A targeted assessment of the syntactic abilities of Transformer models for Galician-Portuguese

Marcos Garcia and Alfredo Crespo-Otero

CiTIUS - Centro Singular de Investigación en Tecnoloxías Intelixentes
Universidade de Santiago de Compostela, Galiza
marcos.garcia.gonzalez@usc.gal, alfredo.crespo@rai.usc.es

Abstract. This paper presents a targeted syntactic evaluation of Transformer models for Galician-Portuguese. We defined three experiments that allow to explore how these models, trained with a masked language modeling objective, encode syntactic knowledge. To do so, we created a new dataset including test instances of number (subject-verb), gender (subject-predicative adjective), and person (subject-inflected infinitive) agreement. This dataset was used to evaluate monolingual and multilingual BERT models, controlling for various aspects such as the presence of attractors or the distance between the dependent elements. The results show that Transformer models perform competently in many cases, but they are generally confounded by the presence of attractors in long-distance dependencies. Moreover, the different behavior of monolingual models trained with the same corpora reinforces the need for a deep exploration of the network architectures and their learning process.

Keywords: Language models · Syntax · Targeted syntactic evaluation.

1 Introduction

The use of modern artificial neural networks gave rise to significant improvements on most NLP tasks [4,23], many of them requiring deep linguistic knowledge, such as machine translation [6] or natural language understanding [21]. This great performance of deep neural networks, together with the fact that they learn from text with no linguistic annotation, has provoked the interest of researchers from different areas, including linguistics and cognitive science. In this regard, there have been several studies that explore how different neural network architectures capture linguistic —mainly syntactic— knowledge [16,11,8].[1]

One of the most prevalent experiments to analyze the syntactic generalizations induced by artificial neural networks is the agreement prediction task, which evaluates whether a model is able to learn a hierarchical morphosyntactic dependency. Therefore, if in a sentence like

> "O *rapaz*$_i$ que jogava com as suas *amigas*$_j$ *está*$_i$| **estão*$_j$ bem."
> 'The *boy*$_i$ who was playing with his *friends*$_j$ *is*$_i$| **are*$_j$ happy.'

a model gives a higher probability to the singular form ("está", 'is') than to the plural ("estão", 'are') it may be an indication that the network is using the hierarchical (syntactic) structure of natural languages instead of a linear one (which would establish an agreement between "amigas" and "estão", both in plural). In this example, the noun "amigas" is used as an *attractor* which may confound the model's behavior with respect to the prediction of the subsequent verb form [3,11].

Inspired by this type of analysis, a new line of research, often dubbed Target Syntactic Evaluation (TSE) [18], has recently emerged defining a variety of analytical probes and releasing datasets in various languages (although mainly in English). In this respect, some authors train *ad-hoc* long short-term memory networks (LSTMs) to observe whether they can generalize syntactic knowledge from raw text [16], while others assess whether models trained on generic language modeling objectives induce syntactic structures [11].

However, to the best of our knowledge, there is no such syntactic evaluation of Language Models (LMs) for Galician-Portuguese.[2] This paper presents a TSE of Galician and Portuguese models based on Transformer [25], one of the best-performing architectures for NLP. We created a dataset to evaluate number, gender, and person agreement dependencies, using instances of subject-verb (e.g., "O *rapaz* [. . .] *está*| **estão*"), subject-predicative adjective (e.g., "A *rapariga* [. . .] é *alta*| **alto*"), and subject-inflected infinitive (e.g., "Preparei a carne para *tu*| **ele*| **nós*|. . . a *comeres*"), respectively. We evaluate monolingual and multilingual models, showing that they behave competently, especially for number and gender, but are unstable regarding the person agreement. Furthermore, the results call for an in-depth analysis of the network architecture and learning process, as different models trained with the same corpus show opposite trends.

The rest of the paper is structured as follows: Section 2 presents related work on the evaluation of the syntactic abilities of neural language models. Then, Section 3 introduces the experiments and the dataset, while the results are discussed in Section 4. Finally, the conclusions of this study are drawn in Section 5.

[1] See [15] and [2] for a review on the syntactic evaluation of neural networks, and on its relation to theoretical linguistics, respectively.

[2] Galician and Portuguese are usually considered varieties of a single language [14,7], but the recent standardization of the former adopting a Spanish-based orthography [22] makes difficult to process it using resources and tools built for Portuguese. Thus, our division of Galician and Portuguese is based solely on their different spellings.

2 Related Work

In their seminal paper, Linzen *et alii* assessed whether LSTMs (a type of recurrent neural networks, RNNs) capture syntactic structure from sequential data in English [16]. They presented the *number prediction task*, where a model should encode both the number and the 'subjecthood' between a long-distance subject-verb dependency (e.g., "The *keys* to the cabinet **is|are* on the table"), and found that, while a generic language modeling objective is not enough to generalize syntax, supervised LSTMs adequately identify this type of structures. However, a subsequent study analyzing 4 languages (Italian, English, Hebrew, and Russian) showed that carefully constructed LSTM-based LMs are able to induce syntactic generalizations in the number prediction task, with only a moderately lower performance than humans, and performing well in non-sensical sentences [11]. Following this path, the authors of [18] presented a test set for TSE in English, including not only instances of subject-verb agreement but also other syntactic phenomena. Several experiments using both RNN language models and syntax-aware supervised RNNs showed that despite performing well in various scenarios, the models' performance is far from that of human annotators.

The impressive results obtained by Transformer-based BERT models [5] in most NLP tasks also aroused the interest in exploring the syntactic knowledge induced by the self-attention mechanism of this architecture [25]. Using previous subject-verb agreement data in English, the results presented in [10] suggest that the non-recurrent architecture of BERT is able to induce long-distance syntactic agreement. Then, Mueller *et alii* evaluated both LSTM and BERT models in a multilingual scenario [19], using a cross-linguistic dataset of subject-verb agreement in English, French, German, Hebrew, and Russian. Their results show that the models use the cues provided by morphologically-rich languages to learn syntactic generalizations and that multilingual models do not seem to transfer syntactic information across languages.

More recently, [20] discussed different approaches to TSE other than using manually selected verb pairs (e.g., "is|are"), while Hall *et alii* [12] questioned the claims about the syntactic generalization capabilities of current LMs, after obtaining lower results on semantically non-sensical sentences.

In this paper, we follow this line of research and create a new dataset to evaluate number, gender, and person agreement in Galician-Portuguese (Gl and Pt). This manually created dataset includes tens of target pairs, and contains lexical variants to minimize the impact of collocational or statistical cues.

3 Materials and Methods

Here we present the experiments, data, and models used in the evaluation.

3.1 Experiments and data

We performed three experiments to explore how Transformer models identify the dependency between a subject and its syntactic head, focusing on the following morphosyntactic features: number and gender, on the one hand, and an additional experiment on the person feature using the inflected infinitive, on the other. Gender and number were evaluated in Gl and Pt, while person agreement was only tested in Galician. The latter decision was made mainly because verbal agreement presents a large variation in Brazilian varieties (e.g., second person pronouns can agree with verb forms in both second and third person) [17], and this may complicate the analysis and interpretation of the results.[3]

To create the dataset, we selected as target (masked) words only those forms appearing in the vocabulary of both monolingual and multilingual models, so that the evaluations of all models can be done on the same number of instances.[4]

Number agreement: For number agreement, we use subject-verb sentences with a relative clause, e.g., "O *rapaz*$_i$ que jogava com as *raparigas*$_j$ *está*$_i$|$*$*estão*$_j$ bem" ('The boy who played with the girls is|$*$are fine'), where we mask the main verb ("está"), which should have the same number as the subject of its clause ("rapaz"). For each of the verbs with singular and plural forms in the mentioned vocabularies (26 for Galician, and 18 for Portuguese), we created a simple sentence with an embedded relative clause, generating new instances with the following conditions: (i) singular and plural subjects (e.g., "o rapaz", "os rapazes"); (ii) masculine and feminine subjects (e.g., "a(s) rapariga(s)"); (iii) 3 variants for each subject (e.g., "a moça", "o menino", etc.); (iv) an attractor, both in masculine and feminine, with a different number (which may confound a sequential model) as the last noun of the relative clause (e.g., "o menino" vs. "os meninos" and vs. "as meninas");[5] (v) 3 variants of the attractors (as in subject); and (vi) sentences with a longer dependency (e.g., "O rapaz que jogava ontem no parque que foi inaugurado recentemente [...]"). This allowed us to create a total of 4,368 (Gl) and 3,024 (Pt) test items.

Gender agreement: Here we evaluate the gender agreement between the subject and a predicative adjective, e.g., "Os *rapazes*$_i$ que jogavam com as *raparigas*$_j$ são *altos*$_i$|$*$*altas*$_j$" ('The boys who played with the girls are tall$_{\text{Masc}}$|$*$tall$_{\text{Fem}}$'), where we mask the adjective ("altos"). As in the first scenario, we assess the impact of attractors, and of the distance of the dependency relation. We used all adjectives with both masculine/feminine and singular/plural forms in the vocabulary (e.g., "alto|alta|altos|altas"), totaling 22 for Galician and 23 for Portuguese. Besides, we generated the same sentence variants as in the first experiment, adapted to gender instead of number agreement. However, in this case, we did not use the number variation for attractors (i.e., both the subject and the attractor have the same number), as the verb inflection would behave as a cue (e.g., "O *rapaz*$_i$ [...] *as raparigas*$_j$ é$_{\text{i}}$|são$_{\text{j}}$ `masked adjective`$_{\text{i}}$"). Therefore,

[3] As this variation hardly exists in European Portuguese, the analysis of the person feature can be easily done for this variety, and we leave this for further work. It is worth mentioning, however, that most neural language models for Portuguese are trained using large amounts of Brazilian data.

[4] Data available at `https://github.com/crespoalfredo/PROPOR2022-gl-pt`

[5] Note that we also included sentences without attractors to observe their impact.

this subset is smaller than the one used for number agreement, having a total of 2,112 instances in Galician, and 2,304 instances in Portuguese.

Person agreement: We observe if the models identify the person (and number) agreement between an inflected infinitive and its subject, e.g., “Preparei a carne para *vós*$_i$| *eu/*nós. . . a *comerdes*$_i$” (‘I prepared the meat for you$_{2^{\text{nd}}\text{Pl}}$|*us. . . to eat$_{2^{\text{nd}}\text{Pl}}$’), masking the subject (“vós”) of the infinitive (“comerdes”). To create this subset we avoided the 1$^{\text{st}}$ and 3$^{\text{rd}}$ person singular pronouns (“eu”, “el|ela”, “vostede”), as the inflected infinitive has the same form as the non-inflected one in these persons. Then, we selected the nominative pronouns which appear in the vocabulary of mBERT: 2$^{\text{nd}}$ person singular (“ti|tu”), 1$^{\text{st}}$ person plural (“nós”), and 3$^{\text{rd}}$ person plural (“eles|elas”). We chose 27 verbs and created the following variants: (i) long and short contexts (which here do not modify the distance between the target dependency); (ii) 2 tenses of the main verb (past and future, e.g., “Preparei|Prepararei”); (iii) 2 persons of the main verb (1$^{\text{st}}$ and 3$^{\text{rd}}$ singular, e.g., “Preparei|Preparou”); (iv) position of the masked pronoun (before/after the infinitive, e.g., “[. . .] para a *comerdes vós*”). This subset has 1,296 instances (Gl).

Table 1. Average sizes (in number of tokens) of the short and long contexts. *Number* and *Gender* include the distances between the dependent elements, where *No* and *Att* refer to contexts without and with attractor, and *Mi* and *Ma* are micro-average and macro-average values, respectively. For person, the values are the length of the sentences, as the distance between the target elements is the same in both contexts.

	Galician								**Portuguese**							
	Short				**Long**				**Short**				**Long**			
	No	*Att*	*Mi*	*Ma*	*No*	*Att*	*Mi*	*Ma*	*No*	*Att*	*Mi*	*Ma*	*No*	*Att*	*Mi*	*Ma*
Number	4.0	7.1	6.6	5.5	6.1	9.1	8.7	7.6	4.1	8.1	7.5	6.1	6.1	10.1	9.5	8.1
Gender	5.1	7.0	6.5	6.1	9.1	11.0	10.5	10.1	5.1	7.9	7.2	6.5	9.0	11.7	11.0	10.3
Person	3.0				9.5				—				—			

Table 1 includes the size of short and long contexts in all subsets. For number and gender, it shows the distances between the dependent elements with and without attractors, along with their averages. For the person agreement task we show the sentence length, as the distance between the target dependent elements is the same in both contexts.

3.2 Models

We used the official multilingual BERT model (base cased, mBERT) [5] as a baseline, which was trained on Wikipedia’s of 101 languages (including Galician and Portuguese), with a cross-lingual vocabulary of 119,547 tokens. Besides, we evaluated the following BERT models (using the *Transformers* library [27]):

Galician: We used the *base* (12 layers) and *small* (6 layers) models described in [9], with 119,547 and 52,000 cased tokens, respectively.[6]

Portuguese: We evaluated the *large* (24 layers) and *base* (12 layers) variants of BERTimbau [24], both of them with a vocabulary size of 30,000 cased tokens.

It is worth noting that the monolingual models were trained with the same corpora for each language, so that differences in the results will be due to the architecture of the networks, or to the different vocabularies of the Galician LMs.

3.3 Evaluation

We performed the most common method for TSE in masked language models [10,19], which involves computing the accuracy by selecting, from the alternatives provided by each instance in the dataset, the one to which the model assigns the highest probability. In the number and gender agreement dependencies, each example includes a pair of alternatives (correct|incorrect, e.g., "is|are"), while for person we may have one or two alternatives as correct, and all the other nominative pronouns as incorrect. For instance, in the following sentence:

"Preparei a carne para a $\textit{comeren}_{3^{\text{rd}}\text{Plur}}$ $\textit{eles}_{3^{\text{rd}}\text{PlurMasc}}|\textit{elas}_{3^{\text{rd}}\text{PlurFem}}$."[7]

in which the inflected infinitive ("comeren") is in the 3rd person plural, both masculine and feminine pronouns with the same person are correct, while all the other nominative pronouns are wrong. We compare the sum of the probabilities of both classes (correct and incorrect), and select the highest one.

4 Results and Discussion

Table 2 shows the average results of the three experiments. In Galician, the models followed the same trend in all cases, with BERT-base obtaining the higher results followed by the small and multilingual models. However, in Portuguese, BERT-base has on average slightly better performance than the large model, and mBERT obtained competitive results. Overall, the accuracy values for number and especially for gender are markedly higher than for person agreement (in Gl).

Number agreement: Figure 1 compares the impact of the attractors in both short and long dependency contexts for number agreement. The first two columns of each model indicate that the attractor produces a low effect in short dependencies (except for mBERT), suggesting that in these cases the model may be identifying the relation between the subject and the verb. The comparison between short and long contexts without attractors (columns 1 and 4) reinforces this finding, as there are no remarkable differences between these values in the monolingual models. However, columns 4 and 5 in Figure 1 indicate that the

[6] We also evaluated the Bertinho models [26], with lower results not discussed here.

[7] "Vostedes" (formal pronoun in the 2nd person plural, agreeing with the 3rd person plural) is not used as it does not appear in the mBERT vocabulary.

Table 2. Micro-average accuracy in the whole dataset for Galician and Portuguese.

	Galician			Portuguese		
	BERT-base	**BERT-small**	**mBERT**	**BERT-large**	**BERT-base**	**mBERT**
Number	0.97	0.95	0.80	0.90	0.91	0.88
Gender	0.98	0.97	0.94	0.95	0.96	0.96
Person	0.73	0.65	0.32	—	—	—

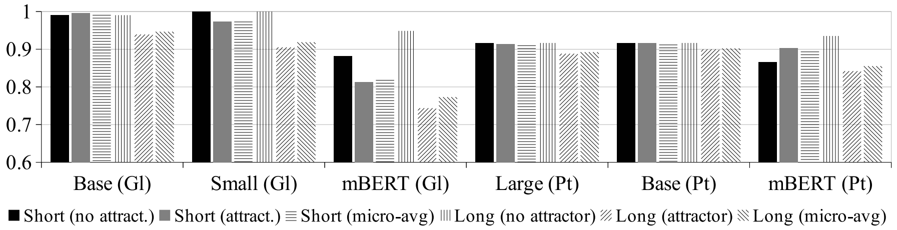

Fig. 1. Accuracy on number agreement in Galician (Gl, left) and Portuguese (Pt, right) for short and long dependencies vs. presence/absence of attractors. For each model, we show six results: first representing short contexts (three bars columns: without attractor, with attractor, and micro-average), and then for long dependency distances.

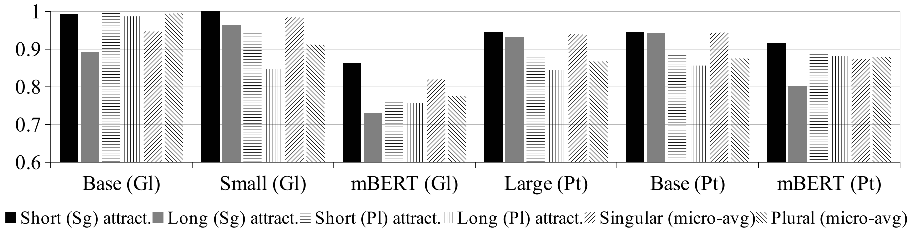

Fig. 2. Accuracy on number agreement in Galician (Gl, left) and Portuguese (Pt, right) for short and long dependencies vs. number of the main verb. There are 6 results per model: the first pair of columns display singular and plural in short dependencies with attractors; then, the same for plural number; last two columns are the micro-average results (with and without attractors) for singular and plural.

performance of the models (especially mBERT and the Galician monolingual models) in long-distance dependencies is affected by the presence of attractors.

The results in Figure 2 allow us to compare the performance of the models with respect to the number of the verb of the main clause. The last two columns of BERT-base (Gl) show that it has some bias towards the plural, while the other models (in Gl and Pt) obtain better results with the singular number. Again, this variation is higher in long-distance dependencies (except for mBERT).

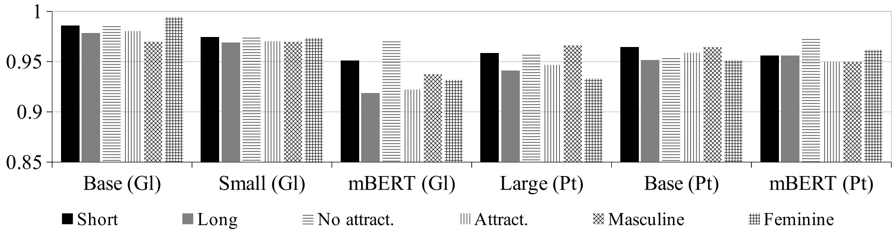

Fig. 3. Accuracy on gender agreement (Gl at left, Pt at right). Each pair of columns on each model displays, respectively, the following scenarios: short and long-distance dependencies; absence and presence of attractors; masculine and feminine subjects.

Gender agreement: Overall, the results of gender agreement are higher than those of number, also for the multilingual model (Figure 3). As expected, the results are slightly lower in long-distance dependencies (columns 1 and 2), and without attractor (columns 3 and 4), except for BERT-base in Portuguese. Regarding the gender of the target relation, BERT-base (Gl) and BERT-large (Pt) show low biases (0.03) towards feminine and masculine, respectively, while the other models seem more stable.

Person agreement: Concerning subject-inflected infinitive agreement, the results in Table 3 show that BERT-base performs noticeably better in the 1st plural, and similarly in the two other cases. Nevertheless, the small model produced very similar results for the 2nd singular and 1st plural, obtaining higher results in the 3rd person plural. mBERT had relatively similar results in the 1st and 3rd persons of the plural, and an extremely low accuracy in the 2nd singular (0.003, with only 3 correct answers out of 960 instances). Even though further analyses are needed to understand the variation between the base and small models (which were trained on the same corpora), the performance of mBERT on the 2nd singular may be due to the low frequency of this person in writing [1], particularly in the Wikipedia corpus used to train this model.

Table 3. Accuracy vs. person of the inflected infinitive (and its subject) in Galician.

	BERT-base	**BERT-small**	**mBERT**
2nd Sing	0.61	0.55	0.00
1st Plur	0.93	0.57	0.42
3rd Plur	0.66	0.81	0.53

Finally, Figure 4 shows pairs of columns to allow for comparisons in the following scenarios: (a) short and long sentences, (b) subject-verb order, and (c) person of the subject of the main verb of the sentence. These variations are in general lower in mBERT, probably due to the low performance of this model in

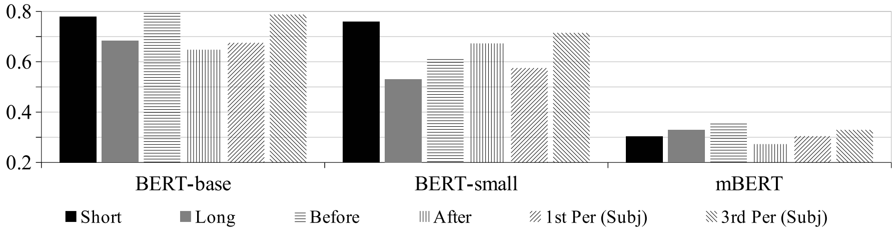

Fig. 4. Accuracy vs. sentence properties in the person agreement test. The variation regarding the tense of the main verb is not shown as it is marginal (average of 0.004).

this experiment. Regarding the length of the sentence, both monolingual models perform notoriously better in short contexts, even though context length does not affect the distance of the target dependency. BERT-base (and mBERT) obtains better results at predicting the subject pronoun when it occurs before the inflected infinitive, while the small model behaves in the opposite way. Furthermore, the models' performance also seems to be affected by the properties of the main clause verb —which does not affect morphosyntactically the target relation—, as all the models have higher accuracies with a 3^{rd} singular person. These results may be due to the higher frequency of the 3^{rd} singular person in writing, but further research is needed to explain this behavior.

In summary, these results show that Transformer models, especially monolingual ones, generalize number and gender agreement even in long-distance dependencies, or with the presence of attractors. However, both conditions seem to mislead the models, whose performance drops significantly in this scenario.

5 Conclusions and Further Work

This paper has presented an evaluation of the syntactic capabilities of Transformer models for Galician-Portuguese. The results of the three experiments conducted here, analyzing number, gender, and person features, show that monolingual and multilingual models perform well on number and gender agreements, while person, assessed using inflected infinitives, seems harder to generalize.

Even if the performance in this last evaluation may be influenced by the relatively low frequency of the inflected infinitive in Galician corpora, the differences of the monolingual models (trained on the same corpora) in the three experiments suggest that the models' architecture is crucial, as previous studies have shown [11,13]. In this regard, Baroni proposes careful analyses of the network architectures, treating them as algorithmic linguistic theories instead of empty devices with no priors [2].

In future work, we plan to extend the dataset to include more syntactic phenomena (including evaluations of the same features using different linguistic structures), naturally occurring sentences from corpora, and also semantically non-sensical —but syntactically well-formed— examples. Moreover, we intend to perform further analyses and evaluations that allow to compare our results with those of other languages in similar scenarios, aimed at gaining new knowledge about the grammatical competence of these models.

Acknowledgments

This research is funded by a *Ramón y Cajal* grant (RYC2019-028473-I), by the Galician Government (ERDF 2014-2020: Call ED431G 2019/04, and ED431F 2021/01), and by a summer internship of the CiTIUS Research Center.

References

1. Ariel, M.: The development of person agreement markers: From pronouns to higher accessibility markers. Usage-based models of language pp. 197–260 (2000)
2. Baroni, M.: On the proper role of linguistically-oriented deep net analysis in linguistic theorizing (2021), arXiv preprint arXiv:2106.08694
3. Bock, K., Miller, C.A.: Broken agreement. Cognitive psychology **23**(1), 45–93 (1991)
4. Collobert, R., Weston, J.: A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th international conference on Machine learning. pp. 160–167 (2008)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1. pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019)
6. Edunov, S., Ott, M., Auli, M., Grangier, D.: Understanding back-translation at scale. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 489–500. Association for Computational Linguistics, Brussels, Belgium (2018)
7. Freixeiro Mato, X.R.: Gramática da Lingua Galega IV. Gramática do texto. A Nosa Terra, Vigo (2003)
8. Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., Levy, R.: Neural language models as psycholinguistic subjects: Representations of syntactic state. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1. pp. 32–42. Association for Computational Linguistics, Minneapolis, Minnesota (2019)
9. Garcia, M.: Exploring the representation of word meanings in context: A case study on homonymy and synonymy. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 3625–3640. Association for Computational Linguistics, Online (Aug 2021)
10. Goldberg, Y.: Assessing BERT's Syntactic Abilities (2019), arXiv preprint arXiv:1901.05287
11. Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., Baroni, M.: 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. pp. 1195–1205. Association for Computational Linguistics, New Orleans, Louisiana (2018)

12. Hall Maudslay, R., Cotterell, R.: Do syntactic probes probe syntax? experiments with jabberwocky probing. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 124–131. Association for Computational Linguistics (2021)
13. Hu, J., Gauthier, J., Qian, P., Wilcox, E., Levy, R.: A systematic assessment of syntactic generalization in neural language models. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 1725–1744. Association for Computational Linguistics (2020)
14. Lindley Cintra, L.F., Cunha, C.: Nova Gramática do Português Contemporâneo. Livraria Sá da Costa, Lisbon (1984)
15. Linzen, T., Baroni, M.: Syntactic structure from deep learning. Annual Review of Linguistics **7**, 195–212 (2021)
16. Linzen, T., Dupoux, E., Goldberg, Y.: Assessing the ability of LSTMs to learn syntax-sensitive dependencies. Transactions of the Association for Computational Linguistics **4**, 521–535 (2016)
17. Lucchesi, D., Baxter, A., da Silva, J.A.A.: A concordância verbal. In: O português afro-brasileiro, pp. 331–371. SciELO Books (2009)
18. Marvin, R., Linzen, T.: Targeted syntactic evaluation of language models. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 1192–1202. Association for Computational Linguistics, Brussels, Belgium (2018)
19. Mueller, A., Nicolai, G., Petrou-Zeniou, P., Talmina, N., Linzen, T.: Cross-linguistic syntactic evaluation of word prediction models. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 5523–5539. Association for Computational Linguistics (Jul 2020)
20. Newman, B., Ang, K.S., Gong, J., Hewitt, J.: Refining targeted syntactic evaluation of language models. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 3710–3723. Association for Computational Linguistics (2021)
21. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving Language Understanding by Generative Pre-Training (2018), `https://openai.com/blog/language-unsupervised`
22. Samartim, R.: Língua somos: A construção da ideia de língua e da identidade coletiva na Galiza (pré-)constitucional. In: Actas do IX Congreso Internacional de Estudos Galegos. Novas achegas ao estudo da cultura galega II: enfoques sociohistóricos e lingüístico-literarios. pp. 27–36. Universidade da Coruña (2012)
23. Schnabel, T., Labutov, I., Mimno, D., Joachims, T.: Evaluation methods for unsupervised word embeddings. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 298–307. Association for Computational Linguistics, Lisbon, Portugal (2015)
24. Souza, F., Nogueira, R., Lotufo, R.: BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In: Cerri, R., Prati, R.C. (eds.) Intelligent Systems. pp. 403–417. Springer International Publishing, Cham (2020)
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need (2017), arXiv preprint arXiv:1706.03762
26. Vilares, D., Garcia, M., Gómez-Rodríguez, C.: Bertinho: Galician BERT Representations. Procesamiento del Lenguaje Natural **66**, 13–26 (2021)
27. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C.,

Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics (2020)