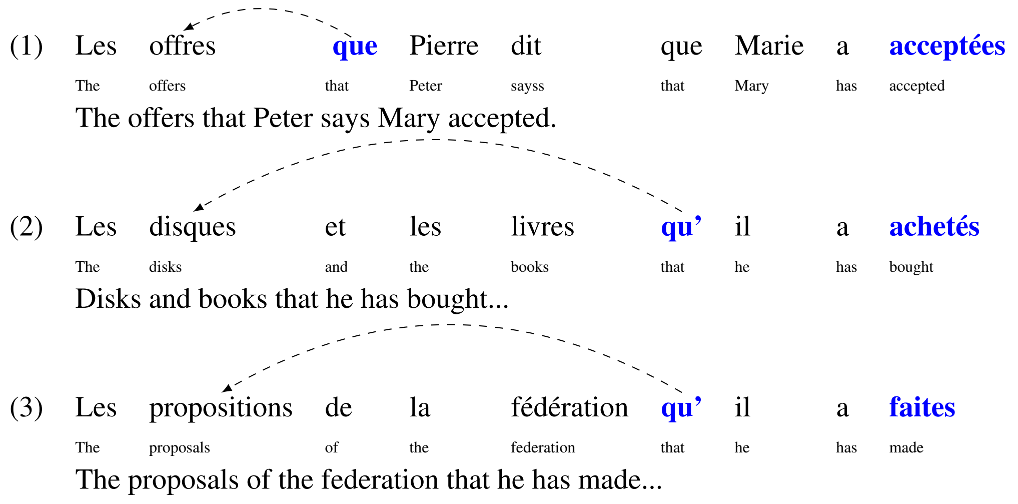
Nonce test set To test the extent to which the lexical or collocational information contribute to model's performance on number agreement task, we adapted the generation procedure of Gulordava et al. (2018) to generate three "colorless green idea" (Chomsky et al., 1957) sentences for each original sentence: each content word of the original sentence is replaced with a random word from the same syntactic category (i.e.,the same POS and morphological features). During the substitution procedure, we excluded the word forms that appeared in the treebank with more than one POS to make sure that the random words used are all with unambiguous POS(e.g., *données* can be a plural noun(data) or the plural past participle of verb *donner* (give)). To respect the argument structure constraints, the target verb could only be replaced by another random transitive word. So the *Nonce Test Set* retains the grammatical syntax of original sentences but are highly semantically implausible.

	hidden/ embedding size	layers	batch size	dropout rate	learning rate	best epoch	ppl
LSTM	768	2	32	0.1	0.001	21	40.5
	768	2	32	0.2	0.0001	38	39.3
	768	2	64	0.2	0.0001	36	37.9
Transformer	768	16	64	0	0.02	41	31.4
	768	16	64	0.2	0.01	50	28.5
	768	16	64	0.1	0.01	49	28.2

Table 4: Hyperparameters and perplexities of top 3 LSTMs and Transformers used in this work

Subsets	Examples	Heuristics	class
4	$_{h4}$Les offres$_{h1}$ que les directeurs$_{h2}$ ont$_{h3}$ acceptées... *The offers_**Pl** that the_Pl directors_Pl have_Pl accepted_**Pl** ...*	h1,h2,h3,h4	Plural
3	$_{h4}$Le nombre d'offres$_{h2}$ qu'ils ont$_{h3}$ acceptées... *The number_Sg of offers_**Pl** that they_Pl have_Pl accepted_**Pl** ...*	h2,h3,h4	Plural
2	Les offres$_{h1h2}$ qu'il a acceptées... *The offers_**Pl** that he_Sg has_Sg accepted_**Pl** ...*	h1,h2	Plural
1	Les offres$_{h1}$ que le directeur a acceptées... *The offers_**Pl** that the_Sg director_Sg has_Sg accepted_**Pl** ...*	h1	Plural
0	Le nombre d'offres que le directeur a acceptées... *The number_Sg of offers_**Pl** that the_Sg director_Sg has_Sg accepted_**Pl** ...*	none	Plural

Table 5: Examples of five subsets according to the number of heuristics that a model could rely on to predict the verb form



(1) Les offres **que** Pierre dit que Marie a **acceptées**

The offers that Peter sayss that Mary has accepted

The offers that Peter says Mary accepted.

(2) Les disques et les livres **qu'** il a **achetés**

The disks and the books that he has bought

Disks and books that he has bought...

(3) Les propositions de la fédération **qu'** il a **faites**

The proposals of the federation that he has made

The proposals of the federation that he has made...

Figure 2: The test set excluded the complex long distance dependencies (1) and ambiguous coordinated object noun phrase (2), but kept the prepositional phrase as antecedent cases like (3)

corpus	size in sentences	LSTMs	Transformers
Nonce Test Set			
overall	68,497	78.1 ± 1.2	92.6 ± 1.9
4 heuristics	32,311	94.3 ± 1.1	98.3 ± 0.7
3 heuristics	13,222	80.3 ± 2.5	93.5 ± 1.9
2 heuristics	8,869	63.2 ± 2.1	89.1 ± 2.9
1 heuristic	10,946	53.0 ± 5.1	84.0 ± 3.5
0 heuristic	3,149	32.3 ± 11	69.1 ± 4.5

Table 6: Accuracy achieved by LSTMs and Transformers on the *nonce test set*, based on prediction difficulty

Table 7 gives an example of a nonsensical sentence converted from its original version.

Mirror test set We generate a singular version of each plural object sentence and vice versa by substituting respectively the antecedent and target verb of each original sentence with their opposite number form. We converted also the adjective and pronoun modifiers of the antecedent to their opposite number form if they are present. At the end, we got a "inverted copy" of the original set in terms of class distribution, 35% singular and 65% plural compared to its original version: 65% singular and 35% plural. Table 7 gives an example of the *mirror* sentence converted from its original version.

D Detailed results

Nonce Set The detailed results on *Nonce Test Set* are reported in Table 6.

Distance Table 8 reports the average prediction accuracy on *Original Test set* as a function of distance between the antecedent and the target verb. The shortest distance (i.e. construction with only two intervening tokens: the relative pronoun and the auxiliary verb) is more challenging for both LSTMs and Transformers due to the attraction effect of the auxiliary. In this non-canonical construction(1,599 examples), the embedded subject in the objective clause occurs after its predicate. Our fine-grained analysis shows that in this non-canonical case, when the number of the intervening auxiliary is different with that of the past participle verb, LSTMs' performance drops to 41.9% and Transformers still achieve an accuracy of 80%, suggesting that Transformer are more robust to resist the lure of adjacent auxiliary attractor.

Test sets	Examples	label
Original	Les **offres** que le directeur a **acceptées**...	Pl
	*The offers_**Pl** that the director has accepted_**Pl** ...*	
Nonce	Les **omellettes** que le professeur a **attachées**...	Pl
	*The omelettes_**Pl** that the professor has attached_**Pl** ...*	
Mirror	L' **offre** que le directeur a **acceptée**	Sg
	*The offer_**Sg** that the director has accepted_**Sg** ...*	
Permuted	directeur a Les que **offres** le **acceptées** ...	Pl
	*director has The that offers_**Pl** the accepted_**Pl** ...*	

Table 7: Examples of test sets used in original and control experiments

	2 tokens	3-4	5-6	7-8	9-10	11-12	13-14
LSTMs	$73.1_{\pm 0.9}$	$82.9_{\pm 1.5}$	$78.7_{\pm 1.2}$	$75.9_{\pm 0.6}$	$74.1_{\pm 0.3}$	$72_{\pm 0.6}$	$69.3_{\pm 1.2}$
Transformers	$88.0_{\pm 3.0}$	$95.1_{\pm 1.2}$	$92.4_{\pm 1.6}$	$89.7_{\pm 1.9}$	$87.8_{\pm 2.2}$	$85.2_{\pm 2.2}$	$83.1_{\pm 1.7}$
# examples	1,599	44,012	14,945	4,799	1729	756	327

Table 8: Accuracy as a function of distance (i.e. number of tokens) between the antecedent and the target verb