This is a BERT. Now there are several of them. Can they generalize to novel words?

Coleman Haley
Department of Cognitive Science
Johns Hopkins University
chaley7@jhu.edu

Abstract

Recently, large-scale pre-trained neural network models such as BERT have achieved many state-of-the-art results in natural language processing. Recent work has explored the linguistic capacities of these models. However, no work has focused on the ability of these models to generalize these capacities to novel words. This type of generalization is exhibited by humans (Berko, 1958), and is intimately related to morphology–humans are in many cases able to identify inflections of novel words in the appropriate context. This type of morphological capacity has not been previously tested in BERT models, and is important for morphologically-rich languages, which are under-studied in the literature regarding BERT's linguistic capacities. In this work, we investigate this by considering monolingual and multilingual BERT models' abilities to agree in number with novel plural words in English, French, German, Spanish, and Dutch. We find that many models are not able to reliably determine plurality of novel words, suggesting potential deficiencies in the morphological capacities of BERT models.

1 Introduction

In recent years, large-scale pre-trained neural network models have transformed the landscape of natural language processing (NLP) research. This approach to NLP became prominent after several models such as BERT (Devlin et al., 2019) achieved new state of the art performance on a wide range of NLP tasks such as natural language inference. The successful performance of BERT and other models like it on natural language understanding tasks suggests that they may be learning valuable general linguistic competencies. However, it is not clear whether these models are able to generalize these competencies to unseen words. With the large training sets of these models (3.3 billion tokens in Devlin et al. (2019)), their state-of-the-art-establishing performance may feasibly have been achieved without ever being tested on a word that was not in the training set.

Nevertheless, BERT may need be concerned about unseen words. Increasingly, there is an interest in creating BERT and BERT-like models trained on large corpora of languages other than English. In comparison to English, many of the world's languages exhibit a much greater amount of inflectional morphology. However, most of the results motivating this explosion of BERT models are in English NLP. It is unclear, then, how well BERT will generalize to languages with complex morphology. While BERT models are being developed for other languages, many of these models have been less comprehensively evaluated than English BERT. For instance, the publicly available Turkish (Schweter, 2020) BERT model (one of the most morphologically complex languages for which a BERT model is available) has only been evaluated on named entity recognition and part-of-speech tagging. It is unclear, then, how well the model would fare on more complex NLP tasks.

In this work, we investigate BERT's ability to capture this type of information by studying its ability to identify the correct plural form of novel words in English, French, Spanish, Dutch, and German. We find that BERT is able to distinguish plural and singular forms to perform number agreement significantly above chance in all languages. However, many BERT models perform substantially worse on novel words than on words in the training set, even when prompted with an example that shows the singular form, a task which humans are known to be capable of (Berko, 1958). This indicates that even simple morphological capacities are not reliably acquired in a human-like way in the BERT training paradigm, showing room for improvement in future models.

Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 333–341
Online, November 20, 2020. ©2020 Association for Computational Linguistics

2 Background

BERT is part of a growing research direction of pretraining deep learning models, often a variant of a "Transformer" architecture (Vaswani et al., 2017), on large amounts of natural language data using some variant of a language modelling objective. This line of research includes other such successful models as ELMo (Peters et al., 2018) and XLNet (Yang et al., 2019). All of these models are trained on very large corpora, with ELMo being the smallest (trained on 1 billion tokens in Peters et al. (2018) – in contrast to BERT's 4 billion tokens in Devlin et al. (2019)). All of these models are also highly computationally intensive to train, so it is desirable to avoid training new BERT models.

BERT uses a transformer-based architecture, making it bidirectionally sensitive. It is trained on a masked language modelling objective, meaning that it takes in as input a sequence with some words replaced with a `[MASK]` token, and is expected to output the original sequence. To enable this, a final fully connected layer and softmax is added after the transformer encoder to produce the desired output. This means BERT is "out of the box" capable of answering exactly those questions that can be posed as replacing `[MASK]` tokens.

BERT-like models are also generally so-called *open-vocabulary* language models, meaning they can assign a probability to any string. This enables them to give probabilities to novel words and novel forms of known words, giving BERT the capacity to learn morphological generalizations. This is achieved through the use of subword segmentation, in which a strategy such as byte-pair encoding (BPE) (Sennrich et al., 2016) or Unigram LM segmentation (such as WordPiece (Kudo, 2018) and the related SentencePiece) is used to turn words into a sequence of multi-character tokens.

These segmentation strategies use statistical methods to determine which multi-character tokens are added to their vocabularies, meaning that high-frequency sub-word strings will more likely be added as tokens. These tokens may or may not correspond to morpheme boundaries. If they do not, then models that rely on them will encounter the same morpheme expressed in many distinct tokens, requiring the model to learn agreement for *all* tokens which may contain, e.g., the plural affix. This may mean that uncommon segments containing inflectional affixes will be less reliable in agreement, since they have no relation in representation to frequently-occurring subwords containing the same inflection.

2.1 BERT and linguistic competence

Previous work has explored the types of generalizations predicted by linguistic and psycholinguistic theory that have been learned by the English BERT models. This work has focused primarily on syntactic generalizations. Initial work by Goldberg found that BERT models showed promise at modelling short- and long-distance subject-verb agreement as well as reflexive anaphora phenomena (Goldberg, 2019). van Schijndel et al. (2019) revisited these results without giving a bidirectional context to BERT and found it performed at best no better than existing LSTM models (contrasting with Goldberg's work). Ettinger (2020) differentiates her work from these works by noting their primarily syntactic focus, and promises to test more diverse linguistic capacities, but focuses on semantic and pragmatic capacities, showing among other things that BERT fails to fully model the meaning of negation. Recently, Mueller et al. (2020) presented cross-linguistic targeted syntactic evaluation of BERT, but only considered multilingual BERT. Most of the work on the formal linguistic capacities has not considered monolingual BERT models for languages other than English (one recent exception being Edmiston (2020)).

Very recently, a few works have considered the morphological aspects of BERT. Bostrom and Durrett (2020) argue that byte-pair encoding less faithfully expresses English morphology than Unigram segmentation, and show a performance improvement in downstream tasks with a unigram-segmentation-based BERT model. Hofmann et al. (2020) show that BERT can be fine-tuned with a classification layer to complete a derivational morphology cloze task, finding that imposing morpheme boundaries with hyphenation on the input side ultimately improved BERT's performance at this task. Finally, Edmiston (2020) investigates several monolingual BERT models for representations of morphological information. Edmiston shows that many morphological features can be extracted by training a simple classifier on a BERT layer. He also identifies a small number of attention heads in each model that seem to pay attention to the morphologically marked words in agreement phenomena over other words. However, this agreement experiment makes no attempt to isolate the morphological information from words which BERT has seen, allowing for the possibility of morphological "memorization" rather than true human-like generalization.

Previous work in psycholinguistics has investigated the human capacity for morphological generalization, and it is this work we intend to build on to explore BERT's morphological capacity. Specifically, Berko (1958) presents the Wug test, a simple test for productive morphology in which speakers are prompted with a sentence containing one form of an unknown word and prompted to complete a sentence with another form. We present a task inspired by this one in which the ability to *recognize* an unseen form of a word is probed through the ability to correctly *agree* with that word's form. In this work, we specifically investigate subject-verb number agreement.

3 Methods

This work focuses on BERT's ability to recognize novel words as singular or plural. This construction was chosen for its testability (through number agreement on verbs) and its disparity in complexity between languages. In English, French, Dutch and Spanish, a large majority of plurals are derived according to rules that can be expressed simply in terms of adding a suffix corresponding to the suffix of the base noun. Further, in French and Spanish, the plurality of a noun is unambiguous if it is preceded by a determiner.

3.1 Plural formations of the languages

In written English, the plural of most nouns is formed by one of three strategies: either 1. *-s* is added to the end of the noun, 2. *-es* is added to the end, or 3. a copy of the final letter followed by *-es* is added to the end. Strategy 2 is used after sibilant sounds, and Strategy 3 is generally used after sibilant sounds which are preceded by a lax vowel. Strategy 1 is used in all other cases (except known irregulars). The words selected for this study were chosen such that their spelling indicates an obvious phonetic realization, and that they are distributed across these 3 strategies.

The French and Spanish plural constructions are arguably simpler than in English. In French, plural nouns are generally formed by adding *-s* to the end; unless the noun ends in *s*, *z*, or *x*, in which case nothing is added, in *eau*, in which case *-x* is added, or in *-al* or *-ail* in which case the suffix may be removed and *-aux* added. In addition to inflecting the word, French marks plurality in its definite determiner, making it unambiguous from the determiner whether a noun is singular or plural.

On the other hand, the German plural construction is significantly more complex than in English. Like French and Spanish, German marks for plurality in the determiner, but the determiner used to indicate plurality in the nominative case is shared with that used to mark feminine noun gender, meaning that noun gender cannot be determined purely from the determiner. Consider for example *the woman→the women*, which in Spanish is *la mujer→las mujeres*, but in German is *die Frau→die Frauen*. Further, German uses several different strategies to form the plural, including adding nothing to the word (-∅), adding *-e*, adding *-(e)r*, adding *-(e)n*, and adding *-s*. These strategies (with the exception of *-(e)n*) may also be combined with "umlautification" of the stressed vowel in the noun, yielding a total of 7 possible plural markers, none of which consitute a majority of examples (Köpcke, 1988; Wiese, 2000).

Singular form	Plural form
das Fett	die Fette
das Brett	die Bretter
das Bett	die Betten
der Sohn	die Söhne
der Thron	die Throne

Table 1: The German plural cannot be predicted from the form of the singular word. Here, we see similar singular words that form the plural in different ways.

The literature on the German plural generally considers it to be a phenomenon over lexical classes which are not phonologically predictable. Several tendencies can be observed in German plural formation, though few are universal. For example, nouns ending in *-e* typically form their plural by adding *-n* (Trommer, 2020). Nevertheless, even near-minimal pairs of nouns may form their plural in distinct ways (see Table 1). Indeed, adult German speakers often vary widely in their choices for novel words (Zaretsky et al., 2013; McCurdy et al., 2020). Accordingly, substantial prior work has suggested the German plural may be a challenging pattern for neural networks to learn (Feldman, 2005; Marcus et al., 1995; McCurdy et al., 2020).

The Dutch plural represents an interesting intermediate case. As in German, the determiner gives

Condition	Stimulus	Candidates
No prime, real words	The author knows many different foreign languages and [MASK] playing tennis with colleagues.	enjoy/**enjoys**
Prime, real words	This is a pilot. the pilots [MASK].	**laugh**/laughs
Prime, non-words	This is a bik. the biks [MASK].	**laugh**/laughs

Table 2: Sample agreement stimuli in representative conditions in English. Correct completion is in **bold**.

Model	Language	Parameters	Training tokens	Tokenization
BERT$_{\text{BASE}}$ (Devlin et al., 2019)	English	110M	3.3B	WordPiece 30k
CamemBERT (Martin et al., 2020)	French	110M	32.7B	SentencePiece 32k
FlauBERT (Le et al., 2020)	French	138M	12.8B	BPE 50k
BETO (Cañete et al., 2020)	Spanish	110M	3B	BPE 32k
BERTje (Vries et al., 2019)	Dutch	110M	2.4B	SentencePiece 30k
Deepset [1]	German	110M	1.8B	SentencePiece 30k
dbmdz[2]	German	110M	2.4B	SentencePiece 30k
mBERT[3]	All	110M	–	WordPiece 110k

Table 3: The models used in this work and associated statistics. Note that the SentencePiece and WordPiece segmentation methods are different implementations of the same algorithm, described in Kudo (2018).

some ambiguous information about plurality, with the determiners *het* and *de* both being used for singular nouns, but only *de* used with plural nouns. The plural in Dutch is constructed using either the ending *-en* or *-s*. Generally, *-en* is used to form the plural of nouns ending with a stressed syllable, and *-s* is used with nouns ending in an unstressed syllable, although this generalization is not perfect (van der Hulst and Kooij, 1998).

3.2 Experimental setup

This experiment probes the ability of BERT to *recognize* the plurals of novel words as such. We probe this indirectly, though a number agreement task following the setup in van Schijndel et al. (2019). As in that study, We use the challenge set from Marvin and Linzen (2018) as a starting point. Number agreement was chosen as a task because it is not fully understood how to treat BERT as a generative model. Therefore, we probe plural recognition as an auxiliary task which BERT has been shown to succeed at (Goldberg, 2019). This task is formulated as a forced choice between a plural verb form or singular verb form.

The Marvin and Linzen (2018) challenge set was translated into English, German, Dutch, Spanish, and French by fluent speakers with an elementary background in formal linguistics. These languages each have a singular-plural distinction and subject-verb number agreement. Syntactic constructions not possible in all five languages were omitted. Some verbs in each dataset were changed to ensure each verb was a single token for all models in that language. The datasets in each language were then modified to replace the subject of the targeted verb with a non-word. For each language, 24 non-words were used. English, French, Spanish, and Dutch non-words were manually created by fluent speakers, while the 24 German non-words were taken from McCurdy et al. (2020), to account for the fact that the German plural of non-words is known to be inconsistent across speakers. The plural formation chosen by a plurality of German speakers in McCurdy et al. (2020) for each German non-word was used; genders were chosen to be distributed uniformly.

The BERT models were evaluated on number agreement on the original datasets and the non-word datasets. Models were evaluated bidirectionally, as in Goldberg (2019), to provide a maximally-charitable estimate of BERT's morphological capacity in each language.

Finally, models were reevaluated on the non-word data with a "prime" for the non-word. In English, the prime takes the form of the sentence

[1] https://deepset.ai/german-bert

[2] https://github.com/dbmdz/berts

[3] https://github.com/google-research/bert/blob/master/multilingual.md

		Real words		Non-words	
Language	Model	No prime	Prime	No prime	Prime
English	$\text{BERT}_{\text{BASE}}$	1.00	1.00	0.87	0.90
French	CamemBERT	0.99	0.98	0.98	0.99
	FlauBERT	0.92	0.97	0.89	0.98
	mBERT	0.97	0.98	0.99	0.99
Spanish	BETO	0.98	0.87	0.90	0.80
	mBERT	0.89	0.89	0.81	0.84
Dutch	BERTje	1.00	0.98	0.85	0.79
	mBERT	0.93	0.93	0.77	0.81
German	deepset	1.00	1.00	0.70	0.72
	dbmdz	1.00	0.99	0.75	0.79
	mBERT	1.00	1.00	0.80	0.75

Table 4: Agreement accuracy on simple sentences (e.g. “The author laughs.”).

“This is a _____”, where the blank was replaced with the singular form of the novel noun in the target sentence and the appropriate determiner for the noun’s gender was selected. This construction was translated into each language.

While it may seem unintuitive that BERT could benefit from the use of this prime at test time, since it is unable to adjust its weights, with self-attention it is theoretically possible to encode a simple “rule” for using the number of a noun seen for the first time (as disambiguated via subject-verb agreement) to influence number agreement for a noun with a similar form. It is this possibility, as well as the human capacity for this type of generalization, that motivates this condition. Examples of stimuli in each condition for English are presented in Table 2.

We consider several cased BERT models, both monolingual and multilingual. The $\text{BERT}_{\text{BASE}}$ size was used for all languages for comparability between models, as not all languages have a $\text{BERT}_{\text{LARGE}}$ model available. The models used are summarized in Table 3 Experiments were run on a single Nvidia GeForce GTX 1080 Ti, and take under an hour to run.[4]

4 Results

Table 4 presents the results for the simple agreement tests with bidirectional context. Here, “simple” refers to sentences consisting of a subject immediately followed by an intransitive verb (e.g., “The man laughed.”). The number of singular sentences ranged between 212-672 depending on language and non-word condition. As in Goldberg (2019), ceiling performance is found on the original dataset in English. CamemBERT also performed near ceiling. Since the task is a forced choice between 2 verb forms (singular or plural), and there are an equal number of singular and plural subjects, chance performance is 0.5. Agreement performance on the non-word sentences was much better than chance, even without the inclusion of a prime for the non-word ($p < 0.001$). This indicates the model is often able to guess whether an noun not seen in training is likely to be singular or plural. Notably, not all models across languages succeeded completely at subject-verb agreement even with real words on simple sentences–FlauBERT for French and mBERT for Dutch and Spanish achieved less than 0.95 accuracy in this simple task.

Cross-linguistically, there is no consistent trend in whether the model is able to use the prime to achieve better performance. While FlauBERT was the only model to achieve statistically significant gains ($p < 0.01$) in the non-word case with the addition of the prime, this gain was also significant ($p < 0.05$) in the real word case, suggesting deeper issues with this model’s agreement capabilities generally. Many models were slightly hurt by the inclusion of the prime, suggesting that they may be spuriously agreeing with the prime, even across a sentence boundary.

As one might expect, the German BERT models had the lowest average performance on the non-word conditions, with no model surpassing 0.80

[4] Code for generating the dataset and replicating the experiments is available at https://github.com/ColemanHaley/BERT-novel-morphology.

accuracy. However, only French models achieved an accuracy of greater than 0.90 in any non-word case. Given that the correct form for French and Spanish agreement can be determined from the noun's article alone, it is surprising that the Spanish models do not fully utilize this heuristic; this heuristic may explain the high French performance.

mBERT performs about as well as the monolingual BERT models in French and German, but performs worse in Dutch and Spanish. In no case did it significantly out-perform a monolingual model ($p > 0.1$).

5 Discussion

To investigate whether the lower novel-word performance was related to the segmentations of the novel words, we measured how often each non-word was associated with an error. We found inconsistent results across models. BERTje was found to perform especially poorly on 4 out of 24 non-words, incorrectly choosing a singular verb for a plural form of the word 93% of the time. On investigating the segmentations, these words were found to be segmented to "`[UNK]`" by the tokenizer. This model uses the standard SentencePiece Unigram tokenizer[5], ostensibly the same as many of these other models. Typically, this tokenizer is considered to be open-vocabulary, yet it fails to segment these subwords, indicating that this is not strictly true for this very popular implementation. If these 4 words are disregarded, in the no-prime case this model achieves an accuracy of 0.93, the highest of all non-French models across languages. While this error is of substantial concern, it affects only the Dutch BERTje results, as no other models were found to have this behavior.

Other models were found to frequently fail on the plural or singular forms of some words, such as BETO, which 50% of the time identified "comanas" as a singular word form. Some models, such as the English model and FlauBERT, instead seemed to be uncertain about the plurality of all forms, making a moderate amount errors at roughly equal rates across non-words. In the case of the English model, the model has a bias towards plurality, with plural accuracy 0.19 greater than singular accuracy; however, FlauBERT makes agreement errors at roughly equal rates regardless of whether the subject is singular or plural.

With the German models, accuracy was > 0.88 for singular non-words, many of which are disambiguated by their determiner. Accordingly, most errors are plural word forms which the model identified as singular. Both monolingual German models showed a pattern of having many plural forms that were identified as singular $> 70\%$ of the time. Most of the remaining forms in each model were correctly identified as plural $> 70\%$ of the time, indicating that these models are relatively certain in their predictions. Unfortunately, no clear relationship to how closely the segmentation pattern matches the morphology was found with whether the correct verb is selected for a given non-word. However, it is possible that the *frequency* of the final subword segment occurring as a plural affix in a German corpus would be more predictive of which segments are likely to result in errors.

5.1 Non-linguistic factors

Although these results are largely consistent with the linguistic hypotheses discussed in Section 3.1, there is an uneven amount of training data across the models and languages. Notably, the French monolingual models used the most data, with German models using the least. However, this relationship is different within the mBERT model itself. This model, being trained on the Wikipedia dumps of each language, has the most data for English, followed by German, then French, then Spanish, then Dutch. While this does not completely disentangle the effects of training size (e.g., for the low Dutch performance), it does indicate that the disparity between model performances in, e.g., French and German cannot be explained solely by this factor. Further, almost all models use the same vocabulary size and number of parameters, with only FlauBERT being substantially larger, so this is also likely not a major factor. Therefore, it seems plausible that a primary driver of the differences in model performance between languages reported here is the language's plural construction.[6]

5.2 Implications for BERTology

While this study is primarily focused on the morphological and novel-word generalization capacities of BERT, it also investigates more models and languages than prior work on BERT's linguistic capacities–no previous work has looked at

[5] `https://github.com/google/sentencepiece`

[6] Additional architectural differences between the models are described in Appendix A.

more than one monolingual model for a single non-English language. The results here strongly suggest that the field of "BERTology" needs to consider the generality of their claims across not just different languages, but even across different BERT models developed on the same language. Even models based on the same architecture and within the same language, such as the German deepset and dbmdz models show different results on this simple task. While French subject-verb agreement can be determined solely from the subject determiner, FlauBERT achieved only 0.92 accuracy on even simple French sentences, while both mBERT and CamemBERT achieve accuracies higher than 0.95. This suggests that even in languages where subject-verb agreement is relatively simple, the BERT training objective alone cannot guarantee total generalization, even when the model performs well on downstream tasks. This suggests a need for greater scruitiny of claims of BERT's linguistic capacity that evaluate only one or two models.

5.3 Potential practical implications

Finally, To consider the relation of these theoretical findings to real-world performance, we ran the German experiments on real nouns again without capitalizing nouns. While all nouns are capitalized in formal German, this case serves as an example of a simple typo that might occur in real data. Agreement accuracy dropped from 1.00 in all 3 German models to 0.90, 0.79, and 0.89 in the mBERT, deepset, and dbmdz models respectively. This indicates that BERT's agreement faculty is highly sensitive to noise, failing to generalize even to highly plausible "non-words" (in this case, uncased nouns). This casts substantial doubt on the generality of BERT's extensively studied agreement competencies.

6 Conclusion

These results suggest that BERT models have some understanding of morphology when applied to novel words (or at least the plurals in a few Germanic and Romance languages). Performance is much significantly better than chance in simple agreement cases, even when no prime is given. This shows that the BERT models have learned something about what plural and singular forms "look like." However, non-word performance is not helped especially by the inclusion of a priming sentence, indicating that the BERT models in question may not have learned to recognize new words and apply rules to them, as humans might. Further work should investigate what types of primes affect the performance and how.

The model performance here represents a *best case* for BERT's morphological capacity on novel words. The plural construction is extremely common in text and is connected to phenomena like agreement which additionally pressures it to be learned on a non-semantic level. Further, the simple sentences studied here allow for the potential of n-gram-level agreement heuristics which are not possible in the general case.

That BERT struggles to capture morphology in this way is likely not due to a lack of training data. There are two potential culprits: the tokenization method and the training objective. The FlauBERT results especially indicate that the masked language modeling objective may not sufficiently encourage agreement. Cross-linguistically, the models seem not to have picked up on how to use the information in the prime. In addition, the subword tokenization methods used by BERT and BERT-like models make the morphology learning problem significantly more complicated. This is because the plural morpheme is connected to some number of final characters of a word as a single token, meaning even plurals formed in the same way may be represented differently. This work points to a need for subword segmentation strategies that more closely mirror a language's morphology than current approaches like Unigram segmentation or byte-pair encoding. In this aspect of language, there remains a large gap between BERT's behavior and human performance.

Acknowledgments

The author would like to thank Paul Schauenburg for his grammaticality judgements and extensive help with preparing the plural non-word data. I would also like to thank Tom McCoy, Tal Linzen, and Colin Wilson for their helpful and clarifying comments. This work was supported in part by a Provost's Undergraduate Research Award from Johns Hopkins University.

References

Jean Berko. 1958. The child's learning of english morphology. *Word*, 14:150–177.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. *Computing Research Repository*, arXiv:2004.03720. To appear in *Findings of ACL: EMNLP 2020*.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, and Jorge Pérez. 2020. Spanish pre-trained BERT model and evaluation data. To appear in PML4DC at ICLR 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Daniel Edmiston. 2020. A systematic analysis of morphological content in bert models for multiple languages. *Computing Research Repository*, arXiv:2004.03032.

Allyson Ettinger. 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.

Naomi Feldman. 2005. Learning and overgeneralization patterns in a connectionist model of the German plural. Master's thesis, University of Vienna.

Yoav Goldberg. 2019. Assessing BERT's syntactic abilities. *Computing Research Repository*, arXiv:1901.05287.

Valentin Hofmann, Janet B. Pierrehumbert, and Hinrich Schütze. 2020. DagoBERT: Generating derivational morphology with a pretrained language model. *Computing Research Repository*, arXiv:2005.00672.

H. G. van der Hulst and J. Kooij. 1998. Prosodic choices in plural formation in Dutch. In W. Kehrein and R. Wiese, editors, *Phonology and morphology of the Germanic languages*, pages 187–198. Niemeyer, Tübingen.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Klaus-Michael Köpcke. 1988. Schemas in German plural formation. *Lingua*, 74(4):303 – 335.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for french. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *Computing Research Repository*, arXiv:1907.11692.

G. F. Marcus, U. Brinkmann, H. Clahsen, R. Wiese, and S. Pinker. 1995. German inflection: the exception that proves the rule. *Cogn Psychol*, 29(3):189–256.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Kate McCurdy, Sharon Goldwater, and Adam Lopez. 2020. Inflecting when there's no majority: Limitations of encoder-decoder neural networks as cognitive models for German plurals. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1745–1756, Online. Association for Computational Linguistics.

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

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Marten van Schijndel, Aaron Mueller, and Tal Linzen. 2019. Quantity doesn't buy quality syntax with neural language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5831–5837, Hong Kong, China. Association for Computational Linguistics.

Stefan Schweter. 2020. *BERTurk - BERT models for Turkish*. Zenodo.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Jochen Trommer. 2020. The subsegmental structure of German plural allomorphy. *Natural Language & Linguistic Theory*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT Model. *Computing Research Repository*, arXiv:1912.09582.

R. Wiese. 2000. *The Phonology of German*. Oxford Linguistics. Oxford University Press.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5753–5763. Curran Associates, Inc.

Eugen Zaretsky, Benjamin P. Lange, Harald A. Euler, and Katrin Neumann. 2013. Differences in plural forms of monolingual German preschoolers and adults. *Lingue e Linguaggi*, 10:169–180.

A Additional model details

This appendix summarizes some additional differences between the models. It is not clear to the author that these would be related to the pattern of results presented here, but they are included so that interested readers need not hunt them down.

The models vary in whether they use an auxiliary task in addition to the masked language modelling (MLM) task described in the background. Some models use next-sentence prediction (NSP), in which the BERT model sees two sentences and must determine whether the second one follows the first. The initial BERT study indicated this improved performance, but subsequent work (Liu et al., 2019) found the opposite to be true, and many subsequent BERT models omit this objective. BERTje instead includes a sentence order prediction (SOP) task, in which the model is presented with two consecutive sentences which may be in their original order or may be swapped, and must predict if they are in the correct order. (Vries et al., 2019) claim the addition of this objective improves their performance on downstream tasks.

Another attribute of the models that vary is how they handle the masking in MLM. The original BERT model masked out a portion of its training data before training, so every time a sentence is encountered the masked segments are the same. Subsequent works such as Liu et al. (2019) utilize *dynamic masking*, where different segments are masked in different training epochs. This is often achieved by masking the training data a fixed number of times and cycling through them during training. Finally, some models utilize sub-word masking (SWM), in which individual subwords are masked independently, while other models use whole-word masking (WWM), where all subwords of a single word are always masked together.

Model	Objective(s)	Masking Strategy
BERT$_{\text{BASE}}$	MLM, NSP	Static, SWM
CamemBERT	MLM	Dynamic, WWM
FlauBERT	MLM	Dynamic, WWM
BETO	MLM	Dynamic, WWM
BERTje	MLM, SOP	Static, WWM
Deepset	MLM, NSP	Static, SWM
dbmdz	MLM, NSP	Static, SWM
mBERT	MLM, NSP	Static, SWM

Table 5: Additional details of models. MLM = masked language modeling, NSP = next sentence prediction, SOP = sentence order prediction, SWM = sub-word masking, WWM = whole-word masking .