Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles
Avignon, France, 27 juin au 1er juillet 2022

Les représentations distribuées sont-elles vraiment distribuées ? Observations sur la localisation de l'information syntaxique dans les tâches d'accord du verbe en français

Bingzhi Li Guillaume Wisniewski Benoît Crabbé
LLF, CNRS, Université Paris Cité, 75013 Paris, France
bingzhi.li@etu.u-paris.fr,
{guillaume.wisniewski,benoit.crabbe}@u-paris.fr

RÉSUMÉ

Ce travail aborde la question de la localisation de l'information syntaxique qui est encodée dans les représentations de transformers. En considérant la tâche d'accord objet-participe passé en français, les résultats de nos sondes linguistiques montrent que les informations nécessaires pour accomplir la tâche sont encodées d'une manière locale dans les représentations de mots entre l'antécédent du pronom relatif objet et le participe passé cible. En plus, notre analyse causale montre que le modèle s'appuie essentiellement sur les éléments linguistiquement motivés (i.e. antécédent et pronom relatif) pour prédire le nombre du participe passé.

ABSTRACT

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

This work addresses the question of the localization of the syntactic information that is encoded in the transformers representations. Considering the object-past participle agreement in French, the results of our probing experiments show that syntactic information is encoded locally in the representations of tokens between the antecedent and the target past participle. Moreover, the fine-grained causal analysis shows that transformers used the encoded syntactic information in a way consistent with the French grammar.

MOTS-CLÉS : représentation distribuée, transformers, information syntaxique, analyse causale.

KEYWORDS: distributed representation, transformers,syntactic information, causal analysis.

1 Introduction

Les transformers (Vaswani *et al.*, 2017) sont devenus une composante clé dans de nombreux modèles TAL, en grande partie en raison de leur capacité à apprendre à construire des représentations distribuées des mots (Hinton *et al.*, 1986) qui sont *contextualisées* : grâce au mécanisme d'auto-attention multi-têtes (Bahdanau *et al.*, 2015), la représentation d'un mot peut, virtuellement, dépendre de la représentation de tous les autres mots de la phrase, et les transformers sont capables d'apprendre des paramètres permettant de sélectionner pour chaque mot d'une phrase les mots les plus pertinents pour l'interprétation de celui-ci.

De nombreux travaux (Rogers *et al.*, 2021) ont cherché à analyser les représentations apprises par

Volume 1 : conférence principale, pages 384–391.

les transformers pour savoir si elles encodent des informations linguistiques. Une des principales méthodes d'analyse est la tâche d'accord à longue distance popularisée par Linzen *et al.*, 2016, qui consiste à évaluer la capacité d'un modèle neuronal à prédire la forme correcte d'un mot (p.ex. un verbe) en fonction des règles d'accord de la langue (p.ex. son sujet). Cette méthode a été généralisée à d'autres types d'accord (Li *et al.*, 2021) et à d'autres langues (Gulordava *et al.*, 2018). Les conclusions concordantes de toutes ces expériences montrent que les transformers sont capables d'apprendre une quantité « substantielle » d'informations syntaxiques (Belinkov & Glass, 2019).

Si la méthode de Linzen *et al.*, 2016 permet de montrer que l'information syntaxique est encodée dans des représentations neuronales, elle ne donne aucune indication sur sa localisation : il n'est pas clair si l'information syntaxique est distribuée sur l'ensemble de la phrase (comme le permet théoriquement l'auto-attention) ou seulement d'une manière cohérente avec la syntaxe de la langue, c'est-à-dire uniquement dans les mots impliqués dans les règles d'accord.

Ce travail [1] apporte une première réponse à cette problématique et s'intéresse, plus précisément, à deux questions dans le cadre de la tâche de prédiction de l'accord : **où** les informations syntaxiques sont-elles encodées dans les représentations du modèle et **comment** utilise-t-il ces informations syntaxiques. Nous abordons ces questions selon deux perspectives, en considérant l'accord objet-participe passé en français (Section 2). Premièrement, dans la section 3.1, à l'aide des sondes linguistiques (Conneau *et al.*, 2018), nous essayons d'identifier les mots dans lesquels l'information syntaxique est encodée afin de trouver sa localisation dans la phrase. Deuxièmement, dans la section 3.2, nous utilisons une méthode d'analyse contre-factuelle consistant à mesurer l'impact sur la prédiction du modèle d'une modification de l'attention afin de mieux comprendre comment les transformers utilisent les informations syntaxiques encodées dans les représentations contextualisées.

2 Tâche d'accord objet-participe passé

FIGURE 1 – Exemple d'accord objet-participe passé en français dans les relatives objet. Le *préfixe* est surligné en bleu, le *contexte* en jaune et le *suffixe* en vert. Dans la grammaire française, le participe passé s'accorde en nombre avec son complément d'objet direct placé avant l'auxiliaire (ici *que*) et donc avec l'antécédent de celui-ci (ici *amis*).

Tâche Nous considérons l'accord objet-participe passé dans les relatives objet en français pour évaluer la capacité des transformers à capturer l'information syntaxique. Cette tâche consiste à comparer les probabilités qu'un modèle de langue attribue aux formes singulière et plurielle d'un participe passé étant donné la séquence de mots qui le précède dans la phrase, c'est-à-dire les mots du *préfixe* et du *contexte* (cf. Figure 1). Comme Linzen *et al.*, 2016, nous considérons que le modèle a correctement prédit l'accord si la forme correcte a une probabilité plus élevée que la forme incorrecte.

1. Cet article est une version traduite et raccourcie de notre article avec le même titre, accepté à ACL 2022

Contrairement à la tâche classique de l'accord sujet-verbe (Linzen *et al.*, 2016), l'accord objet-participe passé en français implique une dépendance à longue distance (*filler-gap dependency*) et le participe passé cible doit s'accorder avec un nom qui ne lui est jamais adjacent. Dans notre cas, il présente en plus une structure syntaxique qui nous permet de mettre en évidence la façon dont l'information est distribuée dans la phrase(§3.1).

La figure 1 donne un exemple des phrases considérées ici. Il s'agit de phrases dont le verbe est au passé composé. Lorsqu'il est utilisé avec l'auxiliaire *avoir*, le participe passé doit s'accorder en nombre et en genre [2] avec son objet direct que lorsque ce dernier est placé avant lui dans la phrase. C'est notamment le cas des relatives objet considérées ici, dans lesquelles l'objet direct est le pronom relatif *que* qui hérite ses caractéristiques de son antécédent. Pour accorder correctement le participe passé dans les relatives objet, il est donc nécessaire d'identifier le pronom relatif objet, son antécédent et l'auxiliaire.

Cadre expérimental Nous réutilisons le jeu de données de Li *et al.*, 2021 qui a extrait, à l'aide d'heuristiques simples, un ensemble de 68 497 phrases comportant des relatives objet du corpus Gutenberg.

Les expériences [3] sont réalisées avec le modèle transformers incrémental décrit dans Li *et al.*, 2021. Celui-ci a été entraîné sur 90 millions de mots de la Wikipédia française, et possède 16 couches et 16 têtes. Les plongements de mots sont de taille 768. Ce modèle est capable de prédire correctement 93,5% des formes du participe passé du corpus Gutenberg, un résultat qui permet à ces auteurs de conclure que les informations syntaxiques sont encodées dans les représentations du modèle.

3 Les informations syntaxiques sont-elles distribuées localement ou globalement dans la phrase ?

Li *et al.*, 2021 montrent que l'information sur le nombre du participe passé est encodée dans les représentations de mots. Les expériences de ces auteur·e·s ne permettent toutefois pas d'identifier où dans la phrase l'information syntaxique est encodée, une question que nous abordons dans la première partie de cette section. Nous cherchons ensuite à déterminer quels mots sont réellement utilisés par le modèle pour prédire la forme du participe passé.

3.1 Sondes linguistiques

Dans une première série d'expériences, nous proposons d'utiliser des sondes linguistiques pour mieux identifier **où** dans la phrase l'information sur le nombre du participe passé est encodée. Une sonde est un classifieur entraîné à prédire des propriétés linguistiques à partir des représentations du langage (Hewitt & Manning, 2019). Plus précisément, nous associons chaque phrase de notre jeu de données à une étiquette décrivant le nombre du verbe cible et nous considérons la tâche qui consiste à prédire cette étiquette en fonction de la représentation du mot. Nous avons utilisé un classifieur de régression

2. Pour plus de clarté, nous ne considérerons que l'accord en nombre dans nos expériences.
3. Les données et le code sont disponibles à l'adresse suivante : https://gitlab.huma-num.fr/bli/syntactic-info-distribution

	Précision moyennée		
	prédictions correctes	prédictions incorrectes	total
préfixe	$60{,}2\%_{\pm 0{,}3}$	$51{,}6\%_{\pm 0{,}5}$	$59{,}4\%_{\pm 0{,}3}$
contexte	$94{,}6\%_{\pm 0{,}9}$	$83{,}9\%_{\pm 1{,}4}$	$94{,}4\%_{\pm 1{,}1}$
suffixe	$72{,}2\%_{\pm 2{,}1}$	$62{,}1\%_{\pm 2{,}2}$	$71{,}6\%_{\pm 2{,}1}$

TABLE 1 – Précisions obtenues par notre sonde sur différents segments de phrases (voir figure 1) de deux ensembles de test : « prédictions correctes » sont des phrases pour lesquelles le modèle a prédit la bonne forme du verbe cible, « prédictions incorrectes » des phrases pour lesquelles le modèle a assigné une probabilité plus élevée à la forme incorrecte.

logistique [4] par catégorie de mots et par segment de phrase (*préfixe*, *contexte* ou *suffixe* comme définis dans la Figure 1) en considérant 80% des exemples comme des données d'apprentissage et les 20% restants comme des données de test. Le tableau 1 rapporte la précision moyenne obtenue par nos sondes sur différents segments de la phrase. Nous observons que l'information sur le nombre du participe passé est essentiellement encodée *localement* dans la zone *contexte* et n'est pas représentée uniformément à travers tous les mots de la phrase.

En effet, comme prévu, [5] dans le *préfixe* (avant l'antécédent) la performance de la sonde reflète principalement la différence entre les probabilités a priori des deux classes. [6] En revanche, la précision devient élevée lorsque les mots du *contexte* sont considérés comme des caractéristiques d'entrée de la sonde, montrant que les informations requises pour prédire la forme correcte du participe passé est répartie sur tous les mots entre l'antécédent (où le nombre du participe passé est spécifié) et le participe passé (où l'information est « utilisée »). Il est assez remarquable que, dès que le participe passé a été observé et que l'information sur le nombre de l'antécédent n'est plus utile, les représentations de mots ne l'encodent plus : dans le *suffixe*, la précision de la sonde chute fortement même si elle reste meilleure que celle observée dans le cas du *préfixe*.

Pour avoir une idée plus précise de la façon dont les informations sur le nombre sont distribuées dans le *contexte*, nous nous concentrons sur une construction spécifique où l'antécédent est séparé du pronom relatif uniquement par un groupe prépositionnel et que le sujet de la relative est un pronom. Le patron du segment *contexte* correspond donc à la séquence `antécédent ADP NOUN que PRON AUX` comme dans l'exemple suivant :

(1) ... bureaux en métal qu' il a trouvés ...
... ANTEC-PL ADP NOUN-SG QUE PRON-SG AUX-SG PP-PL ...

Cette construction représente 3% des exemples du jeu de données original (1 936 phrases). Il y a un potentiel *distracteur* dans ces phrases : le nom entre l'antécédent du pronom relatif et le participe cible peut avoir un nombre différent du nombre du participe passé (dans exemple (1), le mot « métal »).

La Figure 2 présente la précision des sondes à chaque position [7] de ce motif. Dans la zone du *préfixe*

4. Tous les classifieurs sont implémentés avec la librairie Scikit-Learn (Pedregosa *et al.*, 2011) avec les hyperparamètres par défaut. Nous rapportons la moyenne des précisions sur trois répartitions en train/test avec des graines aléatoires différentes.

5. Nous considérons un modèle incrémental dans lequel la représentation d'un mot ne peut dépendre que des mots précédents, les mots suivant étant masqués.

6. Dans le jeu de données, 65% des participes passés cibles sont au singulier.

7. Pour plus de clarté, les positions sont illustrées par les mots d'une phrase d'exemple. Les résultats sont des précisions

(c.-à-d. pour les positions b_I), les précisions sont relativement faibles, sauf pour les deux positions situées juste avant l'antécédent, correspondant souvent à des déterminants ou des adjectifs qui doivent s'accorder en nombre avec celui-ci. Au contraire, dans le *contexte*, les prédictions sont presque parfaites, même lorsque nous sondons des mots marqués explicitement par un nombre différent de celui du participe passé (p.ex. l'auxiliaire ou le distracteur). Les précisions dans la zone *suffixe* chutent rapidement lorsque l'on s'éloigne du participe passé, surtout en présence d'un distracteur. Ces observations confirment que l'information syntaxique n'est pas distribuée dans tous les mots de la phrase mais essentiellement dans le *contexte*.

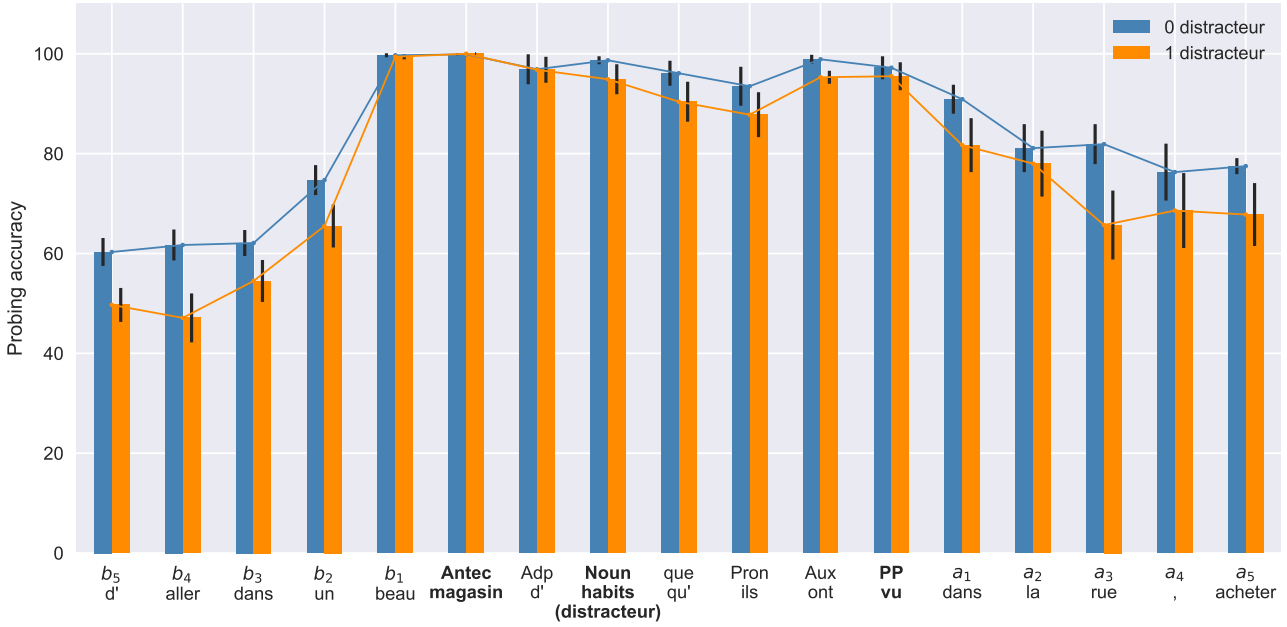

FIGURE 2 – Précisions de sondes à chaque position de phrases du patron fixe. Les positions b_I (resp. a_I) dénotent les I-ème mot avant (resp. après) le patron fixe. Un distracteur se trouve à la position *Noun* pour le sous-ensemble *1 distracteur* comme illustrée par la phrase d'exemple. *0 distracteur* représente des phrases dans les quelles ce nom enchâssé est du même nombre que l'antécédent.

3.2 Intervention causale sur l'attention

Mot masqué	<bos>	Les	cadeau**x**	que	le	directeur	a	accepté·**s** / accepté*
Aucun		-2.8	-9.5	-7.3	-1.8	-6.1	-3.9	**-5.9** / -8.3
que		-2.8	-9.5	-7.3	-1.8	-6.1	-3.9	-13.7 / **-11.9**

TABLE 2 – Exemple de phrase traitée par notre transformer, sans intervention et avec intervention du masquage *que*. Nous rapportons les log-probabilités pour chaque mot des phrases contenant soit la forme plurielle du participe passé cible *acceptés*, ou sa forme singulière *accepté*.

Dans la section précédente, nous avons montré que l'information du nombre est encodée essentiellement dans la partie *contexte* de la phrase. Nous cherchons maintenant à identifier à partir de **quels**

moyennes sur **toutes** les phrases de test et sur trois répartitions différentes en train/test.

Sous-groupes	Taille (en phrases)	Original	Masquer segment *contexte* sauf `Antéc que Aux`	Masquer `Antéc`	Masquer `que`	Masquer `Antéc+que`
total	68 200	$93{,}6\%_{\pm 1,2}$	$85{,}3\%_{\pm 3,1}$	$84\%_{\pm 2}$	$79\%_{\pm 1}$	$76{,}6\%_{\pm 0,7}$
0 distracteur	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 distracteur	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 distracteurs	1 195	$71{,}4\%_{\pm 3,3}$	$68{,}3\%_{\pm 4,8}$	$47\%_{\pm 4,2}$	$36{,}4\%_{\pm 2,1}$	$27{,}2\%_{\pm 1,4}$

TABLE 3 – Précision avant et après les différentes interventions de masquage, selon la difficulté de la prédiction mesurée par le nombre de distracteurs

mots cette information est déterminée. Pour ce faire, nous concevons une expérience causale dans laquelle nous masquons certains mots du segment *contexte* pour mieux déterminer leur rôle dans la prédiction du modèle.

Masquage des mots dans le calcul de l'auto-attention Le mécanisme d'auto-attention permet aux transformers de construire des représentation contextualisées en définissant itérativement (en première approximation) la représentation d'un mot à partir d'une combinaison linéaire des représentations des autres mots de la phrase. Nous proposons de neutraliser la contribution d'un ou plusieurs mots spécifiques dans la construction de la représentation du participe cible en forçant le poids de ce ou ces mots dans cette combinaison linéaire à être zéro. Cette intervention peut être implémentée de manière simple en étendant le mécanisme de masquage utilisé dans les transformers incrémentaux pour interdire à la représentation d'un mot de prendre en compte les mots futurs.

Plus précisément, nous considérons la même tâche d'accord décrite dans la section 2, mais cette fois, lors de la prédiction du participe passé cible (et seulement à ce moment !), nous masquons également soit l'antécédent (et ses dépendants), [8] soit le pronom relatif *que*, soit ces deux mots ou tous les mots du *contexte* sauf ces deux mots. Le tableau 2 fournit un exemple de traitement de phrase avant intervention et après une intervention consistant à masquer *que*. Comme l'intervention n'a eu lieu que lorsque le modèle calculait la représentation du verbe cible, il n'y a aucun effet sur les mots qui le précèdent. Comme on peut le voir dans cet exemple, le transformer attribuait initialement une probabilité plus élevée à la forme plurielle correcte *acceptés* qu'à la forme singulière incorrecte *accepté*. Après l'intervention, la situation est inversée et le modèle prédit la forme singulière (incorrecte).

Cette intervention nous permet de construire des représentations du participe passé qui ne prennent pas en compte certains mots de la phrase et donc de supprimer certaines informations *directes* de la représentation de celui-ci (p.ex. l'information sur le nombre de l'antécédent). Il faut toutefois noter que ces informations peuvent toujours être prises en compte indirectement : la représentation du participe passé s'appuie en effet sur tous les autres mots de la phrase pour lesquels nous ne changeons pas le masque. Il est alors possible, comme pour les expériences d'ablation, de comparer les performances sur la tâche d'accord avec et sans intervention pour évaluer si la représentation d'un ou plusieurs mots spécifiques a un impact direct sur la prédiction de la forme du participe passé.

8. Masquer tous les dépendants prédits par une analyse de dépendances automatique de la phrase nous permet de « cacher » tous les mots avec une indication morphologique du nombre de l'antécédent, comme le déterminant et les adjectifs qui le qualifient.

Résultats La table 3 présente la performance sur la tâche d'accord objet-participe passé lorsque certains mots dans le *contexte* sont masqués. Les précisions sont classées selon le nombre de distracteurs trouvés dans le *contexte*, une approximation de la difficulté de la tâche (Gulordava *et al.*, 2018). Comme dans la majorité des phrases d'évaluation, l'antécédent est le nom le plus proche du participe cible, la précision globale ne permet pas de savoir directement si le modèle est capable d'identifier la dépendance en question. Nous nous focalisons dans notre évaluation sur les sous-ensembles plus difficiles — *1, 2 distracteur(s)*. Les résultats montrent que masquer l'un ou l'autre des deux mots impliqués dans la règle d'accord (le pronom relatif *que* ou l'antécédent avec ses dépendants) dégrade fortement la performance de prédiction. Au contraire, masquer tous les mots dans le *contexte* sauf ceux-ci et le mot précédant le verbe cible (généralement l'auxiliaire) a un impact limité sur les performances, surtout pour le cas le plus difficile. Ceci suggère que les transformers apprennent des représentations qui sont cohérentes avec la grammaire du français.

4 Discussion et conclusion

Pour comprendre comment l'information syntaxique est encodée et utilisée dans des modèles de langue fondés sur les transformers, nous avons réalisé deux séries d'expériences en considérant la tâche d'accord objet-participe passé en français. Premièrement, nos expériences de sondes ont mis en évidence une distribution locale de l'information du nombre dans la partie *contexte* de la phrase, même si le mécanisme d'auto-attention permet à cette information de se propager dans toute la phrase. Deuxièmement, notre intervention de masquage sur l'attention montre un lien causal entre les mots linguistiquement motivés et la décision du modèle, ce qui suggère que les transformers traitent l'accord objet-participe passé en français d'une manière linguistiquement motivée.

Notre travail est un premier pas vers une meilleure compréhension des représentations internes des modèles neuronaux. La conception de nouvelles sondes croisée avec des analyses causales et ses applications à un plus large nombre de langues, pourraient améliorer notre compréhension de ces modèles. En particulier, notre observation sur la distribution de l'information syntaxique linguistiquement motivée dans les représentations des transformers pourrait être étendue à d'autres phénomènes linguistiques et à d'autres langues.

Remerciements

Ce travail a bénéficié d'un accès aux ressources en HPC/IA de l'IDRIS au travers de l'allocation de ressources 2020-AD011012282 et 2021-AD011012408 attribuées par GENCI. Nous tenons également à remercier le Labex EFL (Empirical Foundations of Linguistics), ANR-10-LABX-0083, pour son soutien.

Références

BAHDANAU D., CHO K. & BENGIO Y. (2015). Neural machine translation by jointly learning to align and translate. In Y. BENGIO & Y. LECUN, Éds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

BELINKOV Y. & GLASS J. (2019). Analysis methods in neural language processing : A survey. *Transactions of the Association for Computational Linguistics*, **7**, 49–72. DOI : 10.1162/tacl_a_00254.

CONNEAU A., KRUSZEWSKI G., LAMPLE G., BARRAULT L. & BARONI M. (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)*, p. 2126–2136, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-1198.

GULORDAVA K., BOJANOWSKI P., GRAVE E., LINZEN T. & BARONI M. (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)*, p. 1195–1205, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1108.

HEWITT J. & MANNING C. D. (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)*, p. 4129–4138, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1419.

HINTON G. E., MCCLELLAND J. L. & RUMELHART D. E. (1986). Distributed representations. In *Parallel distributed processing : Explorations in the microstructure of cognition. Volume 1 : Foundations*.

LI B., WISNIEWSKI G. & CRABBÉ B. (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*, p. 4599–4610, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics.

LINZEN T., DUPOUX E. & GOLDBERG Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, **4**, 521–535. DOI : 10.1162/tacl_a_00115.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

ROGERS A., KOVALEVA O. & RUMSHISKY A. (2021). A Primer in BERTology : What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, **8**, 842–866. DOI : 10.1162/tacl_a_00349.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. U. & POLOSUKHIN I. (2017). Attention is all you need. In I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems*, volume 30 : Curran Associates, Inc.