How Well Can BERT Learn the Grammar of an Agglutinative and Flexible-Order Language? The Case of Basque.

Gorka Urbizu[1,2], **Muitze Zulaika**[1], **Xabier Saralegi**[1], **Ander Corral**[1]

[1] **Orai NLP Technologies**
[2] University of the Basque Country
{g.urbizu, m.zulaika, x.saralegi, a.corral}@orai.eus

Abstract

This work investigates the acquisition of formal linguistic competence by neural language models, hypothesizing that languages with complex grammar, such as Basque, present substantial challenges during the pre-training phase. Basque is distinguished by its complex morphology and flexible word order, potentially complicating grammar extraction. In our analysis, we evaluated the grammatical knowledge of BERT models trained under various pre-training configurations, considering factors such as corpus size, model size, number of epochs, and the use of lemmatization. To assess this grammatical knowledge, we constructed the BL2MP (Basque L2 student-based Minimal Pairs) test set. This test set consists of minimal pairs, each containing both a grammatically correct and an incorrect sentence, sourced from essays authored by students at different proficiency levels in the Basque language. Additionally, our analysis explores the difficulties in learning various grammatical phenomena, the challenges posed by flexible word order, and the influence of the student's proficiency level on the difficulty of correcting grammar errors.

Keywords: Neural Language Models, Grammar Learning, Minimal Pairs, BERT, Basque

1. Introduction

Neural Language Models (NLMs) based on the Transformer architecture have demonstrated effectiveness in acquiring skills related to human language use. These skills encompass two primary competencies: formal linguistic competence, which focuses on lexis and grammar, and functional linguistic competence, which involves cognitive skills such as reasoning and world knowledge essential for understanding and using language in real-world contexts. While NLMs exhibit proficiency in both areas, their performance is particularly remarkable in formal linguistic competence (Mahowald et al., 2023b).

NLMs typically acquire these diverse abilities and skills during the pre-training process, leveraging patterns found within training corpora. Focusing on the learning of formal linguistic competence, involving lexis and grammar, one might hypothesize that a language with a more complex grammatical structure poses a greater challenge during the pre-training phase. It becomes pertinent, then, to examine how different pre-training parameters —like corpus size, number of epochs, and model size— affect the process of grammar learning. In this study, we turn our attention to Basque, a language that, like many others, is characterized by its complex morphology (see Table 3) and flexible word order (For example, *"The cat ate the mouse"* in Basque can be formulated in six different orders: *"Katuak sagua jan zuen"*, *"Katuak jan zuen sagua"*, *"Sagua jan zuen katuak"*, *"Sagua katuak jan zuen"*, *"Jan zuen sagua katuak"*, *"Jan zuen katuak sagua"*.). These features could, in theory, complicate the grammar learning process.

In our analysis, we compared the grammatical knowledge of models trained under various pre-training configurations, all utilizing the BERT architecture. To assess this knowledge, we constructed the BL2MP (Basque L2 student-based Minimal Pairs) test set. We constructed this test set based on essays written by students enrolled in Basque courses. These essays contained teacher-provided corrections of grammatical mistakes, facilitating the extraction of pairs of uncorrected and corrected sentences. Utilizing these pairs, we developed the BL2MP. The grammatical phenomena highlighted in this test set, therefore, reflect the common challenges encountered by these learners.

The variables examined in the different pre-training configurations include:

- Training corpus size: How does the size of the training corpus influence grammar learning performance?
- Model size: Does increasing the number of model parameters enhance grammar learning?
- Multi-epoch training: Is grammar learning bolstered by multiple training epochs?
- Tokenization: In inflectional languages like Basque, does lemmatization-based tokenization aid in grammar learning?

LREC-COLING 2024, pages 8334–8348
20-25 May, 2024.

Additionally, we analyzed the relationship between the grammar learning process and the following variables:

- Word order: Does a flexible word order complicate the learning process?
- Grammatical phenomena: Are certain grammatical phenomena more challenging to learn than others?
- L2 student proficiency: Is there a link between the proficiency level of L2 students and the NLM's difficulty in learning specific grammatical phenomena?

The BL2MP dataset, along with the pre-training corpora (both raw and lemmatized), pre-trained models, and evaluation code is publicly available[1].

2. Related Work

2.1. NLMs and Grammar

Since the introduction of pre-trained Language Models with BERT (Devlin et al., 2019), several authors have proved that BERT is able to learn linguistic features like syntax and semantics (Tenney et al., 2019a; Jawahar et al., 2019; Niu et al., 2022). Zhang et al. (2021) shows that a pre-training corpus of 10M to 100M words is enough for a language model to acquire the linguistic capacities of syntax and semantics, much less than what is needed for commonsense reasoning and factual knowledge. Moreover, NLMs are not only able to learn the grammar of a single language, Chi et al. (2020) provides evidence that mBERT learns universal representations of syntactic dependencies and Acs et al. shows that these multilingual NLMs learn morphological information during the pre-training.

Huebner et al. (2021) and Cagatan (2023) examine the grammatical knowledge of small RoBERTa models (Liu et al., 2019) trained on 5M and 10M word corpora of language acquisition data, and they state that NLMs acquire grammatical knowledge comparable to that of a pre-trained RoBERTa-base, despite having significantly fewer parameters and requiring far less pre-training data.

Sinha et al. (2021) shows that NLMs are surprisingly word order invariant and success on downstream tasks mostly due to their ability to model higher-order word cooccurrence statistics. In contrast, Papadimitriou et al. (2022) proves that NLMs learn grammatical features which are needed in cases where lexical expectations are not sufficient and word order is crucial.

[1] https://github.com/orai-nlp/bl2mp

2.2. Evaluating Grammar Knowledge of NLMs

There are several approaches to evaluating the English grammar knowledge of NLMs, leading to the creation of various datasets for this purpose.

One of them is CoLA, The Corpus of Linguistic Acceptability (Warstadt et al., 2019), which is included as a binary classification task in the GLUE benchmark (Wang et al., 2018). Another similar dataset is Mixed Signals Generalization Set (MSGS) (Warstadt et al., 2020b), which contains 20 ambiguous binary classification tasks, suited for evaluating NLMs on linguistic features using the edge-probing technique (Tenney et al., 2019b).

However, both of these datasets rely on fine-tuning the NLMs for a binary classification task, which diverts from assessing the grammar knowledge NLMs acquire during pre-training. Consequently, other datasets have been proposed to evaluate NLMs using zero-shot approaches, which rely on the concept of minimal pairs.

The first benchmark of linguistic minimal pairs for English was BLiMP (Warstadt et al., 2020a), which includes minimal pairs of sentences that differ in a single grammatical aspect while holding all other linguistic factors constant. These sentences can be scored by NMLs (perplexity, loss...), and then compared, to measure the grammatical knowledge and syntactic abilities of NLMs. More recently, supplementary tasks have been published in the context of the babyLM shared task (Warstadt et al., 2023).

Moreover, there is also the Grammar Test Suite Huebner et al. (2021), comparable to BLiMP, which contains pairs of test sentences which isolate specific phenomena in syntax and morphology.

2.3. NLMs and Grammar of Non-English Languages

Most of the works and datasets on the topic focus on English, but evaluating the grammatical knowledge of NLMs is language-dependent, and each language raises its own challenges. Therefore, there have been several works to build similar minimal pair datasets, inspired by BLiMP, for other languages: CLiMP (Xiang et al., 2021) and SLING (Song et al., 2022) for Chinese, Alrajhi et al. (2022) for Arabic, JBLiMP for Japanese (Someya and Oseki, 2023) and LINDSEA for Indonesian and Tamil (Leong et al., 2023).

In addition, there are other datasets following the approach of CoLA, RuCoLA (Mikhailov et al., 2022) for Russian and ScandEval (Nielsen, 2023) for Danish, Swedish, Norwegian Bokmål, Norwegian Nynorsk, Icelandic and Faroese.

Lastly, the only related work for Basque that we are aware of is by Ravfogel et al. (2018), where they demonstrate that LSTMs can capture agree-

ment[2] in Basque. However, they also note that this task is more challenging compared to its English counterpart.

3. Dataset

The concept of Minimal Pairs, integral to the BLIMP dataset (Benchmark of Linguistic Minimal Pairs) (Warstadt et al., 2020a), serves as a linguistic benchmark aimed at evaluating the grammatical and syntactic capabilities of language models. BLIMP features sets of sentences, termed minimal pairs, that differ in just one grammatical aspect. Each pair contains a grammatically correct sentence alongside one that contains an error, allowing for the assessment of a language model's proficiency in discerning between these subtly distinct sentences.

In this paper, we introduce the BL2MP test set, designed to assess the grammatical knowledge of the Basque language, inspired by the BLIMP benchmark. The BL2MP dataset includes examples sourced from the bai&by[3] language academy, derived from essays written by students enrolled at the academy. These instances provide a wealth of authentic and natural grammatical errors, representing genuine mistakes made by learners and thus offering a realistic reflection of real-world language errors.

We randomly selected 1,800 sentences from student essays provided by the bai&by academy, adhering consistently to the "minimal pairs" criterion. To ensure a balanced diversity, we ensured an equal distribution of examples across three proficiency levels[4] (A: Beginner, B: Intermediate, and C: Advanced) and three error types (E1: Declension, E2: Verb, E3: Structure and Order) , as shown in Table 1. This approach aimed to represent a variety of proficiency levels and error types within the dataset. Examples of a minimal pair for each error type are included in Table 2.

Types	Levels	# of sentences
E1: Declension	A	200
	B	200
	C	200
E2: Verb	A	200
	B	200
	C	200
E3: Structure and order	A	200
	B	200
	C	200
Total		1,800

Table 1: Statistics of the BL2MP test set according to levels (A, B, C) and grammatical error types. Each type contains 600 sentences at three different levels, for a total of 1,800 sentences.

E1: Declension

In Basque linguistic morphology, a distinctive feature presents itself through declension, a grammatical phenomenon involving the attachment of suffixes to all phrases in a sentence. These suffixes serve as substitutes for prepositions, with each one corresponding to a specific grammatical case. These cases are defined by interrogative pronouns that establish relationships such as 'who', 'to whom', 'in whom', 'with whom', 'by whom', 'for whom', 'where', 'from whom', among others. Importantly, each case includes three unique forms, varying based on whether it denotes a singular, plural, or indefinite entity.

E2: Verbs

The verb-related errors in the dataset span five main dimensions:

Person: Person in verb conjugation refers to the grammatical category that indicates the relationship between the subject and the verb, who is performing the action.

Number: Number in verb conjugation refers to whether the subject is singular or plural.

Tense: Tenses in Basque indicate different moments in time when the action of the verb occurs (present tense, past tense, imaginary tense).

Mood: Basque verbs can also convey the speaker's attitude or intention towards the action. The indicative mood expresses facts or statements in a neutral way. The subjunctive mood is used to express desires, purposes, or intentions.

Aspect: The aspect of a verb indicates whether the action is completed or ongoing. Basque has three main aspects: completed, incomplete, and event.

E3: Structure and Order

The third category of errors pertains to the structure and order of sentences in the Basque language. Within this category, the errors are divided in two major classes.

The first concerns the grammatically correct utilization of sentence order. It is important to note that, like most agglutinative languages (Mahowald et al., 2023a), Basque exhibits considerable flexibility in the ordering of sentence elements (Laka, 1996). However, this flexibility is notably constrained within certain syntactic constructs, such as noun phrases (Laka, 1996).

The second aspect of structural errors relates to the sequencing of elements in compound sentences, which includes a variety of structures such as completive sentences, indirect interrogative sen-

[2] verb number prediction and suffix recovery

[3] https://www.baiby.com/en/

[4] https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions

tences, relative sentences, and others.

4. Methodology

4.1. Pre-training Corpora

Zhang et al. (2021) demonstrate that a pre-training corpus ranging from 10M to 100M words is enough for a language model to acquire most of the linguistic capacities of syntax and semantics in English. Thus, we selected three Basque corpora within this range. We employed the pre-training datasets defined by Urbizu et al. (2023), comprising 5M, 25M and 125M words, which keeps a constant increase rate among them. Regarding the domain of the texts, the corpora are a mix of 75% news and 25% text from Wikipedia.

4.2. Model Architecture

We employ three sizes of BERT models: BERT_{12L}, BERT_{8L} and BERT_{4L}, with 12, 8 and 4 hidden layers[5] respectively.

We use a cased sub-word vocabulary comprising 50K tokens trained[6] using the unigram-based subword segmentation algorithm proposed by Kudo (2018). The resulting models have 124M, 51M, and 16M parameters, respectively. Models were trained using default hyperparameters[7], applying whole-word masking at Masked Language Model (MLM) training (with a dup-factor[8] of 10), with a batch of 256 and a sequence length of 512, for around 500K steps on a v3-8 TPU.

We trained each model up to the 2^n epochs closer to 500K steps. BERT models trained with the 125M words dataset were trained for 640,000 steps (512 epochs), the models for the 25M dataset for 512,000 steps (2048 epochs) and the models for the 5M dataset for 409,600 (8192 epochs). In every case, the warm-up was set at 6.25% of the training in alignment with the recommendations by Izsak et al. (2021), translating to 40,000 steps (32 epochs) for the 125M, 32,000 steps (128 epochs) for the 25M, and 25,600 steps (512 epochs) for the 5M word corpus.

4.3. Evaluation Method

To evaluate sentences using the BERT architecture, we chose to score entire sentences following the methodology proposed by Salazar et al. (2020). This approach allows for the assessment of the linguistic competence of NLMs in a supervision-free manner, and it does not handicap models trained with predicting unmasked tokens (e.g. BERT, RoBERTa) as other alternative methods like holistic scoring (Zaczynska et al., 2020) does.

The scoring method introduced by Salazar et al. (2020) computes the *Pseudo-log-likelihood* (PLL) scores from NLMs. These scores are derived by summing the conditional log probabilities $logP_{MLM}(\boldsymbol{w}_t \mid \boldsymbol{W}_{\setminus t})$ of each token in a sentence. This process involves masking each token w_t in turn with the [MASK] token in BERT and then computing the probability of the original token given the rest of the sentence context $W_{\setminus t}$.

PLL score for sentence $\boldsymbol{W}$ would be:

$$\text{PLL}(\boldsymbol{W}) := \sum_{t=1}^{|\boldsymbol{W}|} \log P_{\text{MLM}}(\boldsymbol{w}_t \mid \boldsymbol{W}_{\setminus t}; \Theta).$$

Where Θ stands for the parameters of the model.

We used the toolkit *minicons*[9] Misra (2022) to compute the the PLL scores of the sentences. We calculate the PLL scores for both grammatically incorrect and grammatically correct sentences in the test set. For each minimal pair, we compared the two obtained PLL scores, considering it a correct identification if the grammatically correct sentence received a higher score than the incorrect one. Subsequently, we calculated the accuracy based on the proportion of minimal pairs correctly identified.

We discarded example pairs that have an unequal post-tokenization length as done in Zaczynska et al. (2020) for German, to ensure that results are not affected by differences in length measured in number of tokens.

5. Experiments

5.1. How do Corpus Size, Model Size and Epochs Affect Learning Grammar?

In this first experiment, we compared various models that resulted from different configurations used during the pre-training of BERT models. These configurations varied in terms of corpus size and the number of parameters in the model. The objective of this experiment was to understand how the performance in grammar learning changes in relation to the size of the training corpus and to determine whether an increase in the model's parameters enhances grammar learning.

Additionally, we examined when models begin to learn grammar by evaluating various checkpoints and observing whether training over multiple epochs benefits grammar learning.

[5] shrinking other parameters in proportion (hidden dimension, number of attention heads, etc.)

[6] trained for each pre-training corpora

[7] lr = $1e^{-4}$, $\beta 1$ = 0.9, $\beta 2$ = 0.999, L2 w.d.=0.01

[8] Number of times to duplicate the input data (with different masks).

[9] https://github.com/kanishkamisra/minicons

Error-type	Unacceptable Example	Acceptable Example
E1: Declension	"Nik oso pozik nago."	"Ni oso pozik nago." (*I am very pleased.*)
E2: Verb	"Nik daukat zure autoaren giltzak."	"Nik dauzkat zure autoaren giltzak." (*I have your car keys.*)
E3: Structure and Order	"Balkoitik oso ederra bista daukat."	"Balkoitik oso bista ederra daukat." (*I have a wonderful view from the balcony.*)

Table 2: Examples of grammatically unacceptable and acceptable instances for each type of error, with the English translations of the acceptable examples provided in parentheses. Errors and corrections are underlined.

To achieve this objective, we trained a BERT model of each size described in Section 4.2 (12, 8, and 4 layers) using each of the pre-training corpora (5M, 25M, and 125M words) mentioned in Section 4.1. Figure 1 presents the accuracies obtained on the BL2MP test set across different pre-training checkpoints for the nine models. Tables containing results are available in Appendix A.

In our experiments, approximately one-third of the examples were discarded during evaluation because the sentence pairs had varying lengths after tokenization.

The charts show how increasing the pre-training corpora size significantly improves the performance of the models on the BL2MP test set. On the contrary, improvement derived from increasing model size is more limited. There is a considerable gain by increasing model size from BERT_{4L} to BERT_{8L}. However, further increasing the model size to BERT_{8L} yields minimal additional gains.

For each model, we observe that accuracy initially starts around 50% (equivalent to random guessing), decreases during the first epochs, and then begins to increase as the model starts learning towards the end of the warm-up phase, as indicated by the pre-training and development loss curves. This improvement continues until it reaches the optimal level for each model, occurring at different epochs. Notably, the larger the dataset, the sooner this improvement in BL2MP performance is observed within the epochs. It is not only beneficial but also necessary to train for multiple epochs to learn Basque grammar, especially in the context of the low-resource pre-training corpus scenarios examined here. During the pre-training of some models, signs of overfitting are evident, particularly in models trained on the 5M dataset and, to a lesser extent, in the largest model trained on the 25M dataset, as seen in the development loss curve. However, this overfitting does not appear to hinder grammar learning.

As previously noted, BERT_{8L} significantly outperforms BERT_{4L} across all corpus sizes. However, increasing the model size from BERT_{8L} to BERT_{12L} does not yield substantial accuracy improvements. Therefore, to minimize the number of experiments, we chose to proceed with a single model size, selecting BERT_{8L} for its lower computational resource requirements and relatively minor performance difference compared to BERT_{12L}.

There are test sets similar to BLIMP for various languages, as outlined in subsection 2.3. Nonetheless, each test set is developed using distinct methodologies and focuses on particular linguistic phenomena. Consequently, the difficulty levels across these test sets are not comparable, which prevents us from comparing the results obtained on them with each other and with our test.

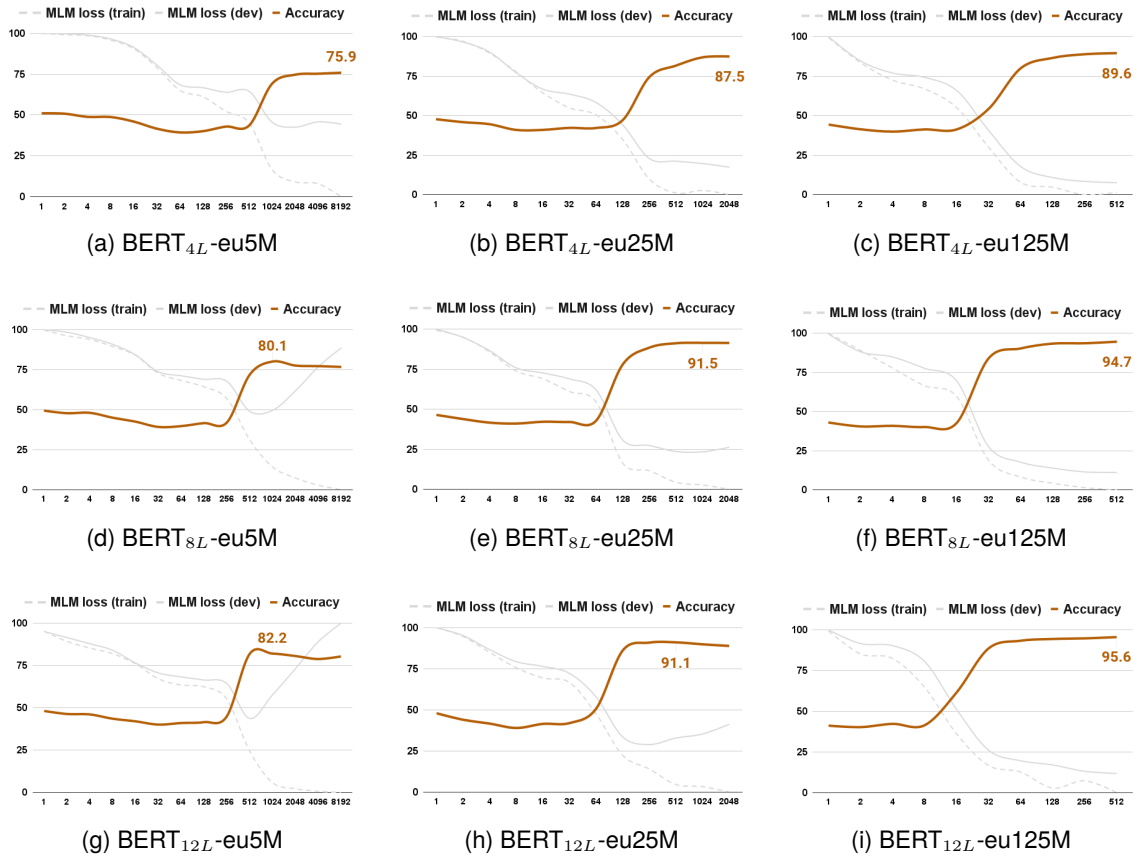
5.2. Are some Grammatical Phenomena Harder to Learn for NLMs?

In Section 5.1, we observe that BERT models are capable of acquiring grammatical knowledge, which improves with the size of the pre-training corpora. However, is the learning of certain grammatical phenomena more challenging than others?

To address this, we analyzed the performance of the BERT_{8L} model, specifically focusing on the different types of grammatical errors from the BL2MP dataset, which reflect the involved grammatical phenomena. These error types are: Declension (E1), Verb (E2), and Structure and Order (E3). The results for each of these error types are presented in charts a, b, and c of Figure 4.

The charts show that the three grammatical error types are learned around the same epochs for each model, but they do not reach always the same accuracies. The BERT_{8L} trained with the 5M word corpus, performs better at structure and order (E3) than at declension (E1) and verbs (E2), reaching 83.3 at E3 at the 1024th epoch, while getting 79.0 and 77.4 accuracy at E1 and E2 respectively.

The charts indicate that the three types of grammatical errors are learned by each model around the same epochs, yet they do not consistently achieve the same accuracy levels. Specifically, the BERT_{8L} model trained on the 5M word corpus provides better performance in Structure and Order (E3) compared to Declension (E1) and Verbs (E2).



(a) BERT$_{4L}$-eu5M (b) BERT$_{4L}$-eu25M (c) BERT$_{4L}$-eu125M

(d) BERT$_{8L}$-eu5M (e) BERT$_{8L}$-eu25M (f) BERT$_{8L}$-eu125M

(g) BERT$_{12L}$-eu5M (h) BERT$_{12L}$-eu25M (i) BERT$_{12L}$-eu125M

Figure 1: Accuracies at BL2MP for the BERT$_{12L}$, BERT$_{8L}$ and BERT$_{4L}$ models pre-trained with corpora of 5M, 25M and 125M each, during training epochs (exponential scale). Each model is trained for around 500K steps (see Section 4.2). Grey lines represent the normalized (Min-max) loss during pre-training on the training and development sets.

It reaches an 83.3 accuracy in E3 by the 1024th epoch, whereas it gets 79.0 and 77.4 accuracy in E1 and E2, respectively.

5.3. Do the Struggles of L2 Students Correlate with BERT?

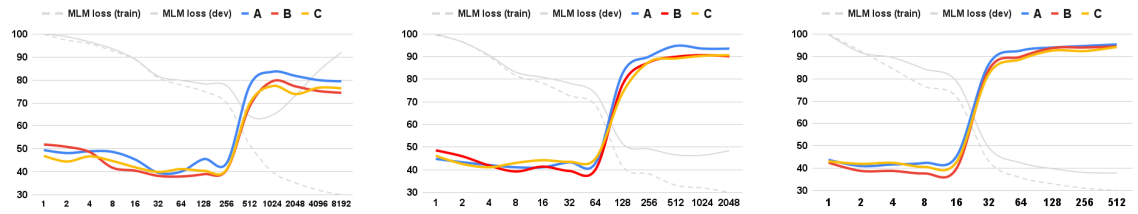
In this experiment, we aim to determine whether there's a correlation between student proficiency levels and the difficulty NLMs face in learning grammatical errors associated with those levels. To achieve this, in Figure 2, we present the accuracies of BERT models with 8 layers (BERT$_{8L}$) trained on corpora of 5 million, 25 million, and 125 million words. These accuracies are categorized by the L2 proficiency levels of the students from whom the BL2MP examples were collected.

The charts indicate that NLMs learn grammatical phenomena from students of all proficiency levels simultaneously, achieving similar learning milestones around the same epoch for each pre-training dataset size. However, the NLM's performance on Set A, which includes grammatical errors made by lower-level L2 students, surpasses its performance on Sets B and C. This discrepancy diminishes when the pre-training corpus expands to 125 million words, at which point the model shows comparable proficiency in handling minimal pairs from all difficulty levels. Despite the expectation that Set C, representing more advanced student errors, would be more challenging than Set B, the BERT model with 8 layers (BERT$_{8L}$) demonstrates similar performance on both sets across the various sizes of pre-training corpora.

5.4. Does Flexible Word Order Hinder the Task of Learning Grammar?

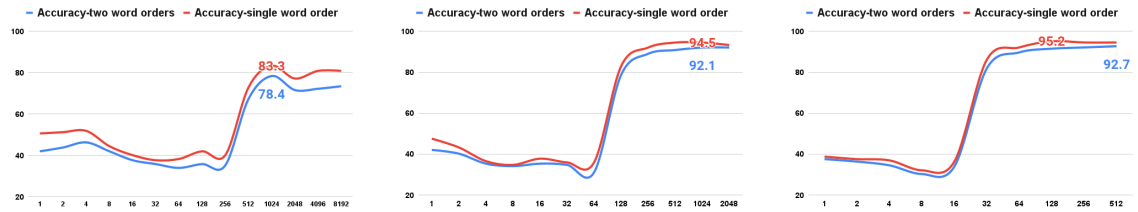
In this experiment, we aim to determine whether the flexible word order hinders the learning process of Basque grammar during the pre-training phase. Specifically, we want to check if the NLMs accurately learn the grammatical phenomena presented in minimal pairs, regardless of the variations in sentence word order.

To analyze this issue, we extract a subset of 200 minimal pairs from the BL2MP test set, which we refer as 200_BL2MP_1. This test set contains only



(a) BERT$_{8L}$-eu5M (b) BERT$_{8L}$-eu25M (c) BERT$_{8L}$-eu125M

Figure 2: Accuracies on BL2MP obtained by BERT$_{8L}$ trained with 5M, 25M and 125M word corpora, separated by the L2 student proficiency level of the example. Grey lines represent the normalized (Min-max) loss during pre-training on the training and development sets.



(a) BERT$_{8L}$-eu5M (b) BERT$_{8L}$-eu25M (c) BERT$_{8L}$-eu125M

Figure 3: Accuracies obtained by BERT$_{8L}$-eu trained with 5M/25M/125M corpora for 200 pairs including single variation of word order (Accuracy-single word order) for 200 pairs including two variations of word order (Accuracy-two word orders).

one word-order variant for each minimal pair. From 200_BL2MP_1, we construct a second test set by adding a new grammatically equivalent pair with a different word order for each existing pair. For instance, from the pair (*Katuak sagua jan zuen.* | *Katuak sagua jan zen.*), we generated the new pair (*Sagua katuak jan zuen.* | *Sagua katuak jan zen.*). We refer to this second test set as 200_BL2MP_2, which includes two word order variants for each pair: the original (e.g., *Katuak sagua jan zuen.* | *Katuak sagua jan zen.*) and the new one created manually (e.g., *Sagua katuak jan zuen.* | *Sagua katuak jan zen.*).

Subsequent evaluations are conducted on the 200_BL2MP_1 and 200_BL2MP_2 test sets. In the evaluation of 200_BL2MP_1, accuracy is calculated based on a single word order variant per pair (Accuracy-single word order in Figure 3). In contrast, the evaluation on 200_BL2MP_2 considers both word order variants for accuracy calculation (Accuracy-two word orders in Figure 3). This means that for a minimal pair to be counted as correct, the NLM must accurately solve both word order variations. For more details, see Appendix B, which contains Tables with numerical results.

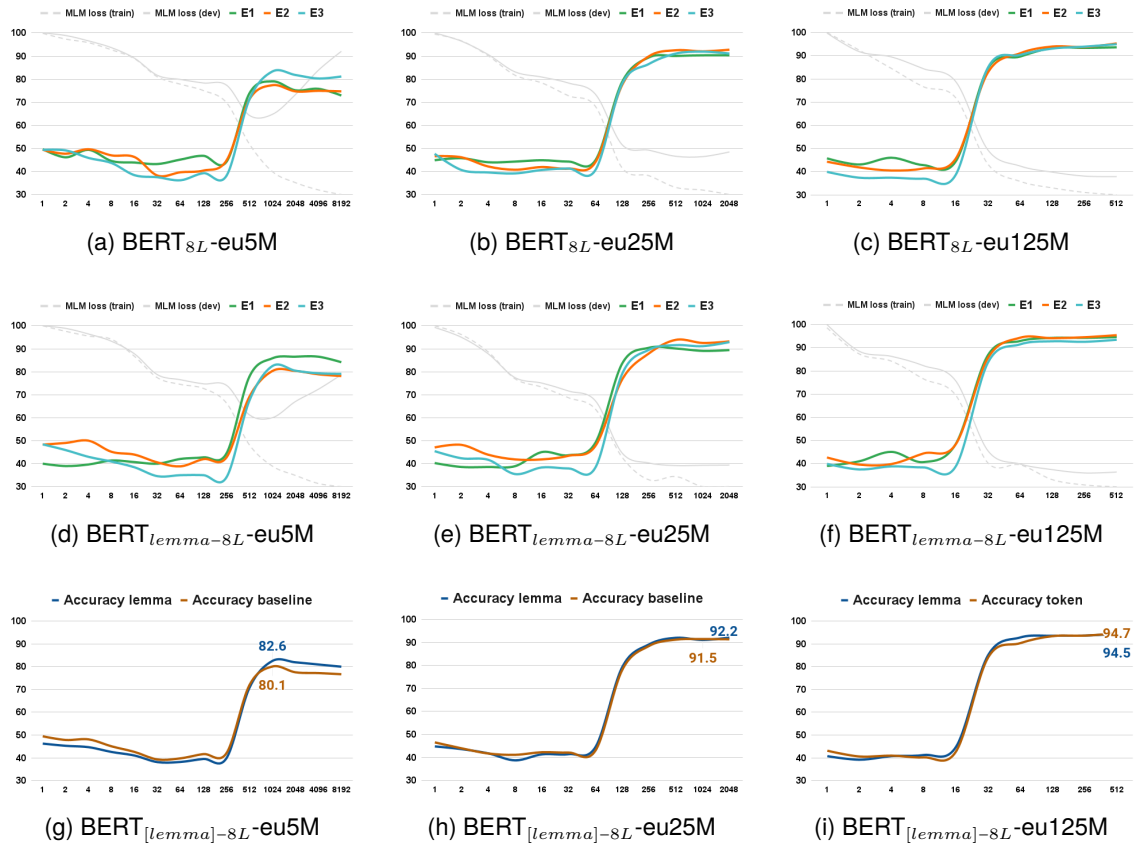
The graphs indicate that BERT$8L$-eu25M and BERT$8L$-eu125M (as shown in charts b and c of the Figure 3) are capable of identifying the same grammatical error across different sentence rearrangements with minimal loss (approximately 2.5 points) compared to identifying the error in a single word order. This loss is more pronounced for BERT$_{8L}$-eu5M, which exhibits a gap of 5.1 points (as shown in charts a of the Figure 3), suggesting that flexible word order poses a greater challenge to the learning of Basque grammar, particularly when pre-training data is insufficient.

5.5. Does Lemmatization Help to Learn Grammar in an Agglutinative Language?

While simply employing a standard sub-word vocabulary usually is good enough for languages without a complex morphology, machine translation models and NLMs for agglutinative languages can benefit from lemmatization (Pan et al., 2020; Mohseni and Tebbifakhr, 2019; Nzeyimana and Rubungo, 2022), which helps NLMs learn better the grammatical regularities and patterns, especially in low resource settings (Pan et al., 2020).

In light of this, we aimed to investigate this approach with the Basque language to determine whether lemmatization enhances the grammar learning process of BERT models.

To achieve that objective, we preprocess the corpora using Eustagger, a morphological analyzer and Part-of-Speech tagger for Basque (Ezeiza et al., 1998), to segment each word into its lemma and declension suffix (refer to Table 3 for an exam-



(a) BERT$_{8L}$-eu5M (b) BERT$_{8L}$-eu25M (c) BERT$_{8L}$-eu125M

(d) BERT$_{lemma-8L}$-eu5M (e) BERT$_{lemma-8L}$-eu25M (f) BERT$_{lemma-8L}$-eu125M

(g) BERT$_{[lemma]-8L}$-eu5M (h) BERT$_{[lemma]-8L}$-eu25M (i) BERT$_{[lemma]-8L}$-eu125M

Figure 4: Accurary results by grammatical phenomena on BL2MP, for the BERT$_{8L}$ models trained with 5M, 25M and 125M word corpora, and their lemmatized counterparts. E1: Declension, E2: Verb, E3: Structure and order. Charts g, h and i contain the total accuracies of the a-d, b-e and c-f chart pairs. Grey lines represent the normalized (Min-max) loss during pre-training on the training and development sets.

ple). Through this segmentation, focused specifically on nouns and adjectives, we aim to reduce the morphological complexity of Basque to levels more comparable to those of languages without complex declension, such as English or Spanish.

The morphology of suffixes corresponding to the same grammatical declension case is not uniformly regular (e.g., 'Ate**an** dago -> ate NUMS_MUGM_INE dago', 'Aldapa**n** dago -> Aldapa NUMS_MUGM_INE dago'). To standardize this morphology, we adopt the most frequently occurring suffix in the corpus as the representative suffix for each morphosyntactic category. In the case of NUMS_MUGM_INE (inessive singular), for example, the suffix 'an' is the most common in the corpus. Consequently, the segmentation for the aforementioned examples would be: 'Ate**an** dago -> ate **an** dago' and 'Aldapa**n** dago -> Aldapa **an** dago'.

In this experiment, we evaluated three BERT$_{8L}$ models trained on 5M, 25M, and 125M word corpora, respectively, comparing their performance on both the original and lemmatized versions of the corpora. The results are depicted in Figure 2 and

Source	Etxe**ko** ate**an** dago[a]
Lem. and morph.	Etxe NUMS_MUGM_GEL ate NUMS_MUGM_INE dago
Segmentation	Etxe ko ate an dago

[a] [It] is at the door of the house

Table 3: Example of morphological segmentation applied prior to tokenizing the pre-training corpora. Morphosyntactic tags are replaced by corresponding representative suffixes ('ko' for 'NUMS_MUGM_GEL' and 'an' for 'NUMS_MUGM_INE'). 'NUMS_MUGM_GEL' denotes the genitive singular of place, and 'NUMS_MUGM_INE' the inessive singular.

Appendix C.

The charts reveal that BERT models trained on lemmatized texts slightly outperform their counterparts trained on the original texts (charts g, h&i), with more than a 2-point improvement observed in the model trained on the 5M corpus. However, the advantage of lemmatization diminishes with larger corpus sizes. For the 25M corpus, lemmati-

Model	Corpus	Acc.
BERT$_{8L}$-eu25M	25M	91.5
BERT$_{8L}$-eu125M	125M	94.7
BERT$_{12L}$-eu125M	125M	95.6
BERTeus	225M	94.8
Roberta-euscrawl-B	289M	95.9
Roberta-euscrawl-L	289M	95.5
ElhBERTeu-medium	350M	95.4
ElhBERTeu-base	350M	**96.5**
mBERT	~30M	58.9
XLM-RoBERTa base	unk.	75.1
XLM-RoBERTa large	unk.	82.4
IXAmBERT	225M	94.4

Table 4: Accuracy results at BL2MP for monolingual and multilingual MLM models for Basque.

zation enhances accuracy by only half a point, and it does not provide any advantage over the original text model for the 125M corpus, which achieves an accuracy of 94.7%.

When we look in depth at the results of the models trained on 5M words by error type (as illustrated in charts a&d), it's observed that the base model excels in handling E3 (Structure and Order) errors more effectively than E1 (Declension) and E2 (Verb) errors. However, with lemmatization applied, while performance on E3 errors remains consistent, there's a slight increase in accuracy for E2 errors, narrowing the gap with E3. Moreover, a significant enhancement is noted in E1 errors, with lemmatization leading to superior performance compared to both E2 and E3 errors.

As the size of the pre-training corpus increases to 25M (charts b&e), the distinctions among error types and the enhancements from lemmatization lessen, and they disappear entirely for the 125M corpus (charts c&f).

5.6. Comparison with Published BERT Models

Lastly, we evaluate most of the publicly available monolingual and multilingual NLM models for Basque on the BL2MP evaluation dataset we created in this work. Table 4 includes the accuracies of the three best performing models of this work, monolingual NLM models for Basque such as BERTeus (Agerri et al., 2020), ElhBERTeu-base|medium (Urbizu et al., 2022), Roberta-euscrawl-base|large (Artetxe et al., 2022), and multilingual NLM models as mBERT (Devlin et al., 2019), XLM-RoBERTa base|large (Conneau et al., 2019) and IXAmBERT (Otegi et al., 2020).

As Table 4 shows, every monolingual model (except BERT$_{8L}$-eu25M) obtains similar results, with an accuracy of around 95%. BERT$_{8L}$-eu125M and BERT$_{12L}$-eu125 trained in this work are competitive with the rest of the State of the Art MLM models for Basque, although they have been trained with a smaller corpus of 125M words. ElhBERTeu-base is the best performing model among all of them, which is the one trained with the biggest corpus. Among the multilingual models, we see how IXAmBERT, which includes 3 languages, outperforms the other massively multilingual models.

6. Conclusions

In summary, after pre-training various BERT models with differing configurations and evaluating them with the BL2MP dataset, we conclude that the training corpus size significantly impacts grammar learning more than the model size. Our findings also underscore the necessity of multi-epoch training for effective grammar acquisition across the 5M-125M word range of pre-training corpora investigated in this study.

Our experiments reveal that lemmatization-based tokenization improves grammar learning in inflectional languages like Basque under very limited resource conditions (5M pre-training words). However, as the dataset size increases to 25M and 125M words, the advantage of lemmatization diminishes. A similar pattern is observed with word order: flexible word order poses challenges in grammar learning with smaller corpora (5M words), but with a sufficient number of training examples, NLMs can learn grammar phenomena regardless of sentence order.

Moreover, while the performance gap among different grammatical phenomena in BERT models is relatively narrow, the model trained on 5M words shows a particular strength in structure and order (E3) over declension (E1) and verbs (E2). In contrast, the model trained on the lemmatized 5M word corpus excels in declension (E1) over verbs (E2) and structure and order (E3). The only notable correlation between L2 student proficiency levels and NLMs' learning challenges is a slight advantage in handling examples from the lowest proficiency level (A), which evens out as the training corpus size increases to 125M words.

Lastly, we believe that the BL2MP dataset, meticulously curated with minimal pairs for evaluating NLMs on Basque grammar, will become a crucial asset for future research aimed at probing linguistic understanding and grammatical precision in neural language models.

Limitations

The experiments have been performed only on BERT models, which implies that the conclusions obtained from the experiments are not fully transferable to other NLMs, be the autoencoder models

(such as RoBERTa or ALBERT), or autoregressive NLMs such as GPT. To determine in a robust way to what extent the conclusions of the analysis presented in this paper are extrapolable to other types of NLMs, it would be necessary to repeat some experiments with these types of models.

The analysis presented in this paper focuses on evaluating the knowledge that a pre-trained BERT model acquires about the grammar of Basque. The evaluation followed is intrinsic, so it does not allow us to determine how the degree of knowledge about the grammar estimated from the pre-trained model is transferred to the performance that such a model would offer in specific downstream tasks. In the fine-tuning process for learning downstream tasks, hyperparameters such as the number of parameters or the number of epochs used in the pre-training also play an important role. It would be necessary to evaluate the different model configurations in different downstream tasks, in order to determine the relationship between learned grammatical knowledge and task performance, also taking into account the different hyperparameters analysed in this work.

The set of minimal pairs for assessing the grammatical knowledge that the pre-trained model acquires has been made on the basis of examples of grammatical errors made by adult learners of a second language (Basque). Most of these learners are native speakers of Spanish. These facts determine to some extent the types of grammatical phenomena collected in the test used in this paper.

Acknowledgements

This work has been partially funded by the Basque Government (ICL4LANG project, grant no. KK-2023/00094). We also acknowledge the support of Google's TFRC program. Finally, we want to thank bai&by language school for providing the annotated essays written by Basque language learners.

7. Bibliographical References

Judit Acs, Endre Hamerlik, Roy Schwartz, Noah A Smith, and Andras Kornai. Morphosyntactic probing of multilingual bert models. *Natural Language Engineering*, pages 1–40.

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. 2020. Give your text representation models some love: the case for basque. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4781–4788.

Wafa Abdullah Alrajhi, Hend Al-Khalifa, and Abdulmalik AlSalman. 2022. Assessing the linguistic knowledge in arabic pre-trained language models using minimal pairs. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 185–193.

Mikel Artetxe, Itziar Aldabe, Rodrigo Agerri, Olatz Perez-De-Viñaspre, and Aitor Soroa. 2022. Does corpus quality really matter for low-resource languages? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7383–7390.

Omer Veysel Cagatan. 2023. Toddlerberta: Exploiting babyberta for grammar learning and language understanding. *arXiv preprint arXiv:2308.16336*.

Ethan A Chi, John Hewitt, and Christopher D Manning. 2020. Finding universal grammatical relations in multilingual bert. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Rudolf PG De Rijk. 1969. Is basque an sov language? *Fontes linguae vasconum: Studia et documenta*, 1(3):319–352.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Mikel L Forcada and Francis Tyers. 2016. Apertium: a free/open source platform for machine translation and basic language technology. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products*.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Philip A Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. Babyberta: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th conference on computational natural language learning*, pages 624–646.

Peter Izsak, Moshe Berchansky, and Omer Levy. 2021. How to train bert with an academic budget.

In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10644–10652.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657.

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.

Itziar Laka. 1996. *A brief grammar of Euskara, the Basque language*. Universidad del País Vasco, Euskal Herriko Unibertsitatea, Euskarazko ….

Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William Chandra Tjhi. 2023. Bhasa: A holistic southeast asian linguistic and cultural evaluation suite for large language models. *arXiv preprint arXiv:2309.06085*.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7:195–212.

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. *arXiv preprint arXiv:1907.11692*.

Zeyu Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A Smith. 2021. Probing across time: What does roberta know and when? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 820–842.

K Mahowald, E Diachek, E Gibson, E Fedorenko, and R Futrell. 2023a. Grammatical cues to subjecthood are redundant in a majority of simple clauses across languages. *Cognition*, 241:105543–105543.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2023b. Dissociating language and thought in large language models: a cognitive perspective. *arXiv preprint arXiv:2301.06627*.

Vladislav Mikhailov, Tatiana Shamardina, Max Ryabinin, Alena Pestova, Ivan Smurov, and Ekaterina Artemova. 2022. Rucola: Russian corpus of linguistic acceptability. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5207–5227.

Mahdi Mohseni and Amirhossein Tebbifakhr. 2019. Morphobert: A persian ner system with bert and morphological analysis. In *Proceedings of The First International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2019) co-located with ICNLSP 2019-Short Papers*, pages 23–30.

Dan Nielsen. 2023. Scandeval: A benchmark for scandinavian natural language processing. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 185–201.

Jingcheng Niu, Wenjie Lu, and Gerald Penn. 2022. Does bert rediscover a classical nlp pipeline? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3143–3153.

Antoine Nzeyimana and Andre Niyongabo Rubungo. 2022. Kinyabert: a morphology-aware kinyarwanda language model. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5347–5363.

Arantxa Otegi, Aitor Agirre, Jon Ander Campos, Aitor Soroa, and Eneko Agirre. 2020. Conversational question answering in low resource scenarios: A dataset and case study for basque. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 436–442.

Yirong Pan, Xiao Li, Yating Yang, and Rui Dong. 2020. Morphological word segmentation on agglutinative languages for neural machine translation. *arXiv preprint arXiv:2001.01589*.

Isabel Papadimitriou, Richard Futrell, and Kyle Mahowald. 2022. When classifying grammatical role, bert doesn't care about word order... except when it matters. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 636–643.

Shauli Ravfogel, Yoav Goldberg, and Francis Tyers. 2018. Can lstm learn to capture agreement? the case of basque. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 98–107.

Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913.

Taiga Someya and Yohei Oseki. 2023. Jblimp: Japanese benchmark of linguistic minimal pairs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1536–1549.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. 2022. Sling: Sino linguistic evaluation of large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4606–4634.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, Rodrigo Agerri, and Aitor Soroa. 2022. Basqueglue: A natural language understanding benchmark for basque. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1603–1612.

Jannis Vamvas and Rico Sennrich. 2021. On the limits of minimal pairs in contrastive evaluation. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 58–68.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Alex Warstadt, Leshem Choshen, Aaron Mueller, Adina Williams, Ethan Wilcox, and Chengxu Zhuang. 2023. Call for papers – the babylm challenge: Sample-efficient pretraining on a developmentally plausible corpus. *Computing Research Repository*, arXiv:2301.11796.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020a. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel Bowman. 2020b. Learning which features matter: Roberta acquires a preference for linguistic generalizations (eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 217–235.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. Climp: A benchmark for chinese language model evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790.

Karolina Zaczynska, Nils Feldhus, Robert Schwarzenberg, Aleksandra Gabryszak, and Sebastian Möller. 2020. Evaluating german transformer language models with syntactic agreement tests. *arXiv preprint arXiv:2007.03765*.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel Bowman. 2021. When do you need billions of words of pretraining data? 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)*, pages 1112–1125.

8. Language Resource References

Ezeiza, Nerea and Alegria, Iñaki and Arriola, Jose Mari and Urizar, Rubén and Aduriz, Itziar. 1998. *Combining stochastic and rule-based methods for disambiguation in agglutinative languages*.

Kanishka Misra. 2022. *minicons: Enabling Flexible Behavioral and Representational Analyses of Transformer Language Models*.

Urbizu, Gorka and San Vicente, Iñaki and Saralegi, Xabier and Agerri, Rodrigo and Soroa, Aitor. 2023. *Scaling Laws for BERT in Low-Resource Settings*.

A. Results Tables for Section 5.1

In this Appendix, we report the detailed results from which we build the charts from Figure 1, on the Tables 5, 6, 7, 8, 9, 10, 11, 12 and 13.

epoch	steps	train loss	dev loss	acc
1	50	11.5	11.5	51.1
2	100	11.5	11.5	50.8
4	200	11.4	11.4	48.9
8	400	11.1	11.2	48.8
16	800	10.6	10.7	46.0
32	1600	9.5	9.7	41.6
64	3200	8.1	8.5	39.3
128	6400	7.8	8.3	40.1
256	12800	6.9	8.0	42.9
512	25600	6.2	8.1	43.5
1024	51200	3.4	6.3	69.2
2048	102400	2.7	6.0	74.7
4096	204800	2.6	6.3	75.3
8192	409600	1.8	6.1	75.9

Table 5: Results of BERT$_{4L}$-eu5M on BL2MP.

epoch	steps	train loss	dev loss	acc
1	250	11.4	11.5	47.8
2	500	11.2	11.2	46.0
4	1000	10.6	10.6	44.7
8	2000	9.6	9.5	41.0
16	4000	8.4	8.6	41.0
32	8000	7.6	8.4	42.3
64	16000	7.2	7.9	42.2
128	32000	5.9	6.7	47.3
256	64000	3.8	4.9	74.6
512	128000	3.0	4.7	81.7
1024	256000	3.2	4.6	87.0
2048	512000	2.9	4.4	87.5

Table 6: Results of BERT$_{4L}$-eu25M on BL2MP.

epoch	steps	train loss	dev loss	acc
1	1250	10.5	10.6	44.4
2	2500	9.3	9.4	41.4
4	5000	8.5	8.8	40.0
8	10000	8.0	8.6	41.3
16	20000	7.1	8.0	41.3
32	40000	5.2	6.0	54.4
64	80000	3.5	4.3	80.1
128	160000	3.3	3.8	86.6
256	320000	2.9	3.6	88.9
512	640000	3.1	3.5	89.6

Table 7: Results of BERT$_{4L}$-eu125M on BL2MP.

epoch	steps	train loss	dev loss	acc
1	50	11.5	11.6	49.4
2	100	11.1	11.4	47.8
4	200	10.9	11.0	48.1
8	400	10.4	10.5	45.0
16	800	9.7	9.8	42.6
32	1600	8.5	8.6	39.3
64	3200	8.0	8.3	39.8
128	6400	7.5	8.1	41.6
256	12800	6.7	7.9	42.1
512	25600	3.8	5.8	72.0
1024	51200	1.9	5.9	80.1
2048	102400	1.1	7.2	77.6
4096	204800	0.6	8.9	77.2
8192	409600	0.3	10.3	76.7

Table 8: Results of BERT$_{8L}$-eu5M on BL2MP.

epoch	steps	train loss	dev loss	acc
1	250	10.8	10.9	46.5
2	500	10.4	10.4	43.9
4	1000	9.6	9.6	41.7
8	2000	8.5	8.7	41.1
16	4000	8.0	8.4	42.2
32	8000	7.3	8.0	42.1
64	16000	6.7	7.4	43.0
128	32000	3.2	4.5	78.1
256	64000	2.8	4.2	88.5
512	128000	2.1	3.9	91.3
1024	256000	2.0	3.9	91.5
2048	512000	1.7	4.1	91.4

Table 9: Results of BERT$_{8L}$-eu25M on BL2MP.

B. Results Tables for Section 5.4

In this Appendix, we report the results from which we build the charts from Figure 3, on the Tables 14, 15 and 16.

C. Result Tables for Section 5.5

In this Appendix, we report the results from which we build the charts *g*, *h* and *i* from Figure 4, on the Tables 17, 18 and 19.

epoch	steps	train loss	dev loss	acc
1	1250	9.7	9.6	43.1
2	2500	8.8	8.8	40.6
4	5000	8.0	8.5	41.0
8	10000	7.1	8.0	40.2
16	20000	6.6	7.4	42.7
32	40000	3.5	4.2	84.1
64	80000	2.7	3.4	90.4
128	160000	2.4	3.2	93.5
256	320000	2.2	3.0	93.7
512	640000	2.1	3.0	94.7

Table 10: Results of BERT$_{8L}$-eu125M on BL2MP.

epoch	steps	train loss	dev loss	acc
1	50	11.3	11.3	48.2
2	100	10.6	10.9	46.4
4	200	10.1	10.4	46.1
8	400	9.7	10.0	43.6
16	800	9.0	9.1	42.1
32	1600	8.0	8.4	40.1
64	3200	7.6	8.1	41.0
128	6400	7.4	7.9	41.5
256	12800	6.6	7.7	44.9
512	25600	2.9	5.2	81.9
1024	51200	0.7	6.8	82.2
2048	102400	0.3	8.7	80.7
4096	204800	0.1	10.6	78.9
8192	409600	0.0	11.9	80.5

Table 11: Results of BERT$_{12L}$-eu5M on BL2MP.

epoch	steps	train loss	dev loss	acc
1	250	10.4	10.4	48.0
2	500	9.9	10.0	44.0
4	1000	9.0	9.2	41.7
8	2000	8.1	8.5	39.0
16	4000	7.5	8.2	41.5
32	8000	7.3	7.8	42.0
64	16000	5.5	6.4	50.8
128	32000	3.1	4.1	86.2
256	64000	2.3	3.7	91.0
512	128000	1.4	4.1	91.1
1,024	256000	1.3	4.3	89.9
2,048	512000	1.0	4.9	88.9

Table 12: Results of BERT$_{12L}$-eu25M on BL2MP.

epoch	steps	train loss	dev loss	acc
1	1250	9.2	9.3	41.2
2	2500	8.1	8.6	40.3
4	5000	7.9	8.5	42.3
8	10000	6.6	7.8	41.4
16	20000	4.4	5.6	61.6
32	40000	3.0	3.7	88.7
64	80000	2.7	3.2	93.4
128	160000	1.9	3.0	94.5
256	320000	2.2	2.7	94.8
512	640000	1.7	2.6	95.6

Table 13: Results of BERT$_{12L}$-eu125M on BL2MP.

epoch	steps	acc (1)	acc (2)
1	50	50.6	42.0
2	100	51.2	43.8
4	200	51.9	46.3
8	400	44.4	42.0
16	800	40.1	37.7
32	1600	37.7	35.8
64	3200	38.3	34.0
128	6400	42.0	35.8
256	12800	40.1	35.2
512	25600	72.8	67.3
1024	51200	83.3	78.4
2048	102400	77.2	71.6
4096	204800	80.9	72.2
8192	409600	80.9	73.5

Table 14: BERT$_{8L}$-eu5M results on 200_BL2MP_1 and 200_BL2MP_2 test sets. acc (1) refers to Accuracy-single word order and acc (2) refers to Accuracy-two word order.

epoch	steps	acc (1)	acc (2)
1	250	47.6	42.1
2	500	43.3	40.2
4	1000	36.6	35.4
8	2000	34.8	34.1
16	4000	37.8	35.4
32	8000	36.0	34.8
64	16000	36.0	31.1
128	32000	82.9	78.7
256	64000	92.1	89.0
512	128000	94.5	90.9
1024	256000	94.5	92.1
2048	512000	93.3	92.1

Table 15: BERT$_{8L}$-eu25M results on 200_BL2MP_1 and 200_BL2MP_2 test sets. acc (1) refers to Accuracy-single word order and acc (2) refers to Accuracy-two word order.

epoch	steps	acc (1)	acc (2)
1	1250	38.8	37.6
2	2500	37.6	36.4
4	5000	37.0	34.5
8	10000	32.1	30.3
16	20000	36.4	33.9
32	40000	86.1	81.8
64	80000	92.1	89.7
128	160000	95.2	91.5
256	320000	94.5	92.1
512	640000	94.5	92.7

Table 16: BERT$_{8L}$-eu125M results on 200_BL2MP_1 and 200_BL2MP_2 test sets. acc (1) refers to Accuracy-single word order and acc (2) refers to Accuracy-two word order.

epoch	steps	acc baseline	acc lemma
1	50	49.4	46.2
2	100	47.8	45.3
4	200	48.1	44.7
8	400	45.0	42.6
16	800	42.6	41.0
32	1600	39.3	38.1
64	3200	39.8	38.2
128	6400	41.6	39.5
256	12800	42.1	39.8
512	25600	72.0	70.8
1024	51200	80.1	82.6
2048	102400	77.6	82.0
4096	204800	77.2	81.0
8192	409600	76.7	80.0

Table 17: Accuracies of BERT$_{8L}$-eu5M and its lemmatized counterpart on BL2MP.

epoch	steps	acc baseline	acc lemma
1	250	46.5	44.8
2	500	43.9	43.6
4	1000	41.7	41.7
8	2000	41.1	38.7
16	4000	42.2	41.3
32	8000	42.1	41.4
64	16000	43.0	44.5
128	32000	78.1	79.3
256	64000	88.5	89.2
512	128000	91.3	92.1
1024	256000	91.5	91.1
2048	512000	91.4	92.2

Table 18: Accuracies of BERT$_{8L}$-eu25M and its lemmatized counterpart on BL2MP.

epoch	steps	acc baseline	acc lemma
1	1250	43.1	40.7
2	2500	40.6	39.2
4	5000	41.0	40.7
8	10000	40.2	41.2
16	20000	42.7	44.8
32	40000	84.1	85.1
64	80000	90.4	92.9
128	160000	93.5	93.6
256	320000	93.7	93.7
512	640000	94.7	94.5

Table 19: Accuracies of BERT$_{8L}$-eu125M and its lemmatized counterpart on BL2MP.