# How Distributed are Distributed Representations? An Observation on the Locality of Syntactic Information in Verb Agreement Tasks

**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

This work addresses the question of the localization of syntactic information encoded in the transformers representations. We tackle this question from two perspectives, considering the object-past participle agreement in French, by identifying, first, in which part of the sentence and, second, in which part of the representation syntactic information is encoded. The results of our experiments using probing, causal analysis and feature selection method, show that syntactic information is encoded locally in a way consistent with the French grammar.

## 1 Introduction

Transformers (Vaswani et al., 2017) have become a key component in many NLP models, arguably due to their capacity to uncover distributed representation of tokens (Hinton et al., 1986) that are *contextualized*: thanks to a multi-head self-attention mechanism (Bahdanau et al., 2015), a token representation can, virtually, depend on the representations of all other tokens in the sentence, and transformers are able to learn a weighting to select which tokens are relevant to its interpretation.

Many works (Rogers et al., 2020) strive to analyze the representations uncovered by transformers to find out whether they are consistent with models derived from linguistic theories. One of the main analysis methods is the long-distance agreement task popularized by Linzen et al. (2016), which consists in assessing neural networks ability to predict the correct form of a token (e.g. a verb) in accordance with the agreement rules (e.g. its subject). This method has been generalized to other agreement phenomena (Li et al., 2021) and other languages (Gulordava et al., 2018). The concordant conclusions of all these experiments show that transformers are able to learn a 'substantial amount' of syntactic information (Belinkov and Glass, 2019).

If the method of Linzen et al. (2016) makes it possible to show that syntactic information is encoded in neural representations, it does not give any indication on its localization: it is not clear whether the syntactic information is distributed over the whole sentence (as made possible by self-attention) or only in a way consistent with the syntax of the language, i.e. only in the tokens involved in the agreement rules.

This work addresses the question: *where* the syntactic information is encoded in transformer representations.[1] We approach this question from two perspectives, considering the object-past participle agreement in French (Section 2). First, in Section 3, using probing and counter-factual analysis, we try to identify the tokens in which syntactic information is encoded in order to find its localization within the sentence. Second, in Section 4, using a feature selection method, we study the localization of syntactic information within the contextualized representation of tokens.

## 2 The Object-Participle Agreement Task

**Task** We evaluate the capacity of transformers to capture syntactic information, by considering the object-past participle agreement in French object relatives. This task consists in comparing the probabilities a language model assigns to the singular and plural forms of a past participle given the beginning of the sentence. The probability of a past participle form is conditioned on all the words in the *prefix* (the words from the beginning of the sentence up to the antecedent ; see Figure 1 for an example) and the *context* (the words from the antecedent up to and excluding the past participle). Following Linzen et al. (2016) the model is considered to predict the agreement correctly if the form with the correct number has a higher probability

[1]Probing datasets and code available at https://gitlab.huma-num.fr/bli/syntactic-info-distribution



*Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*
*Volume 2: Short Papers*, pages 501 - 507

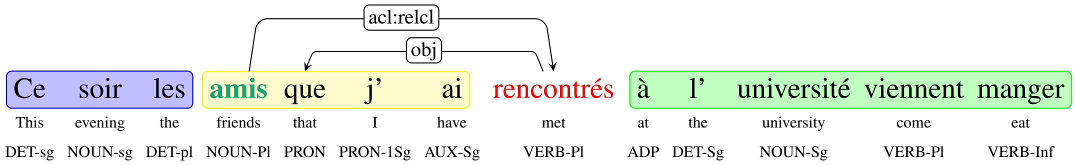


Figure 1: Example of object-past participle agreement in French object relatives. Dependencies between the target verb (in red) and the tokens involved in the agreement rules using the Universal Dependencies annotation guidelines are also shown. The *prefix* is represented in blue, the *context* in yellow and the *suffix* in green. To predict the past participle number, a human is expected to extract number information from the object relative pronoun *(que)* that gets it from its antecedent (*amis* in bold green).

than the form with the incorrect number.

Contrary to the classical subject-verb agreement task (Linzen et al., 2016), the French object past participle agreement involves a filler-gap dependency and the target past participle has to agree with a noun that is never adjacent to it. In our case, it features a syntactic structure that allows us to highlight the way the information is distributed in the sentence (§3.1).

Figure 1 gives an example of the sentences considered here. It involves sentences whose verb is in the compound past *(passé composé)*, a tense composed of an auxiliary and the past participle of the verb. What part of the speech (i.e. the subject, the object or no agreement) the compound verbs must agree with depends on the auxiliary verb used. When the past participle is used with the auxiliary *avoir*, it has to agree in number[2] with its direct object when the latter is placed before it in the sentence. This is notably the case for object relatives considered here, in which the direct object is the relative pronoun *que*, whose number information is the same as its antecedent (even if its morphology–*que*, is the same in singular and plural). To correctly agree the past participle in object relatives, it is therefore necessary to identify the object relative pronoun, its antecedent and the auxiliary.

**Experimental Setting** We reuse the dataset of Li et al. (2021): they have extracted, with simple heuristics a set of 68,497 such sentences after having automatically parsed the Gutenberg corpus with a BERT based dependency parser (Grobol and Crabbé, 2021).

The experiments are carried out with the incremental transformer designed by Li et al. (2021), which was trained on 80 million tokens of French Wikipedia, and has 16 layers and 16 heads. Word embeddings are of size 768. This model is able to predict 93.5% of the past participle agreement, a result that allows these authors to conclude that syntactic information is encoded in the representations.

## 3 Is Syntactic Information Locally or Globally Distributed in the Sentence?

Results reported in the previous section show that information about the number of the past participle is encoded in the token representations but they do not allow to identify which tokens have been used to predict the correct form of the past participle. In this section, we first identify, using linguistic probes, the tokens in which syntactic information is encoded and then, with a causal analysis, the tokens on which transformers mainly rely to predict the form of the past participle.

### 3.1 Probing Experiments

In a first set of experiments, we propose to use linguistic probes to better identify **where in the sentence** the information about the number of the past participle is encoded. A probe is a classifier trained to predict linguistic properties from the language representations: achieving high accuracy at this task implies that these properties were encoded in the representation (Hewitt and Manning, 2019).

More precisely, we label each sentence of our dataset with the number of the target verb (i.e. singular or plural) and consider the task of predicting this label from each token representation of the sentence. We trained one logistic regression classifier per category of word[3] considering 80% of the examples as training data and the remaining 20% as test set.

[2]The past participle must agree in number *and* in gender. For clarity, we will only consider agreement in number.

[3]All classifiers are implemented with the Scikit-Learn library (Pedregosa et al., 2011). See detailed description in Section A of the appendix.

| | Accuracy | | |
|---|---|---|---|
| | correct predictions | wrong predictions | overall |
| *prefix* | $60.2\%_{\pm 0.3}$ | $51.6\%_{\pm 0.5}$ | $59.4\%_{\pm 0.3}$ |
| *context* | $94.6\%_{\pm 0.9}$ | $83.9\%_{\pm 1.4}$ | $94.4\%_{\pm 1.1}$ |
| *suffix* | $72.2\%_{\pm 2.1}$ | $62.1\%_{\pm 2.2}$ | $71.6\%_{\pm 2.1}$ |

Table 1: Mean probing accuracies across different sentence parts (see Figure 1) on two subsets, which differ with respect to whether the transformers correctly or incorrectly predicted the number of the past participle.

Table 1 reports the average accuracy achieved by our probes on different parts of the sentence. We observe that the past participle number information is essentially encoded *locally* within the tokens of the *context* and is not represented uniformly across all the subsequent tokens of the sentence as observed by Klafka and Ettinger (2020).

Indeed, as expected,[4] in the *prefix* (before the antecedent) the performance of the probe mainly reflects the difference between the prior probabilities of the two classes.[5] By contrast, the accuracy becomes high when the tokens of the *context* are considered as input features of the probe, showing that the information required to predict the correct past participle form is spread over all tokens between the antecedent (where the number of the past participle is specified) and the past participle (where the information is 'used'). It is quite remarkable that, as soon as the past participle has been observed and the information on the number of the antecedent is no longer useful, the token representations no longer encode it: in the *suffix* the probe accuracy drops sharply even if it remains better than that observed in the *prefix*. This result contradicts also, at least partially, the observation of Wisniewski et al. (2021) which shows that in a neural translation system, gender information is distributed all over the source and target representations. It should however be noted that this experiment deals with a different kind of information and only considers sentences following a very simple pattern.

To get a more accurate picture of how the number information is distributed within the *context*, we focus on a specific sentence template with a fixed six-word *context*: we only consider sentences in which the antecedent is separated from the relative pronoun by a prepositional phrase made of a preposition and a noun as in the following example:

(1) ... magasin d' habits qu' ils ont vu ...
... store of clothes that they have seen...
... ANTEC-SG ADP NOUN-PL QUE PRON-PL AUX-PL PP-SG ...

This pattern (1,940 sentences) represents 3% of the examples of the original dataset. Note that in these sentences the embedded noun between the antecedent of the object pronoun and the target verb can be an *attractor* noun , i.e. a noun with misleading agreement feature. We trained and tested a separate logistic regression classifier for each position as illustrated by the x-axis labels in figure 2.[6]

We plot in figure 2 the average probing accuracy at different positions of this pattern. In the *prefix* (i.e. b-positions) the probe accuracy is low, except for the position just before the antecedent, which often corresponds to determiners or adjectives that have to agree in number with the antecedent. On the contrary, in the *context*, the predictions of the probe are almost perfect, even when we are probing tokens marked with a number information that is not necessarily related to the number of the past participle (e.g. the auxiliary or the attractor). Accuracy in the *suffix* drops quickly as we move away from the past participle, especially in the presence of an attractor. These observations confirm that the number information is not distributed over all tokens in the sentence as made possible by the self-attention mechanism.

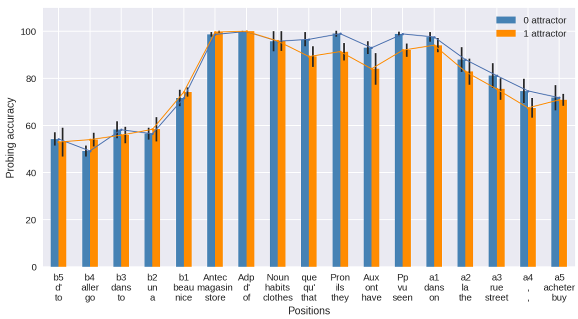


Figure 2: Mean probing accuracy at each position of the six-word *context* pattern. The `bI` (resp. `aI`) position denotes the *I*-th token before (resp. after) the pattern. An attractor occurs at position *Noun* for 1-attractor subset and the agreeing past participle at position *Pp*.

[4] Recall that we are considering an incremental model in which token representations can only depend on the preceding tokens. The following tokens are masked.

[5] In the dataset, 65% of the past participles are singular.

[6] Note that for purpose of clarity, the plot includes tokens of an example sentence. The results are mean accuracies across all test sentences with three different train/test splits.

| Subset | Size (in sentences) | Original | Mask *context* except `Antec que Aux` | Mask `Antec` | Mask `que` | Mask `Antec+que` |
|---|---|---|---|---|---|---|
| Overall | 68,200 | 93.6%$_{\pm 1.2}$ | 85.3%$_{\pm 3.1}$ | 84.0%$_{\pm 2.0}$ | 79.0%$_{\pm 1.0}$ | 76.6%$_{\pm 0.7}$ |
| 0 attractor | 59,915 | 95.4%$_{\pm 0.9}$ | 87.3%$_{\pm 3.0}$ | 87.5%$_{\pm 1.7}$ | 82.9%$_{\pm 0.9}$ | 81.3%$_{\pm 0.6}$ |
| 1 attractors | 7,090 | 82.8%$_{\pm 2.5}$ | 71.3%$_{\pm 3.9}$ | 61.1%$_{\pm 4.2}$ | 53.3%$_{\pm 1.7}$ | 44.6%$_{\pm 1.4}$ |
| 2 attractors | 1,195 | 71.4%$_{\pm 3.3}$ | 68.3%$_{\pm 4.8}$ | 47.0%$_{\pm 4.2}$ | 36.4%$_{\pm 2.1}$ | 27.2%$_{\pm 1.4}$ |

Table 2: Mean accuracies before and after different masking interventions, based on prediction difficulty measured by the number of attractors

### 3.2 Causal intervention on attention

As it stands, we observe that number information is encoded essentially in the *context* part of sentences. Now we test **which** tokens are responsible for providing the number information used to choose the past participle form. To do so, we design a causal experiment in which we mask some tokens of the *context* to better figure out their role in models decision.

**Masking Tokens in Self-Attention Computation** Self-attention is a core component of transformers. In our causal analysis we mask some token representations in the *context* to the self-attention layer. By design, incremental transformers are already masking the end of the sentence with a boolean mask to prevent a token representation to attend to the future tokens. We extend this mechanism to mask, when computing the past participle representation, additional tokens from the sentence prefix such as the antecedent and the relative pronoun.

This intervention allows us to suppress direct access to some tokens such as the antecedent (and thus its number) when building the past participle representation, even if the latter can still access them indirectly: it indeed relies on all other tokens in the sentence for which the mask is kept unchanged. It is then possible, as featured in ablation experiments, to compare performances on the agreement task with and without intervention to evaluate whether the representation of a given token has a direct impact on the prediction of the past participle form.

**Results** Table 2 reports the accuracy on the object-past participle agreement task when some of the tokens in the context are masked. Accuracies are broken down by the number of attractors found in the *context*, a proxy to the difficulty of the prediction (Gulordava et al., 2018). Results show that masking either of the tokens involved in the agreement rule (i.e. the relative pronoun *que* or the antecedent) strongly degrades prediction performance. On the contrary, masking all tokens in *context* except these two and the token before the target verb (generally the auxiliary) has a limited impact on models performance, especially for the most difficult case. This suggests that transformers learn representations that are consistent with the French grammar: the model relies mainly on the same tokens as humans to choose the correct form of the past participle.

## 4 Probing Representations Components

Experiments reported in the previous section show that syntactic information is locally encoded in the *context*. In this section, we address the question of finding **where** this information is encoded **within the transformers representation**. To that end, we repeat the probing experiment on *context* token representations of §3.1 with an $\ell_1$ regularized logistic regression (Tibshirani, 1996). The resulting probe is thus constrained to minimize the number of features used to perform accurate predictions. Given the probe objective function $\sum_{i=1}^{n} - \log P(y_i|\mathbf{x}_i; \mathbf{w}) + \frac{1}{C}||\mathbf{w}||_1$ to minimize, we first determined the lowest bound for C such that the feature coefficients are guaranteed not to be all zeros, from which, we increase C evenly on a log space (i.e. decrease the regularization strength).

**Results** Figure 3 reports the regularization path of the probing classifier. It shows that number information can be extracted with high accuracy (90.1%) solely from a very small number of dimensions, namely 90. Increasing the number of dimensions (by decreasing the regularization strength) only results in a small improvement of model quality: the probe achieves an accuracy of 94.8% when all features are considered. Interestingly, when removing the 90 features selected by the $\ell_1$ regularization from the representation, a probe trained on the remaining features still achieve a very good accuracy of 93.8%, suggesting that the number information is encoded in a redundant way in the contextualised representations.

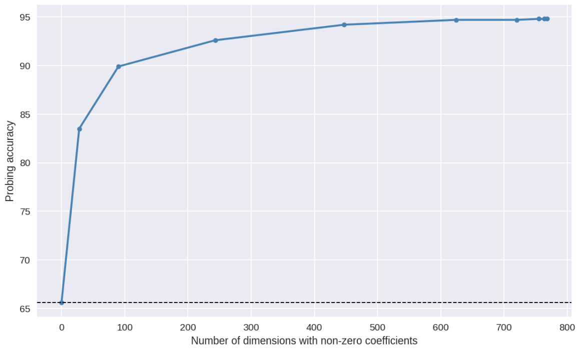


Figure 3: Feature selection by $\ell_1$-logistic regression: probing accuracy of all *context* token representations

## 5 Discussion and conclusion

To understand how syntactic information is encoded and used in transformers-based LM, we carried out three sets of experiments considering the French object-past participle agreement task. First, our probing experiments uncovered clear evidence of a local distribution of number information within the *context* tokens, even though the self-attention mechanism allows this information to be spread all over the sentence. Second, our masking intervention on attention shows a causal link between linguistically motivated tokens and the model's decision, suggesting that transformers process French object-past participle agreement in a linguistically-motivated manner. Finally, we used a $\ell_1$ feature selection method to study the localization of number information within contextualized representations and found that while this information is encoded in a small amount of highly correlated dimensions, it is also fuzzily encoded in a redundant way in the remaining dimensions.

Our work is a first step towards a better understanding of the inner representations of LM. 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 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. We would also like to gratefully acknowledge support from the Labex EFL, "Empirical Foundations in Linguistics" (ANR-10-LABX-0083)

## References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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

Loïc Grobol and Benoit Crabbé. 2021. Analyse en dépendances du français avec des plongements contextualisés (French dependency parsing with contextualized embeddings). In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 106–114, Lille, France. ATALA.

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.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Geoffrey E. Hinton, James L. McClelland, and David E. Rumelhart. 1986. Distributed representations. In *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations*.

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.

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.

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.

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.

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.

R. Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Guillaume Wisniewski, Lichao Zhu, Nicolas Bailler, and François Yvon. 2021. Screening gender transfer in neural machine translation. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 311–321, Punta Cana, Dominican Republic. Association for Computational Linguistics.

## A Probing classifiers

We used a set of logistic regression classifiers to investigate the way the syntactic information is distributed inside the sentences. Each sentence are divided into three parts: *prefix*, *context* and *suffix*, as described in Figure 1. The input for all classifiers are the contextualized token representations built by our pre-trained transformers. We trained one classifier per category of word and per part of the sentences to predict whether the token representation is singular or plural, forcing each probing classfier to specialise on PoS-specific representations of long-distance agreement information. To ensure a fair comparison across parts of sentences, we eliminated the following tokens of PoS tags with less than 100 occurrences: SYM, SCONJ, INTJ, PART, PART and X. Therefore, we have in total 11 categories of tokens in each part of the sentences, resulting in 11*3 probing classifiers, and each classifier is trained with three train/test splits(i.e. random_state $= 0, 20$ and $42$). The averaged results is reported in table 1 of the paper. The detailed results per category of word is in figure 4 below.

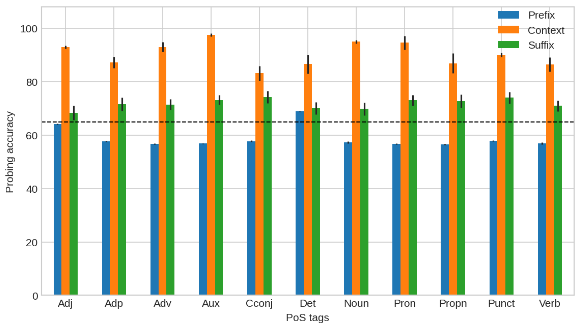


Figure 4: Probing accuracy based on tokens PoS tags and their positions in the sentences, from left to right: *prefix*, *context*, *suffix*