



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

**Gorka Urbizu** ([g.urbizu@orai.eus](mailto:g.urbizu@orai.eus))  
Muitze Zulaika, Xabier Saralegi, Ander Corral

# INDEX

INTRODUCTION  
BL2MP DATASET  
METHODOLOGY  
EXPERIMENTS  
CONCLUSIONS

# Introduction

INTRODUCTION

# Motivation

**LMs** excel at linguistic skills, including **grammar**...

...at least in **English**...

**But** what happens with **other languages** with complex morphology or flexible word order?

**Research Question:**

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

Mendian = In the mountain

beroenetatik = from the hottest

*the cat ate the mouse*

katuak jan zuen sagua

katuak sagua jan zuen

sagua katuak jan zuen

sagua jan zuen katuak

jan zuen katuak sagua

jan zuen sagua katuak

INTRODUCTION

# Main Contributions

**Analysis of grammatical knowledge of BERTs** trained under various configurations:

- Training corpus size.
- Model size.
- Epochs.

**BL2MP** dataset: Basque L2 student-based Minimal Pairs.

We also study **linguistic factors**:

- Morphology and tokenization.
- Word order.
- Grammatical phenomena.
- L2 student proficiency.

**Minimal pair:**

John **loves** chocolate | John **love** chocolate  
Nik **ekarri** dut | **Ni** ekarri dut



# BL2MP Dataset

## BL2MP DATASET

# BL2MP (Basque L2 student-based Minimal Pairs)

- Built following the design of **English BLiMP** dataset (Warstadt et al., 2020).
- Based on **essays** written by **L2 students** learning Basque.
  - Grammar mistakes → their corrections
- We create minimal pairs (hand checked):
  - 3 error types:
    - Declension.
    - Verb.
    - Structure and order.
  - 3 proficiency levels:
    - A: Beginner.
    - B: Intermediate.
    - C: Advanced.
  - Balanced dataset.

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

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



# Methodology

METHODOLOGY

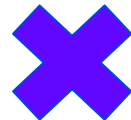
# Training Models

Several BERT models:

- 50K token vocab (unigram).
- Batch size: 256.
- Seq len: 512.
- Warm-up: 6.25%.
- Trained on a v3-8 TPU.

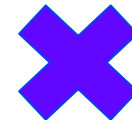
Corpora:

- 5M words.
- 25M words.
- 125M words.



Models:

- BERT<sub>12L</sub>: 124M params.
- BERT<sub>8L</sub>: 51M params.
- BERT<sub>4L</sub>: 16M params.



Epochs:

- ~500K steps.
- 512/2048/8192 epochs.
- Exponential ckpts: 2<sup>n</sup> epochs.

## METHODOLOGY

# Evaluation Method

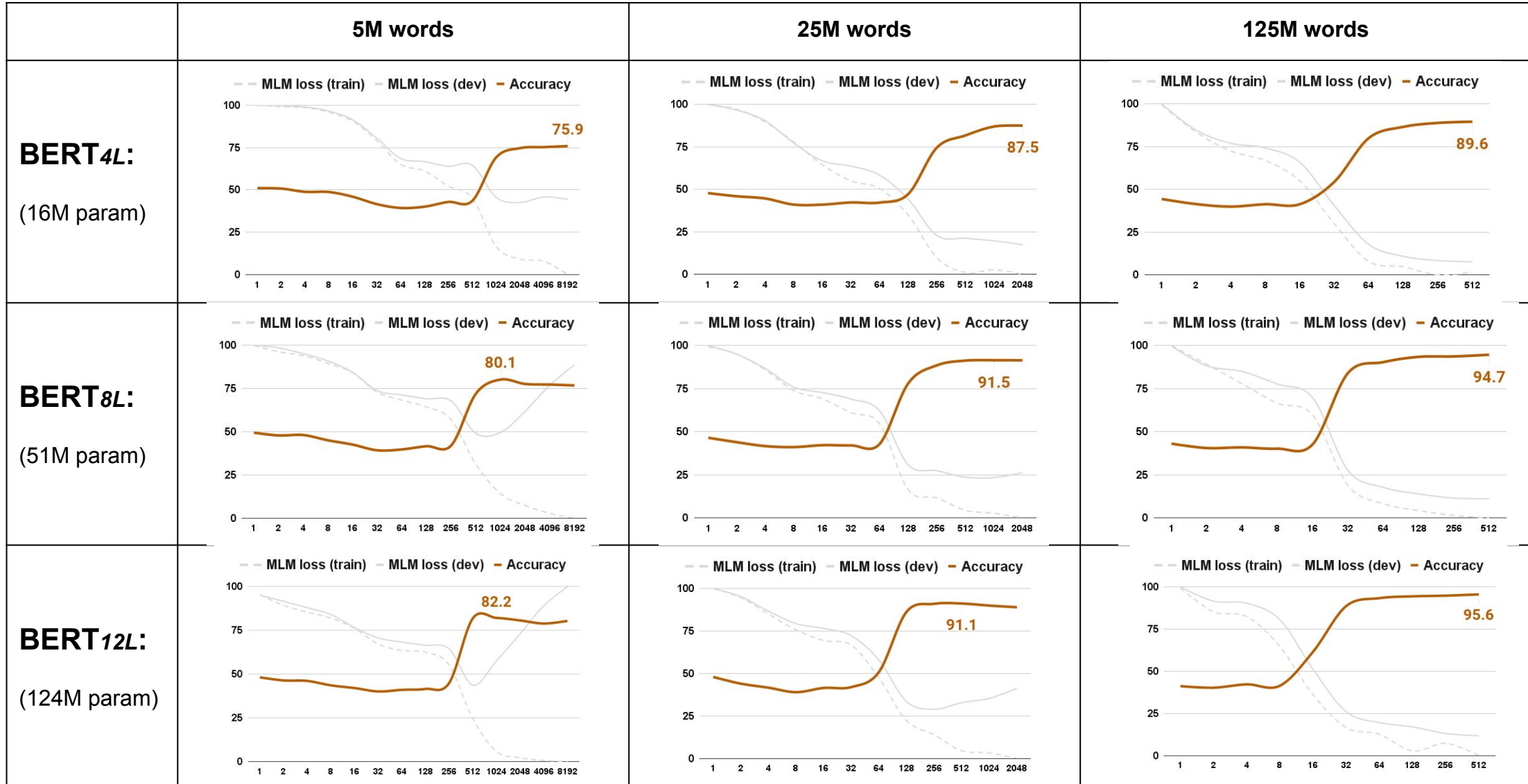
- We score sentences using *Pseudo-log-likelihood (PPL)* (Salazar et al., 2020).

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

- We calculate the PLL scores for both sentences in a minimal pair.
- We calculate the proportion of minimal pairs correctly identified:
  - $\text{PPL}(\text{correct sentence}) > \text{PPL}(\text{incorrect sentence})$ .

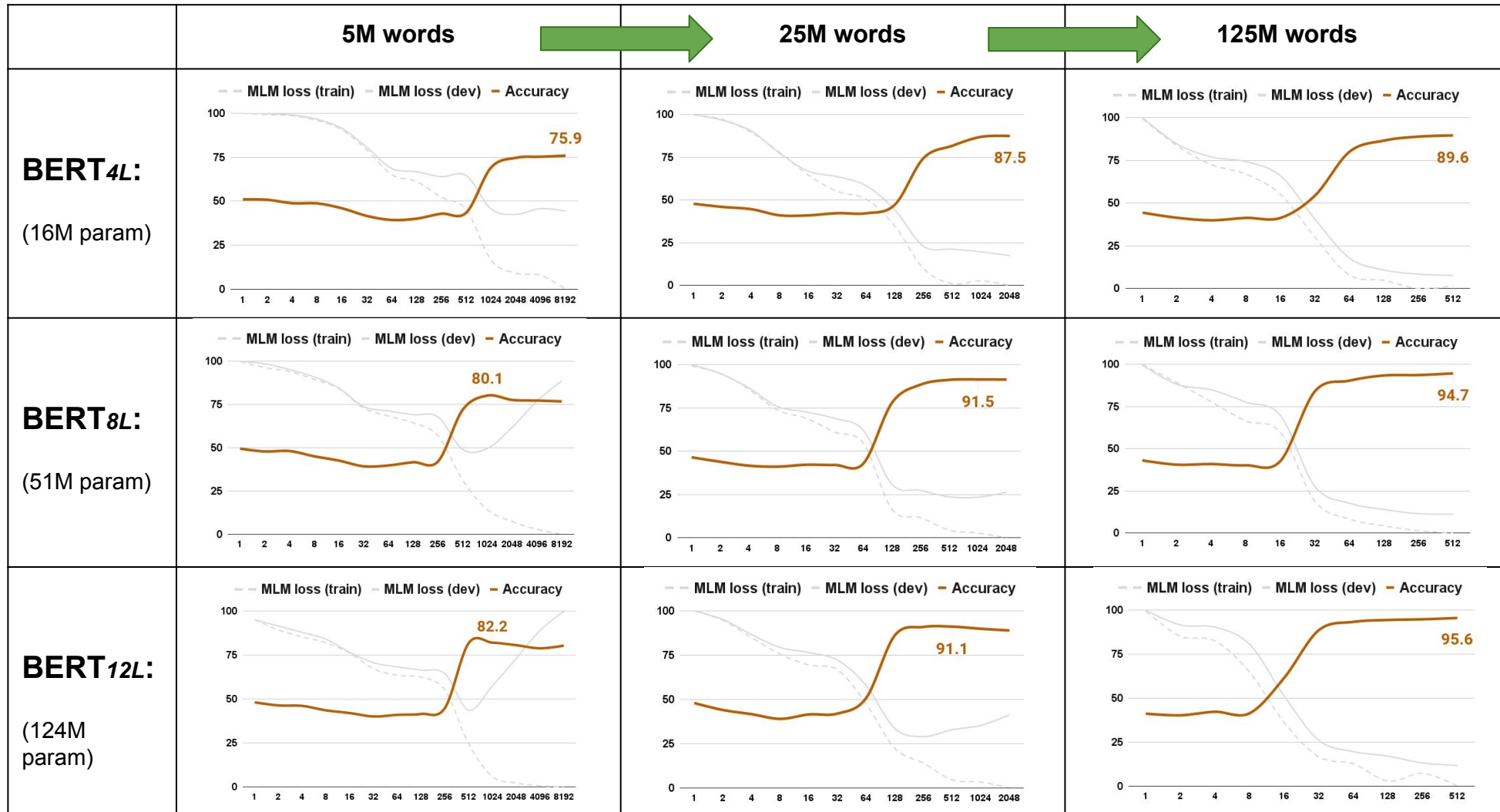
# Experiments

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



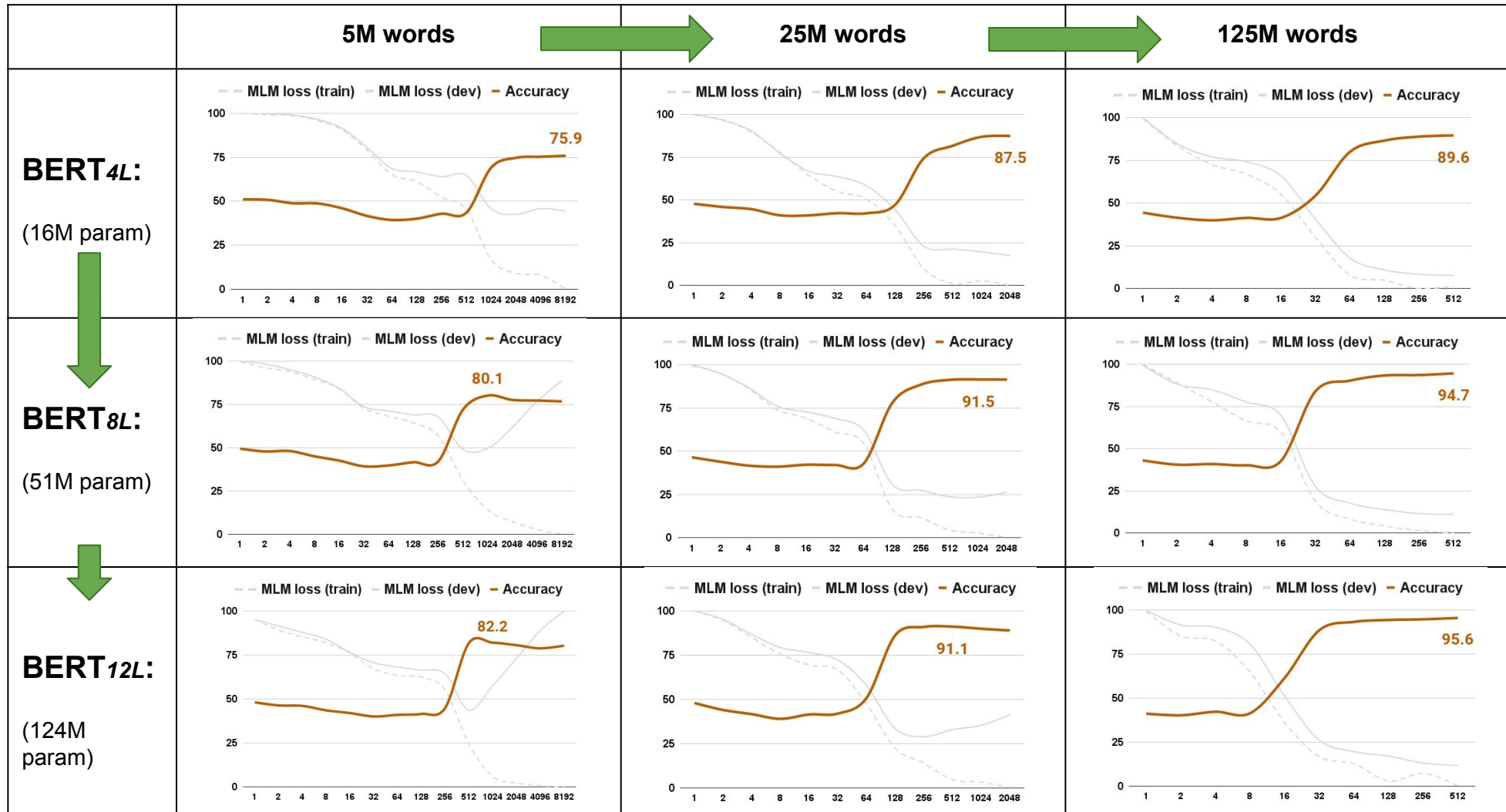
# How do Corpus Size, Model Size and Epochs Affect Learning Grammar?

Increasing the pre-training **corpora size** improves the performance of the models.



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

Improvement from increasing **model size** too, but it is more limited from BERT<sub>8L</sub> → BERT<sub>12L</sub>.

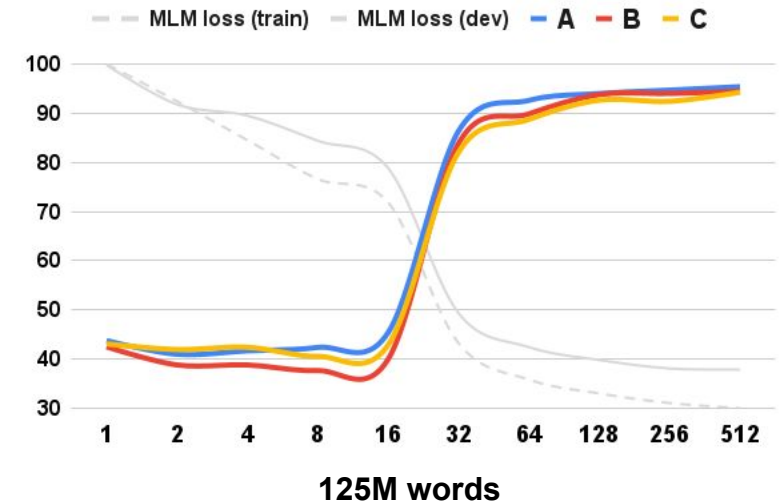
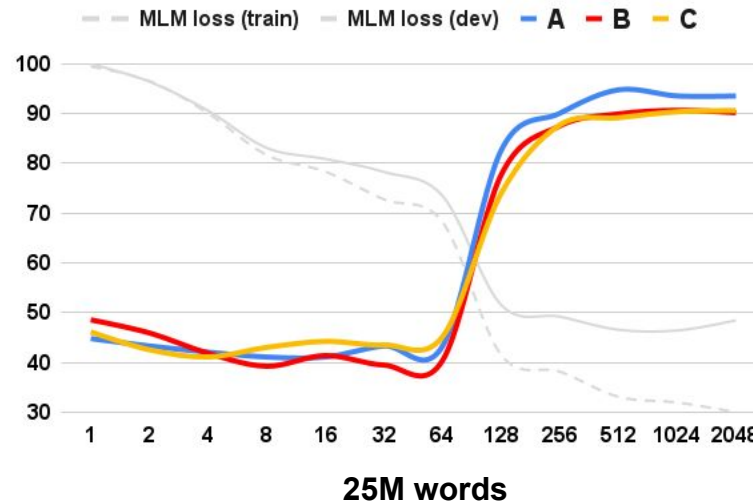
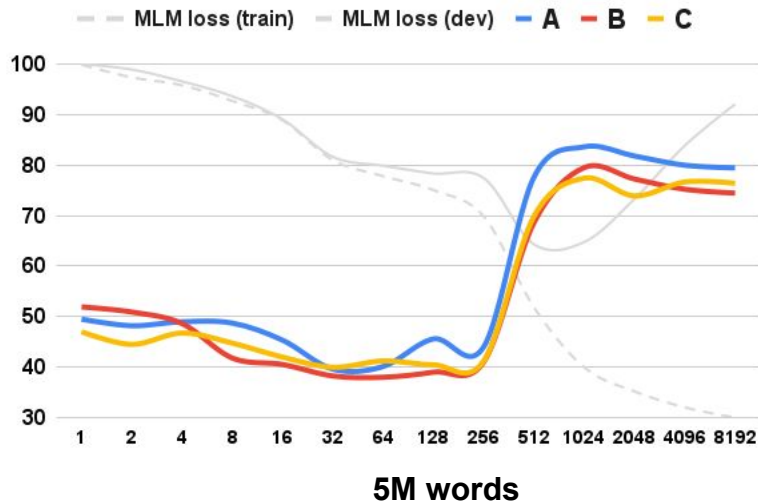


## EXPERIMENTS

# Do the Struggles of L2 Students Correlate with BERT?

We divided the results for medium sized BERTs by different students' proficiency.

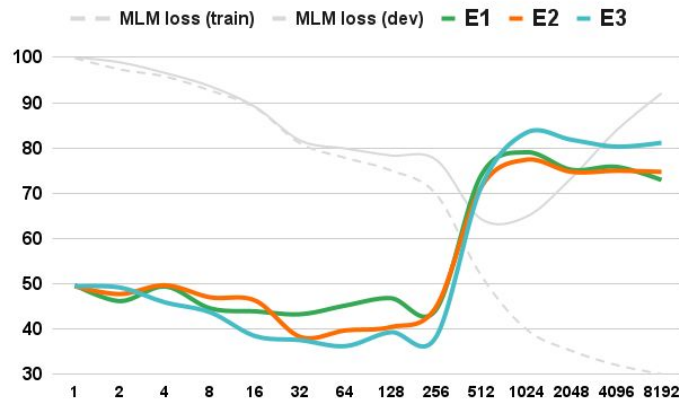
- Performance on Set A (Beginner), higher than on Sets B (Intermediate) and C (Advanced).
  - The gap diminishes when the pre-training corpus expands to 125 million words.



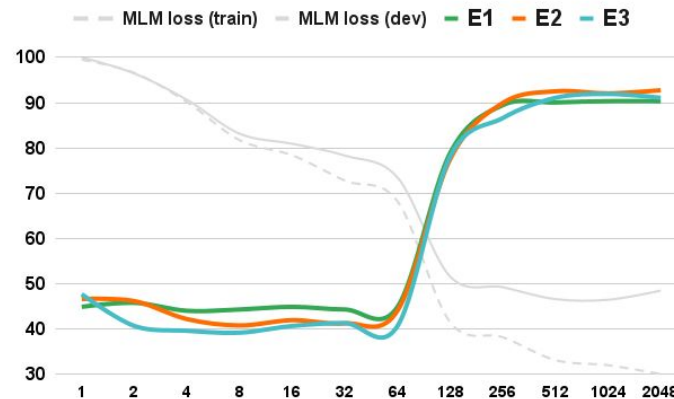
## EXPERIMENTS

# Are some Grammatical Phenomena Harder to Learn?

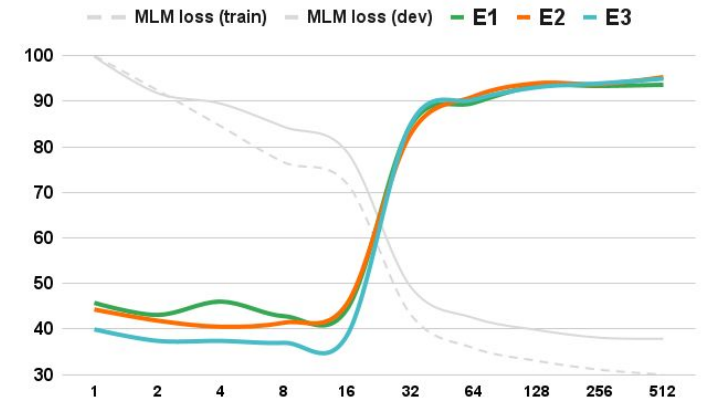
- We analyzed performance on types of grammatical errors:
  - Declension (E1).
  - Verb (E2).
  - Structure and Order (E3).
- All grammatical errors are learned around the same epochs.
- When trained on 5M words it performs better at Structure and Order (E3).
- The difference diminishes with more data.



5M words



25M words



125M words

EXPERIMENTS

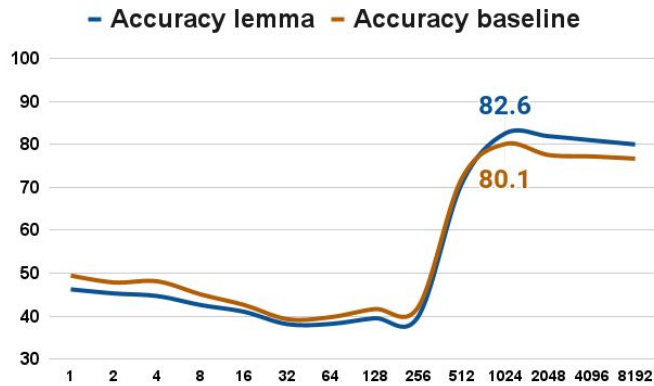
# Does Lemmatization Help in an Agglutinative Language?

- We segmented each word into its lemma and declension suffix using Eustagger.
- Lemmatization benefits BERT<sub>8L</sub> trained on 5M by 2.5 points.
  - But the benefits of lemmatization are diluted when trained on more data.

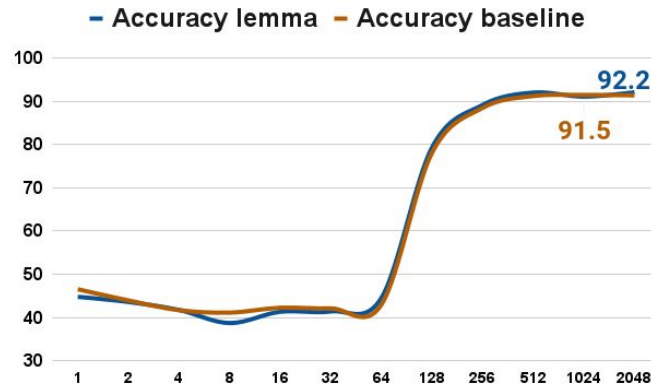


Source	Etxeko atean dago <sup>a</sup>
Lem. and morph.	Etxe NUMS_MUGM_GEL ate NUMS_MUGM_INE dago
Segmentation	Etxe ko ate an dago

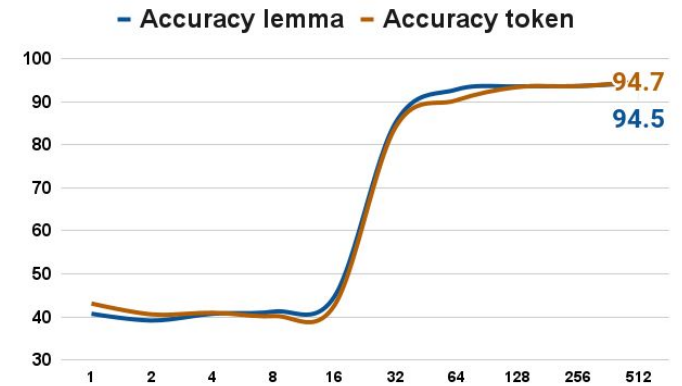
<sup>a</sup> [It] is at the door of the house



5M words



25M words

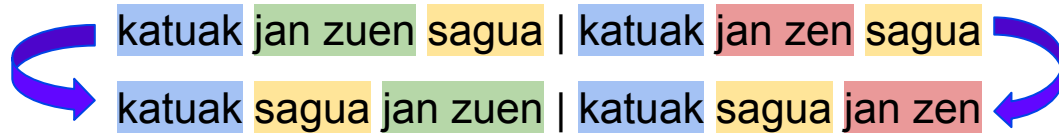


125M words

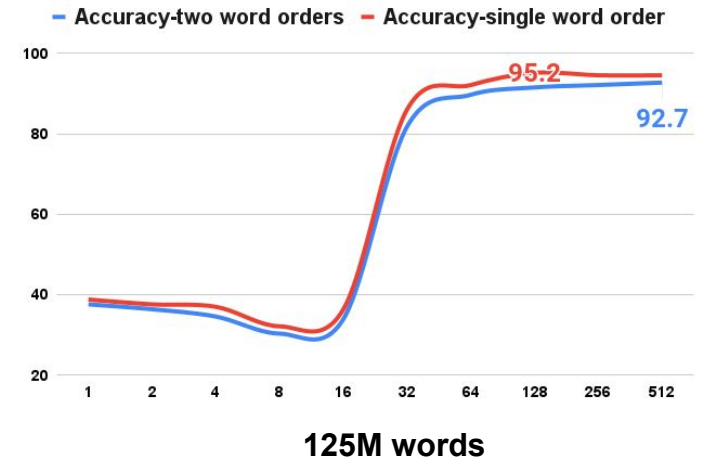
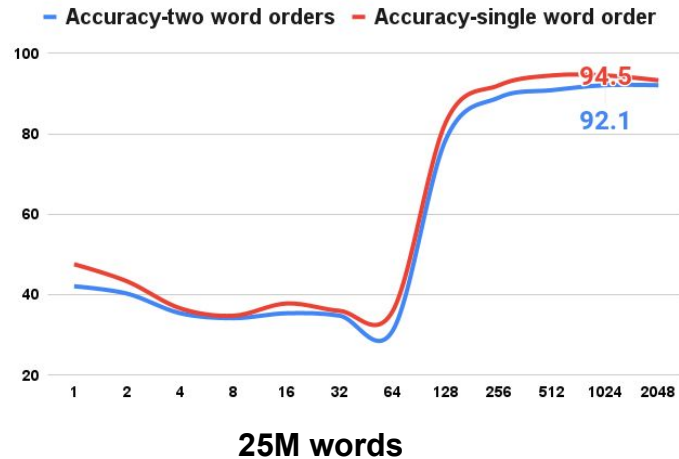
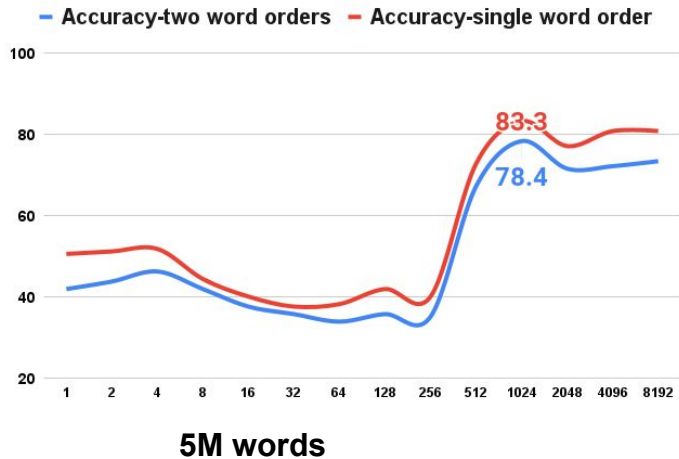
EXPERIMENTS

# Does Flexible Word Order Hinder the Learning Grammar?

- Do LMs accurately learn the grammatical phenomena presented in minimal pairs, regardless of the variations in sentence word order?
- We rewrite a subset of minimal pairs in alternative word orders:



- Flexible word order is a challenge to learn grammar, especially in low data scenarios.

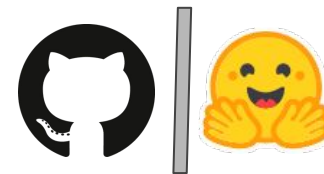


# Conclusions

CONCLUSIONS

# Conclusions

- Increasing corpus size and model size are beneficial for learning Basque grammar.
- More beneficial increasing corpus size than model size.
- Multi-epoch training needed.
- Lemmatization improves grammar learning, but only for very small corpus (5M).
- Flexible-word-order is a challenge in grammar learning, especially when trained on very small corpora.
- Similar performance across different grammatical phenomena and proficiency level.



[orai-nlp/bl2mp](#)

# References

# References

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S. F., & Bowman, S. (2020). BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8, 377-392.

Salazar, J., Liang, D., Nguyen, T. Q., & Kirchoff, K. (2020, July). Masked Language Model Scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 2699-2712).



**Thank you!**

@: [g.urbizu@orai.eus](mailto:g.urbizu@orai.eus)

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

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