Language Resources and Evaluation
https://doi.org/10.1007/s10579-023-09664-1

ORIGINAL PAPER

Assessing linguistic generalisation in language models: a dataset for Brazilian Portuguese

Rodrigo Wilkens[1] · **Leonardo Zilio**[2] · **Aline Villavicencio**[3]

Accepted: 2 May 2023

Abstract

Much recent effort has been devoted to creating large-scale language models. Nowadays, the most prominent approaches are based on deep neural networks, such as BERT. However, they lack transparency and interpretability, and are often seen as black boxes. This affects not only their applicability in downstream tasks but also the comparability of different architectures or even of the same model trained using different corpora or hyperparameters. In this paper, we propose a set of intrinsic evaluation tasks that inspect the linguistic information encoded in models developed for Brazilian Portuguese. These tasks are designed to evaluate how different language models generalise information related to grammatical structures and multiword expressions (MWEs), thus allowing for an assessment of whether the model has learned different linguistic phenomena. The dataset that was developed for these tasks is composed of a series of sentences with a single masked word and a cue phrase that helps in narrowing down the context. This dataset is divided into MWEs and grammatical structures, and the latter is subdivided into 6 tasks: impersonal verbs, subject agreement, verb agreement, nominal agreement, passive and connectors. The subset for MWEs was used to test BERTimbau Large, BERTimbau Base and mBERT. For the grammatical structures, we used only BERTimbau Large, because it yielded the best results in the MWE task. In both cases, we evaluated the results considering the best candidates and the top ten candidates. The evaluation was done both automatically (for MWEs) and manually (for grammatical structures). The results obtained for MWEs show that BERTimbau Large surpassed both the other models in predicting the correct masked

✉ Rodrigo Wilkens
rodrigo.wilkens@uclouvain.be

Leonardo Zilio
leonardo.zilio@fau.de

Aline Villavicencio
a.villavicencio@sheffield.ac.uk

1 Université catholique de Louvain, Leuven-la-Neuve, Belgium

2 Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

3 University of Sheffield, Sheffield, UK

Published online: 02 June 2023

element. However, the average accuracy of the best model was only 52% when only the best candidates were considered for each sentence, going up to 66% when the top ten candidates were taken into account. As for the grammatical tasks, the results presented better prediction, but also varied depending on the type of morphosyntactic agreement. On the one hand, cases such as connectors and impersonal verbs, which do not require any agreement in the produced candidates, had precision of 100% and 98.78% among the best candidates. On the other hand, tasks that require morphosyntactic agreement had results consistently below 90% overall precision, with the lowest scores being reported for nominal agreement and verb agreement, both having scores below 80% in overall precision among the best candidates. Therefore, we identified that a critical and widely adopted resource for Brazilian Portuguese NLP presents issues concerning MWE vocabulary and morphosyntactic agreement, even if it is prolific in most cases. These models are a core component in many NLP systems, and our findings demonstrate the need of additional improvements in these models and the importance of widely evaluating computational representations of language.

Keywords Grammatical structures · Multiword expression · Intrinsic evaluation · Brazilian Portuguese · Contextualised word embeddings · Language models

1 Introduction

Learning computational representations that accurately model language is a critical goal of Natural Language Processing (NLP) research. These representations, or word embeddings, are very helpful for language technology tasks like machine translation, question answering and text adaptation. In addition, they can improve our understanding of human cognition, for instance, by correlating well with human predictions about words (Mandera et al., 2017; Schrimpf et al., 2020), by helping to highlight the existence of explicit and implicit social constructs, including gender and racial bias (Bolukbasi et al., 2016; Kumar et al., 2020), or by helping to detect misinformation (Oshikawa et al., 2020; Su et al., 2020) and cultural differences (Savoldi et al., 2021; Dinan et al., 2020). With the recent popularisation of NLP and the availability of large-scale computing resources, we have witnessed a quick evolution of many new computational architecture proposals for language representation. As a consequence, there are currently pre-trained language models available in an unprecedented scale, for over a hundred languages (Devlin et al., 2018), including low-resourced ones. However, these models lack transparency and interpretability (Marcus, 2020) and are often seen as black boxes, all of which limit their application in tasks that require explainability. This fast-paced evolution raises the need for careful assessment of the kinds of information that a representation incorporates, be it in terms of linguistic, common-sense, or world information. Moreover, the evaluation procedures need to adopt standard protocols that are independent of the proposed architecture and applicable to classes of models in general, setting the standards for what we expect them to have learned and be proficient in.

In fact, much recent effort has been devoted to defining evaluation protocols that allow us to have an insight into the knowledge that the model incorporates (Ettinger,

2020; Warstadt et al., 2020; Vulić et al., 2020). These evaluations may target different phenomena, but their main distinction is between intrinsic and extrinsic evaluations. Intrinsic evaluations usually involve comparing the predictions made by a model with human judgements or with existing resources like WordNet (Miller et al., 1990) or psycholinguistic norms. Extrinsic evaluations are often based on applications (Bakarov, 2018) and examine the ability of embeddings to be used as the feature vectors of supervised machine learning algorithms for tasks like natural language entailment and question answering. The assumption is that better quality and accurate embeddings lead to better quality applications. Recently, the BERT (Devlin et al., 2018) architecture [and its variations, such as RoBERTa (Liu et al., 2019), DistilBERT (Sanh et al., 2019), XLNet (Yang et al., 2019) and ALBERT (Lan et al., 2019)] has been used as a language encoding component in several applications that enhanced the state of the art in several NLP fields. These developments can be seen as an extrinsic evaluation of the BERT architecture, as they seem to suggest that BERT more successfully encodes syntactic and semantic information than alternative models. However, what these evaluations do not indicate is which specific information is encoded, and if it is used for a given task in ways that are compatible with the use of language by humans. In this paper we focus on intrinsic evaluations with an aim at gaining a better understanding of how linguistically proficient these models really are.

A wide variety of intrinsic evaluation protocols is needed to cover various levels of linguistic description and phenomena, such as those targetting agreement and negation (Warstadt et al., 2020; Ettinger, 2020). This is particularly important to assess the extent to which a model is able to handle idiosyncrasies and preferences in language, like those involved in multiword expressions (MWEs), especially idiomatic MWEs. They have provided a rich and varied test ground as they range from more productive to idiosyncratic cases that require knowledge that goes beyond information about individual words and their compositional combinations (Cordeiro et al., 2019). In this paper, evaluation also targets MWEs of varying degrees of idiomaticity and frequency.

For different languages, the stage of development and proficiency of the models varies considerably, and so do the datasets available for evaluation. For Portuguese, in particular, only a few evaluation initiatives are available, such as Souza et al. (2020), Abdaoui et al. (2020), and Schneider et al. (2020)[1]. Despite being valuable evaluation contributions to NLP in Portuguese, these correspond to only extrinsic evaluations and target very specific tasks. Overall, the Portuguese versions of BERT have not been subjected to many intrinsic evaluations in comparison to their English counterparts, and there is great uncertainty about the coverage and quality of the linguistic information related to Portuguese that is encoded in these models. As a consequence, there is also uncertainty about the stability and quality of the output produced by systems built using these models.

In this paper, we address some of these issues, assessing the linguistic generalisation of language models for Portuguese. In particular, we propose a model-agnostic test set to measure the ability of a model to capture expected linguistic patterns found in Brazilian Portuguese. In order to observe how generalisation works when these

[1] We highlight that the model by Abdaoui et al. (2020) is a distillation from mBERT, therefore aiming to achieve a model that has similar representations.

models are applied, we designed two different classes of tests based on masked-word prediction: the first class consists of highly idiomatic multiword expressions (MWEs), which are hardly generalisable and need to be learned individually, as their meaning cannot be derived from their parts—for the tests in this class, the model has only one correct option for replacing the mask; the second class comprises several grammatical tests, which correspond to generalisable tasks, because the grammatical patterns allow for a broader range of lexical units to be used for replacing the mask. This is an exploratory study that aims at a first attempt to probe BERT-like models for Portuguese in terms of their generalisation capabilities, and our main research question is: how well can BERT-like models for Portuguese perform in predicting words when faced with generalisable and non-generalisable linguistic patterns?

The main contributions of this study are:

- A dataset for testing generalisation related to multiword-expression (MWE) and grammatical information. The MWE dataset contains 330 test instances targeting highly idiomatic (i.e., non-compositional) MWEs, while the tests for grammatical information are divided into 6 categories and amount to 1232 test instances.
- An analysis that generated a language profile of BERT-like models for Brazilian Portuguese.
- A comparison of the BERT-like models' quality considering only the best generated candidates and the ten best candidates, also highlighting cases where the model puts more confidence on wrong candidates.

The proposed dataset is available on Github,[2] along with the manual evaluation that was carried out.

The following sections are organised as follows: Sect. 2 summarises the most influential model-agnostic proposals for intrinsic evaluation, including those targeting Portuguese. Section 3 presents the BERT-like models for Portuguese, describing their characteristics, and the models used in this work. Then, in Sect. 4, we detail the methodology that was used to create all seven tasks comprised in our test set. The results of the different models are presented task by task in Sect. 5, along with a discussion of the main findings. Finally, in Sect. 6, we sum up our findings and discuss potential future work.

2 Related work

The performance that state-of-the-art, pre-trained language models achieve on language tasks is seen as an indication of their ability to capture linguistic information. However, their lack of transparency and interpretability prevents an in-depth analysis of the linguistic patterns and generalisations that they capture. Therefore, considerable attention has been devoted to the development of intrinsic evaluation protocols for inspecting the information captured by these models.

[2] Github page of the dataset: https://github.com/rswilkens/Assessing-Linguistic-Generalisation-in-Language-Models.

Intrinsic evaluations usually involve experiments in which word embeddings are compared with human judgements regarding relations between words. They may be divided into various categories. For example, Bakarov (2018) arranges them as follows:

(a) Conscious evaluation: for tasks like semantic similarity, analogy, thematic fit, concept categorisation, synonym detection, and outlier word detection;
(b) Subconscious evaluation: for tasks like semantic priming, neural activation patterns, and eye movement data;
(c) Thesaurus-based evaluations: for tasks that use thesaurus vectors, dictionary definition graph, cross-match test, semantic difference and semantic networks; and
(d) Language-driven evaluation: for phonosemantic analysis and bi-gram co-occurrence frequency.

In this paper, we concentrate mainly on the conscious evaluation tasks. These vary from testing morphosyntactic agreement patterns, like number and gender (Warstadt et al., 2020), to more semantic-related information, such as the implications of negation (Ettinger, 2020; Kassner & Schütze, 2020).

As an approach to evaluate sensitivity to syntactic structures in English, Linzen et al. (2016) proposed an evaluation focusing on number agreement for subject-verb dependencies, which is something that we also address in this paper. They evaluated four tasks:

(a) Number prediction: a binary prediction task for the number of a verb;
(b) Verb inflection: with a variation of the number prediction when the singular form of the upcoming verb is also given to the model;
(c) Grammaticality judgements: as another classification task, but the model must indicate the grammaticality of a given sentence; and
(d) Language modelling: in which the model must predict a correct word with the highest probability than a wrong word in a given sentence.

A related dataset, the "colorless green ideas" test set, was proposed by Gulordava et al. (2018) including four languages: Italian, English, Hebrew, and Russian. Its test items focus on long-distance number-agreement evaluation, testing the accuracy in cases of subject-verb agreement with an intervening embedded clause and agreement between conjoined verbs separated by a complement of the first verb. Their test set is composed of syntactically correct sentences from a dependency treebank which are converted into nonce sentences by replacing all content words with random words with the same morphology, aiming to avoid semantic cues during the evaluation.

In an attempt to prevent too unnatural sentences (e.g., the apple laughs), Marvin and Linzen (2018) use non-recursive, context-free grammars to automatically generate 350,000 English sentence-pair templates, consisting of grammatical and ungrammatical sentences. This automatically constructed dataset can be used for evaluating the grammaticality of predictions made by a language model in relation to subject-verb agreement, reflexive anaphora and negative polarity. Marvin and Linzen (2018) follows a similar direction as Gulordava et al. (2018), but they are more moderate by avoiding sentences that are very implausible or violate selectional restrictions. For that, they specialised the generation templates for animate and inanimate subjects and verbs.

To evaluate the performance obtained by BERT in these datasets [i.e., Gulordava et al. (2018); Linzen et al. (2016); Marvin and Linzen (2018)], Goldberg (2019) fed a masked version of a complete sentence into BERT, then compared the score assigned to the original correct verb to the score assigned to the incorrect one.

Mueller et al. (2020) evaluated LSTM language models and the monolingual and multilingual BERT versions on the subject-verb agreement challenge sets. They used CLAMS (Cross-Linguistic Assessment of Models on Syntax), which extends Marvin and Linzen (2018) to include English, French, German, Hebrew, and Russian. To construct their challenge sets, Mueller et al. (2020) use a lightweight grammar engineering framework (attribute-varying grammars) that allows for more flexibility when compared to the hard-coded templates of Marvin and Linzen (2018). Their experiments on English BERT and mBERT suggest that mBERT learns syntactic generalisations in multiple languages, but not in the same way in all languages. In addition, its sensitivity to syntax is lower than that of monolingual BERT.

Regarding evaluations for Portuguese, Bacon and Regier (2019) evaluated BERT's sensitivity to four types of structure-dependent agreement relations for 26 languages, including Portuguese. They showed that both the monolingual and multilingual BERT models capture syntax-sensitive agreement patterns. Bacon and Regier (2019) evaluated the models in a cloze task, in which target words would share the morphosyntactic features of the word with which they agree. In the cloze task, the data comes from version 2.4 of the Universal Dependencies (UD) treebanks (Nivre et al., 2017), using the part-of-speech and dependency information to identify potential agreement relations. BERT responses are then annotated with morphosyntactic information from both the UD and the UniMorph projects (Sylak-Glassman, 2016) to be compared with the morphosyntactic features of the word with which they agree. With respect to Portuguese, Bacon and Regier (2019) evaluate 47,038 sentences in the cloze test and 2,107 in the feature bundles using mBERT, and the results obtained show that it performed remarkably well, achieving an accuracy close to 100%.

In another multilingual evaluation, Şahin et al. (2020) introduced 15 probing tasks at type level for 24 languages, finding that some probing tests correlate to downstream tasks, especially for morphologically rich languages. They performed an extrinsic evaluation using downstream tasks, assessing the performance in POS tagging, dependency parsing, semantic role labelling, named-entity recognition, and natural-language inference targeting German, Russian, Turkish, Spanish, and Finnish. Moreover, they propose the contextless prediction of the morphological feature of a given word, aiming to identify the feature indicated in the UniMorph dictionary (Sylak-Glassman, 2016). For that, Şahin et al. (2020) removed ambiguous forms and partially filtered infrequent words. This evaluation included features like the following for Portuguese: number, gender, person, and tense. However, they did not report the models' performance for Portuguese.

Evaluations related to MWEs have often focused on specific tasks (Constant et al., 2017) like, for example, on MWE token identification (Scholivet & Ramisch, 2017; Pasquer et al., 2020), evaluating the sensitivity of a model to the occurrence of MWEs in sentences or looking at their interpretation (Cordeiro et al., 2019; Garcia et al., 2021a), assessing to what extent models are able to represent their potential idiomaticity (Tayyar Madabushi et al., 2021, 2022). In particular the results obtained for MWE

interpretation by a variety of off-the-shelf embeddings with different levels of contextualisation (including transformer-based models) indicated that idiomaticity was not yet accurately captured even when adopting large pre-trained language models, for any of the two languages covered, English and Portuguese (Garcia et al., 2021a; 2021b). One recent evaluation included these two specific tasks: SemEval 2022 Shared Task 2 (Tayyar Madabushi et al., 2022), with Subtask-A dedicated to token identification and Subtask-B to MWE representation and interpretation, for three languages, English, Portuguese and Galician. The results obtained by different systems indicated that although considerable increases in performance could be obtained with fine-tuning, there were still challenges when dealing with unseen MWEs (Tayyar Madabushi et al., 2022). In this paper we approach these two tasks jointly by investigating the ability of a model to identify an MWE in an idiomatic context, focusing on Portuguese.

3 BERT models for Portuguese

Large pre-trained language models are valuable assets for language processing. These models support technological and scientific improvements. For English, for example, we can easily find thousands of models accessing the Huggingface website.[3] However, this abundance of models is not true for most languages. For example, there are only a few hundred models for Portuguese available in the same repository,[4] and most of these are for Machine Translation, Automatic Speech Recognition, and Text2Text Generation. In this section, we highlight three models dedicated to language processing: smaller mBert and BERTimbau, which are both based on Portuguese for general purposes, and BioBERTpt, which is based on Portuguese for special purposes.

Souza et al. (2019, 2020) trained language-specific models for Brazilian Portuguese (nicknamed BERTimbau), using data from brWaC (Wagner Filho et al., 2018). Their models were trained over one million steps using the BERT-Base and BERT-Large Cased variant architectures. In order to evaluate their models, they resorted to three probing tasks (sentence textual similarity, recognising textual entailment, and named-entity recognition) in which BERTimbau improved the state of the art compared to multilingual models and previous monolingual approaches.

Targeting clinical NLP, Schneider et al. (2020) trained BioBERTpt, a deep contextual embedding model for Portuguese, by fine-tuning mBERT. BioBERTpt is trained using clinical narratives and biomedical-scientific papers in Brazilian Portuguese.

Aiming to evaluate the BERT models for Brazilian Portuguese, we selected two models that were trained using brWaC (Wagner Filho et al., 2018) and multilingual BERT, which can serve as a comparison point with BERT models trained in other languages. We used both the BERTimbau Base and Large models (Souza et al., 2019, 2020), which have, respectively, 12 layers and 110 M parameters, and 24 layers and 335 M parameters. These models were chosen because they are uniquely trained in Brazilian Portuguese, thus reducing issues from different variations, and they are also widely adopted, being the most popular models for Portuguese on the Huggingface

[3] https://huggingface.co/models?filter=en. Accessed on 31/05/2022.

[4] https://huggingface.co/models?filter=pt. Accessed on 31/05/2022.

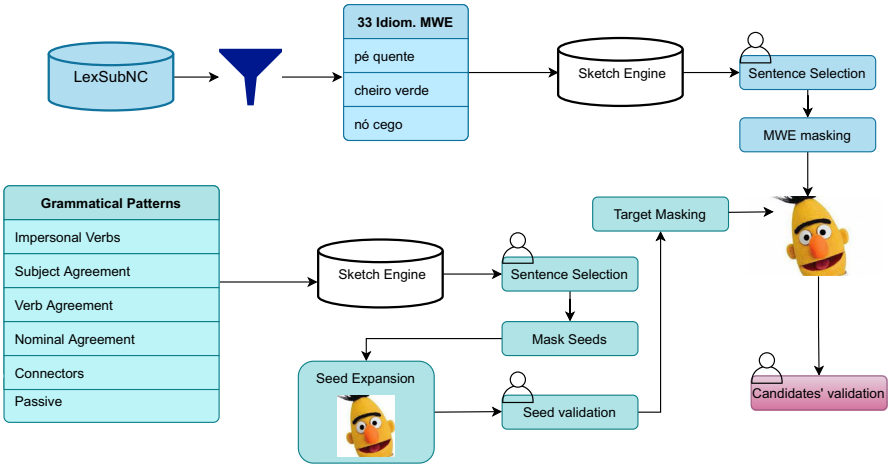

Fig. 1 Methodology for creating the test dataset

webpage at the time of this research[5]. Therefore, the findings of this paper may have a wide impact in downstream applications for Brazilian Portuguese.

4 Methodology

Our main goal in this study is to explore how well BERT-like models for Portuguese handle generalisable and non-generalisable linguistic information. In this section, we describe the steps performed for developing the dataset that was used for testing the models. We first compiled a multiword expressions (MWEs) test set (see Sect. 4.1) by selecting idiomatic MWEs, which are not generalisable and need to be learned individually, and five context sentences for each of them. As part of the generalisable information, we developed a series of grammatical tests (see Sect. 4.2), for which we also selected context sentences and looked for specific grammatical patterns.

Once we obtained the sentences for both tasks, we created different seeds for each of the sentences, aiming to use these seeds as cues to a masked word. Finally, we fed the selected BERT models with our test sentences, and manually evaluated the models' output (see Sect. 5). Figure 1 presents an overview of this methodology.

4.1 Multiword expression tests

Multiword expressions (MWEs), such as nominal compounds (NCs) and idioms, can be defined as linguistic structures that cross word boundaries (Sag et al., 2002). Due to their many potential linguistic and statistical idiosyncrasies (Sag et al., 2002), knowledge about MWEs is often considered a mark of proficiency for language learners.

[5] On 31 May 2022, BERTimbau Base had already been downloaded 229 k times, and BERTimbau Large, 51.3 k times.

In the MWE task proposed here, our goal is to assess the MWE inventory of a model by measuring its ability to identify an MWE in a given context. The context here is a sentence where the MWE originally appears, and we also keep a single word of the MWE to be used as cue for the model.

We first collected all the highly idiomatic, two-word NCs available in LexSubNC (Wilkens et al., 2017), such as *pão duro* [stingy person; lit. *hard bread*] and *pé quente* [person who brings good luck; lit. *hot foot*]. The MWEs in LexSubNC were annotated by human judges using a Likert scale from 0 (idiomatic) to 5 (compositional). In this study, we used all MWEs for which the average compositionality score was lower than 1, i.e., the most idiomatic MWEs in LexSubNC. We selected only the most idiomatic ones because the models would not be able to predict them by means of generalisation, as one word of the NC would require the model to specifically predict the other word in order to form a plausible sentence. This allows us to explore whether the models store information about compound lexical units. This resulted in a selection of 33 highly idiomatic MWEs. Based on this selection, we then used Sketch Engine (Kilgarriff et al., 2004) to choose 5 sentences from the Corpus Brasileiro (Sardinha, 2010) for each MWE. The sentence selection was done by shuffling concordance results on Sketch Engine and selecting 5 full sentences of different lengths for each of the 33 MWEs, which resulted in 165 different sentences for the MWE test set.

For assessing the models' capacity to retrieve the target MWE given a compatible context, we adopted a masking protocol, in which we fed the model with an input sentence using a mask to replace one word of the MWE. Since all MWEs that were used contained two words (a noun and an adjective), we generated two test sentences for each of the 165 selected sentences, by first masking the adjective and using the noun as cue for the model, and then masking the noun and using the adjective as cue. For example, the sentence *Presidente, trago uma notícia em* ***primeira mão***. [*President, I bring you first-hand news.*], which contains the NC *primeira mão* [*first-hand*], is processed twice:

– *Presidente, trago uma notícia em [MASK]* ***mão***.
– *Presidente, trago uma notícia em* ***primeira*** *[MASK].*

As such, both words were masked once in each sentence, so the full MWE test set was composed of 330 masked sentences.

Our expectation was that these sentences, along with one of the MWE component words, should provide enough context about the MWE to prompt the model to produce the missing word. The output of this process resulted in 10 candidate words for each test sentence, i.e., 20 candidate words for each MWE: 10 for the first component word of the compound and 10 for the second. The model was then evaluated in terms of whether or not it could predict the correct word for replacing the mask as its best candidate or among its 10 best candidates.

4.2 Grammatical tests

To assess the models' grammatical representation, we developed a broad range of cloze tests, each one targeting a specific phenomenon, with several sub-phenomena, where each sub-phenomenon was tested using a set of sentences and a series of vari-

ations. The targets of these tests were: impersonal verbs, subject agreement, verb agreement, nominal agreement, passive and connectors[6]. By selecting a broad, even if non-exhaustive, range of grammatical tests, we were able to explore how the model generalises linguistic information, as the linguistic patterns that were provided allow for different lexical units to be suggested as replacement for the missing word.

The tests were inspired by the work of Kurita et al. (2019), who propose a template-based method to quantify social bias in BERT. Similarly to what was done with the MWE test set, we selected full sentences from Corpus Brasileiro (Sardinha, 2010) for each of the grammatical tests, and then each test sentence was converted into several test instances, with the change of a cue word or phrase, and the addition of a mask. As such, each test sentence contains a mask and a set of seeds as parameters. Again, similar to what we did in the MWE test, the mask is the "blank", and the seeds serve as a linguistic cue to the expected answer. As an example, let us observe the following test sentence that was used for testing nominal agreement[7]:

> É precisamente na revelação <*SEED*> [MASK] que reside o valor dessas cartas, diz.
> [It is precisely in the revelation <*SEED*> [MASK] that lies the value of those letters, he/she says.]

This single base sentence is used 41 times in the dataset, and each time it contains a different seed, which can account for different morphological variations (e.g., gender and number), such as *das relações* [*of the relations*$_{FemPlural}$], *das capacidades* [*of the capacities*$_{FemPlural}$], *da alma* [*of the soul*$_{FemSing}$], *da condição* [*of the condition*$_{FemSing}$], *dos seres* [*of the beings*$_{MascPlural}$], *dos segredos* [*of the secrets*$_{MascPlural}$], *do ser* [*of the being*$_{MascSing}$], and *do espírito* [*of the spirit*$_{MascSing}$]. In this example, we use four types of seeds related to the expected type of agreement. Each type of seed then targets a different output in the masked word (masculine singular, masculine plural, feminine singular and feminine plural). The different seed sets are specific for each sentence, because they have to take into consideration its syntactic and semantic restrictions. Our approach also shares similarities with the one proposed by Marvin and Linzen (2018). However, their seeds approach uses a grammar with attributes to guide the sentence generation while our work uses pre-selected seeds. Our seeds are related to their attributes, and our generation of grammatical sentences is the combination of seed and original target word. Although our methodology can generate ungrammatical examples, we consider that it is easier to evaluate and filter out these ungrammatical cases instead of designing a sentence generation grammar that controls the process.

In terms of creating seeds for the grammatical tests, we used several different methods, depending on the type of seed that was needed. For **verb and subject agreement**, and for **impersonal verbs**, we automatically generated a set of seeds based on the UNITEX-PB dictionary (Muniz et al., 2005; Vale & Baptista, 2015), and then also

[6] We will explain each of the different grammatical tests in more detail in Sect. 5.2, as we go through each of the tests. Here in the methodology, we will focus on the general process used to create the test sets.

[7] As a general reference for the examples of grammatical tests presented in this paper, we will use angle brackets (<>) and italics for indicating the seeds, and square brackets ([ ]) to indicate the words predicted by the model. As a reference, we also present a translation that is very close to literal in most of the cases.

validated each seed in context[8]. In the case of **passive voice**, due to the complexity of the seeds, we had to manually generate sets of seeds for each sentence. Finally, in the case of **connectors**, we could not use any type of seed, because this would require us to generate full clauses, so we only used five sentences for each connector that was tested.

Nominal agreement was a special case, as we automatically generated seed sets using a mask where the original seed would be. This allowed us to have a more extensive coverage. In this process, we selected the top 10 candidates from BERTimbau Base, used the UNITEX-PB dictionary to generate extra seeds by varying gender and number of the words, and then we manually evaluated the candidates to eliminate bad fits (i.e., we removed the candidates with agreement errors or a different meaning). Before we applied this methodology, we debated about using BERTimbau to generate seeds that would be used to predict the masked word, as this would generate a bias that could help the model. However, as we further discuss in Sect. 5.2.3, in most cases, the model was not able to take advantage of this bias, which is in itself an interesting result that requires further study.

Using the manually validated templates composed of sentences and their seeds, we generated 10 candidate answers for each masked word using BERTimbau Large.[9] These candidates were also annotated with part-of-speech using the Unitex-PB dictionary (Muniz et al., 2005; Vale & Baptista, 2015).[10] Then, all candidate words from BERTimbau Large were evaluated by a linguist according to their syntactic and semantic suitability in context. Although we used seeds as cues for the expected type of answer (for instance, we used generic passive structures to induce the model to produce a nominal participle as candidate in the passive voice tests), the actual prediction of the model could be from a different category, and still the sentence would be grammatically correct. As such, the evaluation took into account any answer that would fit in the context, not necessarily only the ones that were expected. Even so, the answers that were correct, but deviated from the expected target answer, were identified during the evaluation, and are further discussed in the results.

In total, after following the procedure described here, the dataset for grammatical evaluation consists of 6 dimensions that contain 1232 test instances.

5 Results

This section presents the evaluation results considering the two different subsets of tests: multiword expressions (MWEs) and grammatical information. The MWE results consider three models: BERTimbau Large, BERTimbau Base and mBERT. These models were evaluated in terms of accuracy of the best prediction and accuracy among the top ten candidates. The grammatical evaluation was done on BERTimbau Large,

[8] Due to the nature of the UNITEX-PB dictionary, we generated seeds based mainly on traditional-grammar-based verb tenses, which, for Portuguese, are all simple tenses, so we do not look at compound tenses for these cases in our study.

[9] Given the size of the manual evaluation task, with more than 12 k items, we only used the BERTimbau Large model, as it performed best in the MWE task, as will be discussed in the next section.

[10] We do not use a parser in this step for simplicity.

Table 1 Accuracy of different models for predicting Word1 and Word2 in 165 sentences. acc@10 means the accuracy considering the top 10 candidates

Model	Word	ACC	ACC@10
BERTimbau Base	1	40.00%	57.58%
BERTimbau Base	2	45.45%	64.24%
BERTimbau Large	1	52.73%	65.45%
BERTimbau Large	2	51.52%	67.27%
mBERT	1	6.67%	16.97%
mBERT	2	3.64%	11.52%

and considered results in terms of precision of the best candidate (or accuracy) and accuracy among the top ten candidates.

5.1 MWE tests

Aiming to assess the model prediction capabilities concerning MWE vocabulary, we calculated the accuracy of the top candidate for each masked sentence, and we also used a more lenient approach, by looking for the correct word in the top ten predictions of the model (acc@10).

As can be seen in Table 1, the multilingual model, mBERT, performed consistently worse than the dedicated models trained for Portuguese. This can be explained by the curse of multilinguality (Conneau et al., 2019), which states that a small set of languages positively impacts the model's performance, but this effect becomes negative when the model has to deal with several languages. Moreover, BERTimbau Large performs better than the base version.[11] In all cases, it was the larger model that performed better, suggesting that the ability to learn MWEs is related to the size of the model. In terms of the difficulty of the task, apart from the multilingual BERT, the models were able to predict the missing MWE component as the first alternative in 40–52.73% of the cases, which indicates that the models have some level of MWE representation, but there is still need for further improvement. In a more lenient scenario, when we analyse the ability of a model to predict the missing word among the top 10 most likely words, the results substantially improve. While mBERT has an average increase of up to 9% in accuracy in comparison with only the top prediction, the BERTimbau models have a much more substantial increase: BERTimbau Base has an improvement of 17.58% and 18.79% for predicting word 1 and word 2, respectively, in relation to the top candidate; and BERTimbau Large improves 12.72% for word 1 and 15.75% for word 2. Although mBERT shows good capacities in different NLP tasks, the model captured little to no information to predict idiomatic items that are specific to Brazilian Portuguese. Additionally, we observed that the gain of including more candidate words is mainly restricted to the top two candidates.

The changes from BERTimbau Base to BERTimbau Large allowed the model to learn a larger MWE inventory, considering the target MWEs, and have more accu-

[11] We compared the three models using McNemar's test, where we observed p-values < 0.0001. Analogously, we compared the models regarding the task of predicting W1 and W2 separately, where we observed p-values of 0.0001 and 0.034 respectively for the Base to Large comparison and 0 for the other comparisons.

Table 2 Performance comparison between BERTimbau models: MWEs which were predicted by BERTimbau Large, but not by BERTimbau Base*

Word 1 is masked	Word 2 is masked
livro aberto	livro aberto
montanha russa	montanha russa
nó cego	olho mágico
olho mágico	pau mandado
pão duro	pavio curto
pé frio	pé frio
pé quente	pé quente
peso morto	peso morto
planta baixa	planta baixa
sangue azul	saia justa
sangue frio	sangue frio

*MWEs are presented in alphabetical order

Table 3 Performance comparison between BERTimbau Large and BERTimbau Base: MWEs that were not predicted by either model*

Word 1 is masked	Word 2 is masked
bode expiatório	bode expiatório
cheiro verde	gato pingado
elefante branco	longa metragem
ovelha negra	nó cego
pavio curto	olho gordo
pente fino	pente fino
roleta russa	vista grossa
saia justa	

*MWEs are presented in alphabetical order

rate prediction given the cues in the masking task (Table 2). However, there were two cases in which performance decreased using the larger model: *sangue azul* [blue blood] and *pão duro* [stingy person; lit. *hard bread*]; and other cases that are still not accurately represented by either model (Table 3). Overall, the difference in performance between BERTimbau Large and Base was not as big as the difference between them and mBERT, but BERTimbau Large was able to learn more MWEs and displayed more confidence in correct predictions. However, we also noticed that the values of accuracy at 10 for BERTimbau Base had a higher increase when compared to BERTimbau Large, indicating that exploring more outputs from the Base model might lead to the same performance of BERTimbau Large, which requires more processing power.

5.2 Grammatical tests

This section presents the results of the grammatical tests to which BERTimbau Large was submitted[12]. As there were six very different grammatical tests, we report and

[12] See Sect. 4 for a discussion of why we chose this model.

discuss their results individually as precision at 1 (or accuracy) and precision at 10, aiming to analyse the model's proficiency. At the end of this section, we make a brief comment on the overall performance of the model.

5.2.1 Impersonal verbs

The set of sentences with impersonal verbs tested whether the model would produce candidates that were not subjects. The verbs used as cue represented existential verbs (such as *to exist*, *to have been*) and meteorological verbs (such as *to rain*, *to snow*) (Perini, 2010), which do not accept any subject in Portuguese, and are therefore defective in their conjugation. As such, we expected the model to predict candidates that were not nouns or pronouns. Here is an example of impersonal verb test instance from our dataset:

> [MASK] <*Choveu*> na cidade durante 12 horas seguidas, de sábado a domingo.
> [[MASK] <$It_{Not\ Represented}$ $rained_{Impersonal}$> in the city for 12 straight hours, from Saturday till Sunday.]

In these cases, the cue was simply the verb in its several tenses (e.g., *choveu* [it rained], *choverá* [it will rain]).

As we can see in Table 4, the models perform well in this task, with meteorological verbs having slightly worse results, as the model still predicted some candidates that did not fit. Results considering the top candidate for meteorological verbs were near 100%, and the average precision considering the top 10 candidates decreases to 76.04%. For existential verbs, results were much higher, with a precision of 97.50% among the top 10 candidates and 100% for the top 1. The model was able to generate punctuation marks, such as parenthesis or quotation marks, in most of the templates, which were a good fit for the test sentences.

Looking at the different tenses, to check whether there is any impact of the verb form used as a cue for the model, we see that only the simple pluperfect form produced results below 100% for the top candidates, and was also the worse cue on the precision at 10 evaluation, with 65.33%. This could be a reflection of the fact that the pluperfect tense in Brazilian Portuguese is usually written in a compound form, which is formed by the auxiliary verb *ter* [to have] conjugated in the imperfect tense, associated with the past participle of the main verb, whereas the simple pluperfect form, which was the form used in the test sentences, is not widely used anymore, except in highly formal contexts. As such, assuming that lower frequencies have an impact on the quality of the representation, the simple pluperfect tense might not have been learned well by the model. Nonetheless, the performance confirms the proficiency of the model with impersonal verbs.

5.2.2 Subject agreement

For the subject agreement task, the model was expected to produce a subject that would agree in person and number with the provided verb seeds. Here is an example of masked sentence that was used for testing subject agreement:

Table 4 Impersonal verbs: verb types

Case name	P@1	P@10
Meteorological verbs	97.56%	76.04%
Existential verbs	100.00%	97.50%

Table 5 Impersonal verbs: results by tense and mode

Case name	P@1	P@10
Past indicative	100.00%	82.00%
Present indicative	100.00%	85.63%
Future indicative	100.00%	80.00%
Imperfect indicative	100.00%	95.71%
Pluperfect indicative	86.67%	65.33%
Future subjunctive	100.00%	79.00%

Neste contexto, como [MASK] <*comentamos*>, o tempo de compreender se evapora.
[In this context, as [MASK] <*discussed*>, the time to understand evaporates.]

Considering the type of subject that was expected given the verb conjugation, as seen in Table 6, results both in the top 1 and in the top 10 were varied, ranging from 64.10% for the second person singular to 100% for the third person singular. It was expected that the model would have a hard time to produce good candidates when the expected subject should fit a second person singular or plural, because these are less commonly used conjugations in Brazilian Portuguese[13], however it is interesting to see that the model also has higher confidence in wrong answers when we look at the results for the third person plural, which is not an unusual conjugation at all.

When we look at the tense and mode of the conjugated seed (Table 7), results also varied, and it is interesting to notice that the tense that yielded worst results in the top 1 candidates (barely above 75% precision) was present indicative, while future indicative was the tense with best results (above 92% precision). Among the top 10 results, pluperfect indicative was the worst, with 71.78%, and the best result was again the future indicative, with above 90% precision.

Given that Portuguese allows for hidden or indeterminate subject, we were expecting that the model would produce results that would not be subjects, but that would fit in the sentences. This was the case for 1450 instances (considering the top 10 candidates), which accounts for 52.54% of the all instances that were evaluated. However, among all these 1450 instances, only for 37 of them the model had a confidence score above 50%. An example of this can be seen in one of the instances of the sentence presented above, where the mask was substituted by an adverb:

[13] Usually the third person singular and third person plural are used instead of the respective second person conjugations.

Table 6 Subject agreement: results by expected person and number

Case name	P@1	P@10
First person singular	97.06%	93.46%
Second person singular	64.81%	61.73%
Third person singular	100.00%	95.85%
First person plural	94.59%	92.33%
Second person plural	66.67%	63.70%
Third person plural	85.00%	93.52%

Table 7 Subject agreement: results by tense and mode

Case name	P@1	P@10
Present indicative	75.61%	75.13%
Past indicative	78.72%	87.10%
Pluperfect indicative	81.25%	71.78%
Future indicative	92.86%	90.46%
Past tense future indicative	82.05%	79.78%
Imperfect indicative	79.49%	75.28%

Neste contexto, como [já] <*comentamos*>, o tempo de compreender se evapora.
[In this context, as we$_{Hidden}$ [already] <*discussed*$_{3rd\,Person\,Plural}$>, the time to understand evaporates.]

In this case, we would expect to see a first person plural subject (i.e., *nós* [we]), however, the model produced the adverb *já* [already], which fits the context and doesn't make the sentence incorrect, because the syntax of Portuguese allows for hidden subjects. Compared to this, the model produced only 792 correct instances (28.70% of the evaluated instances) that were actual subjects, but 79 of them had confidence scores above 50%.

Considering the precision of this task, the model has a good capacity of generating elements that fit in the sentences, especially when considering only the top candidates. Three verb conjugations (second person singular and plural, and third person plural) and two tenses (present and pluperfect indicative), however, proved to be an extra challenge for the model.

5.2.3 Nominal agreement

For the task of nominal agreement, we looked into the adjective agreement using a noun as cue. Since Portuguese adjectives agree with nouns in gender and number, we had four different test categories: masculine singular, masculine plural, feminine singular and feminine plural. An example of masked sentence used in the dataset was already presented in Sect. 4.2, when we discussed the methodology for creating different test sentences using one base sentence, but here is another one, as we will use it later in this section:

Table 8 Nominal agreement: gender and number

Case name	P@1	P@10
Feminine singular	78.00%	92.00%
Masculine singular	86.84%	91.58%
Feminine plural	87.18%	88.46%
Masculine plural	91.89%	93.22%

Nada mais incômodo do que não rir <*da piada*> [MASK] de um palhaço profissional.
[Nothing is more embarrassing than not laughing <*at the*$_{FemSing}$> [MASK] <*joke*$_{FemSing}$> of a professional clown.]

In this specific case, we expected the system to produce an adjective in the feminine singular form as replacement for the mask.

Table 8 shows that results were fairly good across all categories, reaching up to 93.22% in masculine plural when all 10 candidates were considered for each mask. It is interesting to see that, in this test, the suggestion in which the model has most confidence is sometimes among the few wrong ones that it produces, as the results on p@1 were consistently worse than when the top 10 candidates were considered. This is especially true for feminine singular, in which the model achieved only 78% precision for the top 1 candidates.

In terms of producing candidates that would fit the context but were not adjectives, the model only had one candidate that fell into this category. For the example sentence presented above, among its top ten candidates, the model produced *diante* [in front], which is an adverb that, along with the preposition *de* [of], creates a prepositional phrase that actually fits in the context. This was the only correct instance of a non-adjective candidate in the full set of 1,640 generated candidates.

As a final point of discussion for the nominal agreement test, we would like to recall that the seeds used for these tests were originally suggested by BERTimbau Base, as we masked the seed in the original sentences to get a set of seed candidates, as we previously discussed in Sect. 4.2. We were expecting the model to have a very high precision on the top 1 candidates, as it would potentially produce the original sentence. However, what we observed was that, in most cases, the model was not able to predict it, sometimes not even among the top 10 candidates. This might have been due to our second step, which was processing the seeds with the UNITEX-PB dictionary, or it might have been because we used BERTimbau Base as generator, and tested on BERTimbau Large. In any case, it remains as an interesting side result from this study, that the model was not able to predict words that were used for suggesting the seeds.

5.2.4 Verb agreement

The task of verb agreement was designed to check whether the model can produce verb candidates that correctly agree with the cue. The sentences used for this test had temporal cues or specific language patterns that would induce or require certain verb conjugations. As seeds, we used pronouns and nouns in singular and plural, but also

used some verb structures to check for the production of infinitives and gerunds. We present below two examples of test sentences. The first one uses a pronoun as seed, which was used as cue for a conjugated verb, and the second example uses a verb as seed, cuing an infinitive[14]:

> Ontem <*nós*> [MASK] Maria Aparecida para receber o dinheiro.
> [Yesterday <*we*> [MASK] Maria Aparecida in order to receive the money.]

> O aluno Hélio gostou do convite e resolveu imediatamente que não <*precisaria*> [MASK].
> [Student Hélio liked the invitation and immediately decided that he <*would not need to*> [MASK].]

Table 9 shows the results for each expected verb form. Indicative forms (first two rows) had much better results than subjunctive forms (last three rows), while the non-conjugated forms (rows 3 and 4) had the best results overall, reaching up to 100% precision considering the top 1 candidate.

In this specific case, it was observed that some cues were not as effective as others, so we investigated the results based on the different pronouns and nouns that were used as cues. In Table 10, pronouns such as "tu" [you_{Sing}], "nós" [*we*] and "vós" [you_{Plural}] proved to be challenging cues, as the model did not seem to be able to produce many candidates that agree with them. When considering only the top 1 candidates, "tu" had an abysmal result, as the model was not able to predict a single correct instance—and only one correct candidate among the top 10—, and "vós", a very formal and archaic pronoun, did not reach 8% of precision. While "tu" and "vós" are pronouns that are less commonly used in Brazilian Portuguese, being frequently substituted, respectively, by "você" and "vocês"—which had a much better precision—, it is harder to explain why the model does not produce good candidates for a cue like "nós". One possibility is that this form may be replaced by a more informal option (i.e., *a gente* [we; lit. *the people*]). This could be a bias of the corpus used for training the model. However, a study comparing both forms in the attention layer is still needed to shed more light into this matter.

Although responses with the expected verb tenses were induced in most cases, 5.00% of the responses provided by the model were correct, meaning that they fit the sentence well, but were not verbs or did not agree with the cue provided.

5.2.5 Connectors

Connectors (also referred to as connectives) have been studied under various perspectives. From a Linguistic point of view, they have usually been more studied under the umbrella of Text Linguistics, under the scope of cohesion or coherence (see, for instance, De Beaugrande and Dressler (2011), for an overall introduction to Text Linguistics and the concepts of cohesion and coherence). Several studies have proposed a typology of connectors, such as Halliday and Hasan (2014), Koch (1988), Koch (1999), Louwerse (2002), some focus more on traditional conjunctions, such as *and*,

[14] In the English translation of the second example, the negation *not* was included in the seed for simplicity, but the seed in Portuguese is only the verb, and the negation *não* is part of the base sentence.

Table 9 Verb agreement: different expected tenses and modes

Case name	P@1	P@10
Past indicative	87.51%	78.44%
Present or future indicative	94.63%	68.59%
Infinitive	100.00%	96.10%
Gerunds	94.81%	74.88%
Present subjunctive	58.00%	33.40%
Past subjunctive	62.78%	25.08%
Future subjunctive/conditional	68.15%	57.10%

Table 10 Verb agreement: breakdown of pronouns

Case name	P@1	P@10
Eu [I]	80.00%	52.67%
Tu [You$_{Sing}$]	0.0%	0.77%
Você [You$_{Sing}$]	88.89%	58.89%
Ele/Ela [He/She]	79.41%	60.59%
Nós [We]	30.77%	13.08%
Vós [You$_{Plural}$]	7.69%	6.92%
Vocês [You$_{Plural}$]	85.71%	44.29%
Eles/Elas [They]	85.71%	44.29%

because, *however*, and others involve also what Koch (1988) addresses as *sequential particles*, which encompasses adverbial phrases that helps in creating textual cohesion, such as *finally*, *before that*, *on the one hand*.

In this paper, we are considering only the former category, which includes the logical connectors, as basis for selecting input sentences. This is a more computational approach, which is also present in tools for textual analysis, such as Coh-Metrix (Graesser et al., 2004) and its adaptation to Portuguese, Coh-Metrix-Port (Scarton & Aluısio, 2010). These tools use the typology by Louwerse (2002) and organise the connectors in categories such as *causal*, *temporal*, *additive*.

In the specific case of connectors, typologies are useful as a starting point, but they can only be applied after the text is already produced, as they serve for the purpose of textual analysis. Since connectors can work as the main element that provides the meaning of a link between two clauses, they are hardly enforceable in a specific context. As such, the same sentence that contains a temporal connector (e.g., He was speaking *while* a crowd started to form around him.) could very well have a connector from a different category in its place (e.g., He was speaking *and/but/because* a crowd started to form around him.), and it would still be a plausible sentence, even if it would have a different meaning. Because of this, trying to create a dataset that covers all categories of connectors would not necessarily work for testing the information that the model has about the different types of connectors.

In this study, since we are testing whether the language model is able to generate plausible sentences, that have grammatical accuracy, we used a connector typology

only for selecting a few types of connectors that would serve as basis for the input sentences in our dataset, but we did not aim at an exhaustive description of all possible connectors that would fit in the sentence. Among the selected connectors, we have two conditional connectors (*caso* [if]; *se* [if]), three causal connectors (*pois* [because]; *porque* [because]; *portanto* [therefore]), two alternative connectors (*nem* [nor]; *ora* [now]), two adversative connectors (*contudo* [however]; *todavia* [however]), two temporal connectors (*enquanto* [while]; *quando* [when]), and one conformative connector (*conforme* [according to]). For each of them, we searched and selected five sentences, for a total of 60 different sentences for testing connectors. As we discussed, and as it can be seen in the evaluated data,[15] these had a good coverage of the different types of connectors for Portuguese. An example of test sentence is the following (which had *Contudo* [However] in the original masked position):

> [MASK] o referido processo já foi objeto de estudo nos pareceres supracitados, cabendo apenas à CAPLAN fixar o número de vagas.
> [[MASK] the referred process has already been scrutinised in the aforementioned reviews, and CAPLAN alone is responsible for setting up the number of places available.]

As Table 11 shows, the model was able to predict connectors with very good precision, reaching 100% in all cases among the top 1 candidates, and then varying in precision among the top 10 candidates. It is important to emphasise again that the connectors shown in the table are just for a reference of what was present originally in the base sentence, as the model might have provided other candidates that where plausible within the context. Considering all correct candidates among the top 10 suggestions, 10.77% were not conjunctions in the traditional sense, but most of these were textual connectors, such as *finalmente* [finally] and *também* [also], which could be classified as sequential particles (Koch, 1988). There were, however, two instances, among all 600 candidates that had not connector (or connector-like) functions, but that fit the context: *todo* [all], which is a predeterminant that was suggested for the example sentence above, and *mal* [hardly], which is an adverb that fit the context in another sentence that had *enquanto* [while] as masked element.

As discussed above, many of the sentences could not enforce a very specific requirement in terms of connector. However, sentences such as the ones with "ora" (now), which is a dual connector (*ora..., ora* [∼ now..., now]), had no margin for other plausible options within that context. Also, the fact that one connector can potentially be replaced by another with different meaning does not mean that there is a great amount of plausible options, as the number of connectors is finite. This could explain the very poor precision for the top 10 candidates in several cases. Even so, looking at the overall results obtained, we can see that the model can proficiently generate connectors, because cohesive elements were the top option in 100% of the cases.

5.2.6 Passive

Passive voice in Portuguese has an important characteristic, shared with other Romance languages, which is the agreement of the nominal participle with the subject. This can

[15] Available in this Github repository: https://github.com/rswilkens/Assessing-Linguistic-Generalisation-in-Language-Models.

Table 11 Connector prediction

Case name	P@1	P@10
Caso [if]	100.00%	26.67%
Conforme [according to]	100.00%	46.67%
Contudo [however]	100.00%	86.67%
Enquanto [while]	100.00%	66.67%
Nem [nor]	100.00%	50.00%
Ora [now]	100.00%	20.00%
Pois [because]	100.00%	50.00%
Porque [because]	100.00%	40.00%
Portanto [therefore]	100.00%	90.00%
Quando [when]	100.00%	26.67%
Se [if]	100.00%	50.00%
Todavia [however]	100.00%	96.67%

lead to long distance agreement between these two components. The following example illustrates this type of agreement: *A escolha foi feita.* [The$_{FemSing}$ choice$_{FemSing}$ was made$_{FemSing}$.] This is different to what happens, for instance, with compound verb tenses, such as the compound pluperfect, where the verbal participle of the main verb is used, and thus there is no requirement for agreement (for a broader discussion about nominal and verbal participles, see Perini (2010)). An example of this can be seen in this example: *Ela tinha escolhido aquela cor.* [She$_{FemSing}$ had chosen$_{NoAgreement}$ that$_{FemSing}$ colour$_{FemSing}$.]

To test the generation of nominal participles, we used 13 base sentences, which were modified with a varied number of linguistic cues composed of verbal constructions associated with the verb "ser" (*to be*). This resulted in a total of 175 test sentences. An example of test sentence for the passive voice is the following:

> Contrariando ao ideal materno dos romanos, os filhos das classes médias e altas <*eram*> [MASK] às amas de leite que os alimentavam e se incumbiam dos cuidados gerais da criança.
> [Contrary to the maternal ideal of the Romans, children$_{MascPlural}$ of middle and upper class citizens <*were*> [MASK] to wet nurses, who fed them and were responsible for the general care of the child.]

The results in Table 12 show that the model was able to produce correct candidates most of the time, with 86.65% precision among the top 1 candidates and 78.38% among the top 10 candidates. Interestingly, among the candidates that were not correct, 44.99% were a participle, but had incorrect nominal agreement with the subject of the sentence. The results in the table also point to the model having trouble producing good candidates for the feminine singular cases. This is a rather interesting phenomenon, as it was unexpected for the model to fail on only one type of agreement. A deeper study would have to be conducted in terms of why this happened, but we can hypothesise that, because the verbal form of these participles are generally written in a form that coincides with the masculine, the model made the wrong generalisation and assumed the masculine form also for the nominal participles when they are in singular form.

Table 12 Passive: gender and number breakdown

Case name	P@1	P@10
Feminine singular	17.86%	50.71%
Masculine singular	95.89%	89.32%
Feminine plural	100.00%	55.00%
Masculine plural	98.53%	81.47%

Another possibility would be that there are simply not enough instances of feminine singular nominal participles in the corpus that was used to train the language model.

Finally, within the list of correctly generated candidates, not many deviated from the expected word form, as only 5.07% of the candidates were adjectives that fit the context resulting in a grammatically correct option, instead of the targeted nominal participle.

5.2.7 Overview of the grammatical tests

Summing up the results of the grammatical tests, we can see in Table 13 that tasks that require no agreement had the best results, with 100% precision for connectors and 98.78% for impersonal verbs. Where morphosyntactic characteristics of the language played a role, we see that the results fall below 90% even considering only the candidate in which the model has the most confidence.

Interestingly, the nominal agreement test was the only one that shows better results among the top 10 candidates in comparison to the top 1. This could possibly mean that, for selecting adjectives, lexical cues in the context are stronger than the morphosyntactic information in the cue word. Considering the results reported by Bacon and Regier (2019) for Portuguese on mBERT, we see that the model here had much worse performance for nominal agreement. This could be because we evaluated all candidates, and not only the ones that had the same part of speech as the masked word.

As the only closed class task, we expected connectors to present the largest difference in evaluation from the top 1 to the top 10 candidates, because the model would eventually run out of options to fit in the context. However, the worse performance overall was seen in the task of verb agreement, where the model had problems finding good candidates for a few personal pronouns, both at the top 1 and at the top 10. Moreover, frequency also seems to play a role in model performance, and prediction accuracy decreases for less frequent forms, including very formal pronouns or those that are frequently omitted.

6 Conclusions and future work

In this work we addressed the problem of intrinsic evaluation of large scale language models and proposed a battery of model-agnostic tests to assess linguistic proficiency in Portuguese. This paper focused on some widely adopted neural language models of Brazilian Portuguese, and evaluated their performance in lexical, syntactic and

Table 13 Summary of grammatical proficiency tests

Test	P@1	P@10
Nominal agreement	77.53%	89.67%
Verb agreement	79.56%	59.55%
Subject agreement	83.38%	75.70%
Connectors	100.00%	54.17%
Impersonal verbs	98.78%	86.77%
Passive	86.65%	78.38%

semantic tasks. We developed a dataset with evaluation items that cover two particular areas: MWE inventory of idiomatic items, and proficiency in 6 gmatical tasks.

Overall, the larger and language-specific models performed better in the MWE task. Although mBERT shows good capacities in different NLP tasks, the model captured little to no information to predict idiomatic items in Brazilian Portuguese. This is most likely due to the curse of multilinguality (Conneau et al., 2019). Despite the small difference in performance between BERTimbau Large and Base, the larger version presented a better recognition of MWE. However, exploring more outputs from the BERTimbau Base might lead to the same performance as BERTimbau Large.

The grammatical tests showed that BERTimbau Large has a good overall precision in generating candidates for masked items, especially when we looked only at the top 1 candidates, going up (or very close) to 100% precision both for connectors and impersonal verbs. Even so, for tasks that required morphosyntactic agreement, precision decreased, and the worse results (below 80% among the top 1 candidates) were reported for nominal and verb agreement. The case of verb agreement was especially challenging for the model, because it consistently failed to produce good results for certain personal pronouns (first person plural, and second person singular and plural), which could be a sign of poor morphosyntactic generalisation, or be a side effect of the training corpus.

We adopted two evaluation criteria, one more strict, considering only the best candidate, and a broader one, which includes the top 10 candidates. Moreover, by considering not only the expected target word forms during the evaluation, but also considering alternative, grammatically correct outputs, we were able to detect the capacity of the model to produce different types of word forms for different contexts. Although these "deviant" word forms did not represent much of the correct responses, amounting to around 5% in most cases, they showed that a syntactic cue might not be as strong as the overall context of the sentence, as argued by Gulordava et al. (2018).

The results obtained confirm that the model achieves proficiency levels in tasks that do not require morphosyntactic agreement. However, it still lacks quality in certain items, in particular related to feminine singular forms. We also observed that there are instances (e.g., nominal agreement) in which the model has higher confidence (i.e., higher probability) in inadequate responses. All these evaluations led to a profile of the model's linguistic information.

This study contributed towards understanding deep learning models' representation of the Portuguese language. For that, we opted for a black-box perspective by

assessing the linguistic generalisation of language models. One of the advantages of this approach is its independence from model architectures. In BERT architecture, we observe evidence that points to a superior performance of the model in generalisable tasks than in non-generalisable ones, but the results are not enough for a final conclusion about the model learning process. In this sense, further studies should be done on this architecture in order to verify how the model creation process differentiates generalisable and non-generalisable cases. This could be done, for instance, by exploring explainability methods, such as studying attention layers.

As future work, the battery of tests can be extended to other linguistic aspects, such as selectional preferences for verbs, and inventory of collocations and terminology. In addition, an investigation can be conducted to analyse whether possible biases in the distribution of the training data can affect the performance of these patterns. Finally, a multilingual version of the test set can be developed, adapting it to closely related languages that share some of these linguistic patterns, and assessing if language proximity can be beneficial for few-shot learning scenarios.

Acknowledgements This research has partially been funded by a research convention with France Éducation International. It is also partly funded by EPSRC (project EP/T02450X/1 Modeling Idiomaticity in Human and Artificial Language Processing), by the Royal Society (project NAF/R2/202209), and by Research England, in the form of the Expanding Excellence in England (E3) programme.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Abdaoui, A., Pradel, C., & Sigel, G. (2020). Load what you need: Smaller versions of mutililingual bert. In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pp. 119–123.

Bacon, G., & Regier, T. (2019). Does bert agree? Evaluating knowledge of structure dependence through agreement relations. arXiv:1908.09892.

Bakarov, A. (2018). A survey of word embeddings evaluation methods. arXiv:1801.09536.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, Curran Associates Inc., pp. 4356–4364,

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. arXiv:1911.02116.

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., & Todirascu, A. (2017). Multiword expression processing: A survey. *Computational Linguistics, 43*(4), 837–892.

Cordeiro, S., Villavicencio, A., Idiart, M., & Ramisch, C. (2019). Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics, 45*(1), 1–57.

De Beaugrande, R.-A., & Dressler, W. U. (2011). Einführung in die textlinguistik. In *Einführung in die Textlinguistik*. Max Niemeyer Verlag.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805.

Dinan, E., Fan, A., Wu, L., Weston, J., Kiela, D., & Williams, A. (2020). Multi-dimensional gender bias classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 314–331.

Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics, 8*, 34–48.

Garcia, M., Kramer Vieira, T., Scarton, C., Idiart, M., & Villavicencio, A. (2021a). Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels. 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)*. Association for Computational Linguistics, , pp. 2730–2741.

Garcia, M., Kramer Vieira, T., Scarton, C., Idiart, M., & Villavicencio, A. (2021b). Probing for idiomaticity in vector space models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, , pp. 3551–3564.

Goldberg, Y. (2019). Assessing bert's syntactic abilities. arXiv:1901.05287.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36*(2), 193–202.

Gulordava, K., Bojanowski, P., Grave, É., 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)*, pp. 1195–1205.

Halliday, M. A. K., & Hasan, R. (2014). *Cohesion in English*. London: Routledge.

Kassner, N., & Schütze, H. (2020). Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 7811–7818.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwel, D. (2004). The sketch engine. In *Proceedings of the Eleventh EURALEX International Congress*.

Koch, I. G. V. (1988). Principais mecanismos de coesão textual em português. *Cadernos de Estudos Linguísticos, 15*, 73–80.

Koch, I. G. V. (1999). *A coesão textual*. London: Editora Contexto.

Kumar, V., Bhotia, T. S., Kumar, V., & Chakraborty, T. (2020). Nurse is closer to woman than surgeon? Mitigating gender-biased proximities in word embeddings. *Transactions of the Association for Computational Linguistics, 8*, 486–503.

Kurita, K., Vyas, N., Pareek, A., Black, A. W., & Tsvetkov, Y. (2019). Quantifying social biases in contextual word representations. In *1st ACL Workshop on Gender Bias for Natural Language Processing*.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, arXiv:1909.11942.

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.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv:1907.11692.

Louwerse, M. (2002). An analytic and cognitive parameterization of coherence relations.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting?: a review and empirical validation. *Journal of Memory and Language, 92*, 57–78.

Marcus, G. (2020). The next decade in AI: four steps towards robust artificial intelligence. *CoRR*, arXiv:2002.06177.

Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1192–1202.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography, 3*(4), 235–244.

Mueller, A., Nicolai, G., Petrou-Zeniou, P., Talmina, N., & Linzen, T. (2020). Cross-linguistic syntactic evaluation of word prediction models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5523–5539.

Muniz, M. C., Maria das Graças, V. N., & Laporte, E. (2005). Unitex-pb, a set of flexible language resources for brazilian portuguese. In *Workshop on Technology on Information and Human Language (TIL)*, pp. 2059–2068.

Nivre, J., Agić, Ž., Ahrenberg, L., Antonsen, L., Aranzabe, M. J., Asahara, M., Ateyah, L., Attia, M., Atutxa, A., Augustinus, L., & others (2017). Universal dependencies 2.1.

Oshikawa, R., Qian, J., & Wang, W. Y. (2020). A survey on natural language processing for fake news detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 6086–6093.

Pasquer, C., Savary, A., Ramisch, C., & Antoine, J.-Y. (2020). Verbal multiword expression identification: Do we need a sledgehammer to crack a nut? In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, pp. 3333–3345.

Perini, M. A. (2010). *Gramática do português brasileiro*. London: Parábola Ed.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2002)*. Springer, pp. 1–15.

Şahin, G. G., Vania, C., Kuznetsov, I., & Gurevych, I. (2020). LINSPECTOR: Multilingual probing tasks for word representations. *Computational Linguistics, 46*(2), 335–385.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. arXiv:1910.01108.

Sardinha, T. B. (2010). Corpus brasileiro. *Informática, 708*, 0–1.

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., & Turchi, M. (2021). Gender bias in machine translation. *Transactions of the Association for Computational Linguistics, 9*, 845–874.

Scarton, C., & Aluısio, S. M. (2010). Coh-metrix-port: A readability assessment tool for texts in brazilian portuguese. In *Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language, Extended Activities Proceedings, PROPOR*, vol. 10.

Schneider, E. T. R., de Souza, J. V. A., Knafou, J., Oliveira, L. E. S. e., Copara, J., Gumiel, Y. B., Oliveira, L. F. A. d., Paraiso, E. C., Teodoro, D., & Barra, C. M. C. M. (2020). BioBERTpt: A Portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, pp. 65–72.

Scholivet, M., & Ramisch, C. (2017). Identification of ambiguous multiword expressions using sequence models and lexical resources. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*. Association for Computational Linguistics, pp. 167–175.

Schrimpf, M., Blank, I., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J., & Fedorenko, E. (2020). Artificial neural networks accurately predict language processing in the brain. *BioRxiv*.

Souza, F., Nogueira, R., & Lotufo, R. (2019). Portuguese named entity recognition using bert-crf. arXiv:1909.10649.

Souza, F., Nogueira, R., & Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *Brazilian Conference on Intelligent Systems*. Springer, pp. 403–417.

Su, Q., Wan, M., Liu, X., & Huang, C.-R. (2020). Motivations, methods and metrics of misinformation detection: An nlp perspective. *Natural Language Processing Research, 1*, 1–13.

Sylak-Glassman, J. (2016). The composition and use of the universal morphological feature schema (unimorph schema). Johns Hopkins University.

Tayyar Madabushi, H., Gow-Smith, E., Garcia, M., Scarton, C., Idiart, M., & Villavicencio, A. (2022). SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics, pp. 107–121.

Tayyar Madabushi, H., Gow-Smith, E., Scarton, C., & Villavicencio, A. (2021). AStitchInLanguageModels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, pp. 3464–3477.

Vale, O., & Baptista, J. (2015). Novo dicionário de formas flexionadas do unitex-pb: Avaliação da flexão verbal (new dictionary of inflected forms of unitex-pb: Evaluation of verbal inflection). In *Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology*, pp. 171–180.

Vulić, I., Baker, S., Ponti, E. M., Petti, U., Leviant, I., Wing, K., Majewska, O., Bar, E., Malone, M., Poibeau, T., Reichart, R., & Korhonen, A. (2020). Multi-SimLex: A large-scale evaluation of multilingual and crosslingual lexical semantic similarity. *Computational Linguistics, 46*(4), 847–897.

Wagner Filho, J. A., Wilkens, R., Idiart, M., & Villavicencio, A. (2018). The brWaC corpus: A new open resource for Brazilian Portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 4339–4344.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics, 8*, 377–392.

Wilkens, R., Zilio, L., Cordeiro, S. R., Paula, F., Ramisch, C., Idiart, M., & Villavicencio, A. (2017). LexSubNC: A dataset of lexical substitution for nominal compounds. In *IWCS 2017: 12th International Conference on Computational Semantics: Short papers*.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems, 32*, 1–10.