Tomato, Tomahto, Tomate: Do Multilingual Language Models Understand Based on Subword-Level Semantic Concepts?

Crystina Zhang[1]*, **Jing Lu**[2], **Vinh Q. Tran**[2], **Tal Schuster**[2], **Donald Metzler**[2], **Jimmy Lin**[1]
[1]University of Waterloo [2]Google DeepMind

Abstract

Human understanding of text depends on general semantic concepts of words rather than their superficial forms. To what extent does our human intuition transfer to language models? In this work, we study the degree to which current multilingual language models (mLMs) understand based on subword-level semantic concepts. To this end, we form "semantic tokens" by merging the semantically similar subwords and their embeddings, and evaluate the updated mLMs on five heterogeneous multilingual downstream tasks. Results show that the general shared semantics could get the models a long way in making the predictions on mLMs with different tokenizers and model sizes. Inspections of the grouped subwords show that they exhibit a wide range of semantic similarities, including synonyms and translations across many languages and scripts. Lastly, we find that the zero-shot results with semantic tokens are on par with or even better than the original models on certain classification tasks, suggesting that the shared subword-level semantics may serve as the anchors for cross-lingual transfer.

1 Introduction

Human understanding of text depends on general semantic concepts of words that are robust to their superficial forms (Figure 1). The most obvious examples are semantically equivalent words in different languages: code-switching "they" to "они", or "tomatoes" to "tomate".[1] The sentence meaning is preserved for people who understand both languages. The robustness in understanding may also apply to some of the inflectional changes: swapping "tomatoes" for "tomato", while losing the plural form, still conveys the overall sentence meaning. Finally, even the words are replaced with ones

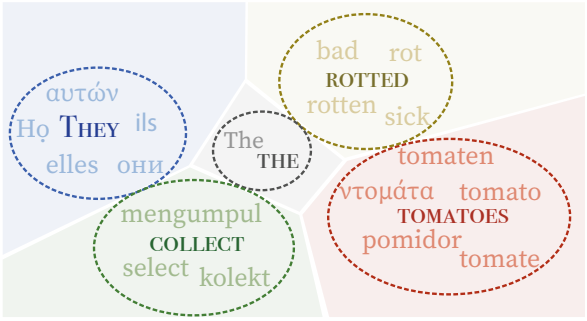

Figure 1: Words with similar meanings or inflectional changes fall under semantic concepts, as indicated by the colors. The sentence "They collected the rotted tomatoes." is from XNLI (Conneau et al., 2018).

that are only vaguely related in semantics, e.g., from "rotted" to "bad", resulting in a phrase that is less precise, the sentence meaning could still be interpreted and in line with the original sentence. While the definition and representation of concept have not been universally agreed upon (Jackendoff, 1988; Gabora et al., 2008; Goddard and Wierzbicka, 2013; Gardenfors, 2014; Fumagalli et al., 2019; Sajjad et al., 2022), and their alignment varies across languages depending on domain and cultural context (Thompson et al., 2020), it is generally accepted that concepts underlying words could be shared or even universal across languages (Bundy and Wallen, 1984; Goddard and Wierzbicka, 2013) and have a pivotal role in many aspects of language understanding (Murphy, 2004; Fumagalli et al., 2019).

On the other hand, language models learn distinct embedding vectors for these subwords, which share similar underlying semantics yet may not share similar context. We thus ask: *To what degree do current multilingual language models (mLMs) understand based on subword-level semantic concepts?* Based on the existing mLMs, we form "semantic tokens" by grouping the subwords based on the similarity of their word embeddings, which

*Work is done while Crystina Zhang was a student researcher at Google DeepMind.

[1] "они": "they" in Russian; "tomate": "tomato" in Spanish

Findings of the Association for Computational Linguistics: NAACL 2025, pages 1821–1837
April 29 - May 4, 2025

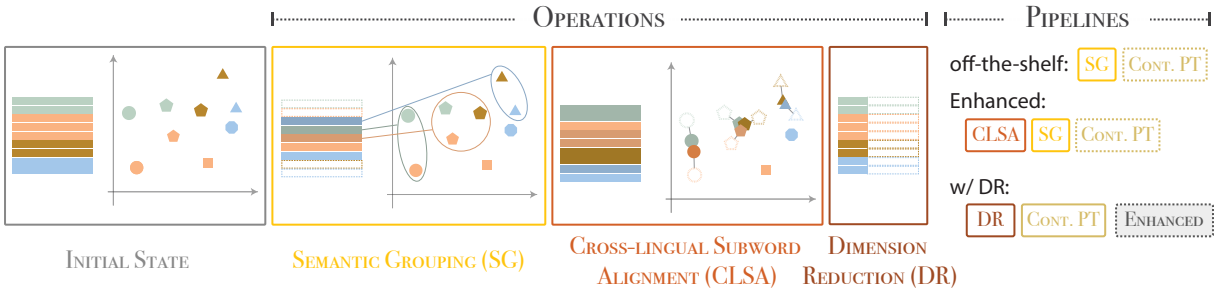

Figure 2: Illustrations of the operations that modify the word embeddings and the pipelines composed of these operations. "Cont. Pt": Continual Pretraining, whose illustration is skipped due to its prevalence. The colored rows represent the embeddings of the subwords and the coordinates depict their spatial distances. The shapes indicate the underlying semantics and the colors indicate the languages of the subwords. Under the "Pipelines", the solid boxes denote the required operations that are core to the pipeline or essential to the model functionality; the dashed boxes denote the optional operations that are only to provide additional improvement. Better viewed in color.

henceforth share the same "semantic embedding".[2] The updated mLMs are evaluated on five downstream tasks, which cover thirty languages in total and include classification and embedding tasks in different granularities.

We find that with a small number of semantic tokens and their embeddings, the mLMs can preserve most of the downstream effectiveness: semantic tokens in 5% of the original vocabulary size achieve 90% of the effectiveness on classification tasks, and 20% of the semantic tokens achieve over 85% effectiveness on the embedding tasks. These findings suggest that while nuances exist in the meaning of each subword, the general semantics representations could get the models a long way in prediction-making.

Next, we eliminate the confounding factor of embedding size: While forming semantic tokens, the number of the word embedding parameters is also reduced. Does the change in effectiveness stem from changes in subwords or reduced embedding parameters? We thus apply the semantic grouping to the word embeddings with reduced parameter size via truncating the embedding dimension, finding that the same observation persists even when the parameters could not be further reduced via dimension reduction. This suggests that the above results are not confounded by the embedding size but rather are a result of the semantic grouping.

[2] We understand that the formed semantic tokens are not perfect and may contain subwords of loosely related or unrelated meanings. On one hand, we provide inspection of the formed semantic tokens in Figure 8 and 13, which suggest that the grouped tokens could reflect coherent semantics. On the other hand, we consider our results as a lower bound, where techniques that form more accurate semantic tokens are likely to further improve the downstream effectiveness.

Additional experiments suggest that the findings generalize to mLMs with different tokenizers and model sizes. Inspection shows that the grouped subwords indeed exhibit a wide range of semantic similarities: numbers, punctuation, synonyms, and translations across multiple languages under different scripts. Lastly, we find that the zero-shot results on certain classification tasks with semantic tokens are on par with or even better than the original models, suggesting that the shared subword-level semantics may serve as the transfer anchors for cross-lingual generalization.

Our contributions are as follows: (1) We find that mLMs can preserve a majority of the downstream effectiveness with a small number of shared subword-level semantics (Section 4.1, 4.2). (2) We show that the findings are general across mLMs with different tokenizers, model sizes, and other aspects (Section 4.3). (3) Inspection reveals that the grouped subwords exhibit a wide range of semantic similarities (Section 4.4). (4) The zero-shot results suggest that the shared subword-level semantics may serve as transfer anchors for cross-lingual generalization (Section 4.5).

2 Operations and Pipelines

The experimental setting in this paper can be categorized into several pipelines composed of multiple independent operations, centering on semantic grouping. All operations, except for the continual pretraining, only affect the vocabulary V and word embeddings $E \in R^{(|V|,D)}$, where $|V|$ is the vocabulary size and D is the initial word embeddings dimension. Figure 2 illustrates all operations except the continual pretraining due to its prevalence, as well as the pipelines composed of the operations.

Dataset Name	Task Name	Task Type	Granularity	# L.	Languages
MasakhaNER	NER	classification	word-level	10	am, ha, ig, rw, lg, sw, wo, yo, luo, pcm
XNLI	NLI	classification	sentence-level	15	ar, bg, de, el, en, es, fr, hi, ru, sw, th, tr, ur, vi, zh
TyDi QA	QA	classification	sentence-level	11	ar bn, en, fi, d, ja, ko, ru, sw, te, th
MIRACL	P Retrieval P Reranking	classification embedding	passage-level	18	ar, bn, de, en, es, fa, fi, fr, hi, id, ja, ko, ru, sw, te, th, yo, zh

Table 1: Downstream Tasks and Datasets. Granularity suggests the information level required from input data to perform the task. "P Retrieval": Passage Retrieval; "P Reranking": Passage Reranking.

2.1 Semantic Grouping (SG)

Given the vocabulary V and its word embeddings E, multiple semantically similar subwords are grouped into a single "semantic token" and henceforth share the same "semantic embedding". The semantic embedding is then initialized by the averaged embeddings of the grouped words. That is, after the grouping, the updated LM has a new vocabulary V' composed of semantic tokens and new word embeddings $E'_{vocab} \in R^{(|V'|,D)}$, where $|V'| < |V|$. We define the grouping ratio as $r_G = |V'|/|V|$.

K-Means. Subwords are grouped via K-Means based on the cosine distance of their word embeddings. We choose K-Means due to its flexibility in the produced number of groups $|V'|$, and use the cosine distance as experiments show that it has better performance, especially at high word embeddings dimensions (Ap. E). The inspections of the groups show that they could reflect coherent semantics (Figure 8 and 13). In this work, we set $|V'|$ to correspond to a grouping ratio $r_G \in \{5\%, 10\%, 20\%, 40\%\}$.[3]

First-k. As the baseline, we keep only the first-k emerged subwords while training the tokenizer, where $k = |V'|$. In this way, the size of the vocabulary and word embedding vectors remains the same as in the corresponding semantically grouped models, yet each embedding corresponds to only a single subword as the original LM.

2.2 Cross-lingual Subword Alignment (CLSA)

Other than clustering on the off-the-shelf word embedding, we investigate manually aligning the embeddings of subwords across different languages. Specifically, we gather cross-lingual word pairs from bilingual dictionaries (i.e., MUSE; Conneau et al., 2017; PanLex[4]) and concept lists of multilingual words (i.e., Concepticon; List et al., 2016; ColexNet; Liu et al., 2023). We only preserve the words that are tokenized into a single subword. Then, the word embeddings are trained using InfoNCE loss (Oord et al., 2018) with in-batch negatives. Note that only the parameters of word embeddings are used and updated in this operation, while the rest of the model remains untouched. The training configurations of the CLSA operation are provided in Ap. 11, as well as the ablations on the source datasets used for CLSA training.

2.3 Dimension Reduction (DR)

Section 4.2 involves experiments that reduce the dimension of word embeddings, where we simply *remove* the final $D - d$ dimensions of the word embeddings to form the new word embeddings $E'_d \in R^{(|V|,d)}$ ($d < D$), and pad each embedding vector with zeros on the fly. Note that the positional and token-type embeddings are not affected by this operation. While there are alternative options to reduce the word embedding parameters, we adopt DR for its simplicity and to minimize changes in model architecture. Due to its great modifications to the word embeddings, this operation is always followed by continual pretraining (Section 2.4).

2.4 Continual Pretraining

The above operations may lead to a potential mismatch between the word embeddings and the rest of the model parameters. To address the mismatch, we continually pretrain the entire LM using Masked Language Modeling (MLM) objectives to align the updated embeddings and language model parameters (Devlin et al., 2019). All continual pretraining uses Wikipedia data of 28 languages,[5] where the languages are selected based on the coverage of downstream tasks. Details on the configurations are provided in Ap. A.

[3] We also investigate grouping based on bilingual lexical mappings in pilot studies. See results comparison in Ap. F.

[4] https://panlex.org/

[5] ar, bg, bn, de, el, en, es, fa, fi, fr, ha, hi, id, ig, ja, ko, lg, ru, rw, sw, te, th, tr, ur, vi, wo, yo, zh

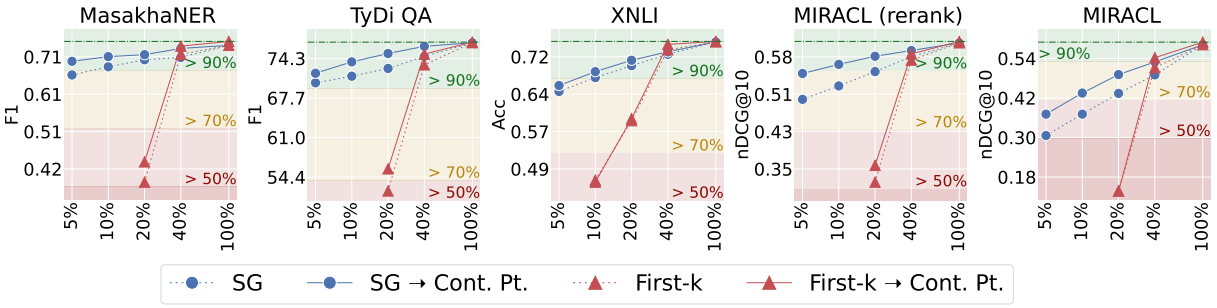

Figure 3: Results of mBERT with vocabulary and embeddings after semantic grouping (SG) or simply reduced size (First-k). **x-axis**: the grouping ratio r_G in log scale. The background colors indicate the relative performance to the oracle results, i.e., continual pretrained mBERT with full vocabulary, indicated by the green dashed lines on top. green: >90%, yellow: 70%–90%, red: 50%–70%. The scores of First-k at $r_G = \{5\%, 10\%\}$ are skipped in the figures as they greatly skew the y-axis.

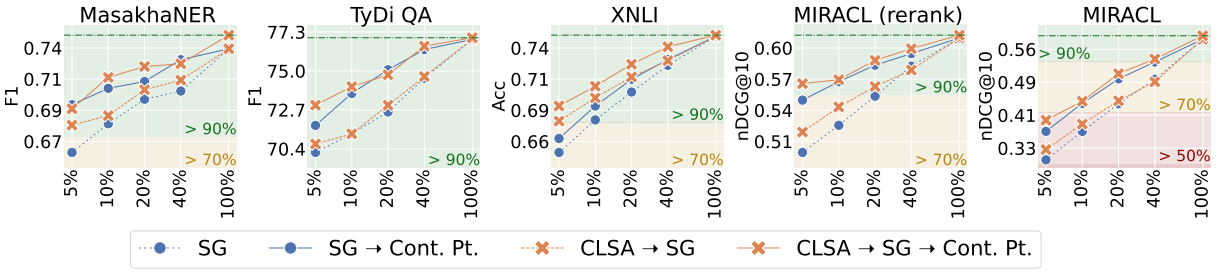

Figure 4: Results of mBERT after semantic grouping (SG) with and without applying cross-lingual subword alignment (CLSA). Background colors design is identical to Figure 3.

3 Downstream Tasks Evaluation

We evaluate all configurations on five multilingual downstream tasks, where four are classification and one is embedding. The classification tasks include word-level understanding, sentence-level understanding, and passage reranking, whereas the embedding task includes passage retrieval. All evaluation datasets cover at least 10 diverse languages. See Table 1 for details. We hope these provide a comprehensive evaluation regarding the task nature, model structures, and languages.

MasakhaNER (Adelani et al., 2021) is a named entity recognition (NER) benchmark including 10 under-represented African languages. We use its version 1.0 in this work.

TyDiQA-GoldP (Clark et al., 2020) is the gold passage task of TyDiQA (Clark et al., 2020), a question answering (QA) dataset that includes 11 topologically different languages. It requires predicting correct answer based on gold answer passages. We refer to it as TyDiQA for simplicity.

XNLI (Conneau et al., 2018) is a natural language inference (NLI) dataset that extends the development and test sets of MultiNLI (Williams et al., 2018) to 15 diverse languages.

MIRACL (Zhang et al., 2023) is a monolingual information retrieval dataset that provides training data for 16 diverse languages and evaluation data for an additional 2 languages. Two tasks are performed on MIRACL: passage retrieval and passage reranking, which fall under embedding and classification tasks respectively. In the rest of this paper, MIRACL refers to the passage retrieval task and MIRACL (rerank) refers to the passage reranking task. We use the classic DPR (Karpukhin et al., 2020) and monoBERT (Nogueira and Cho, 2019; Nogueira et al., 2019) models for each task.

4 Results and Analysis

4.1 Semantic Grouping (SG)

Figure 3 illustrates the results of applying SG on mBERT (Devlin et al., 2019) across a spectrum of grouping ratios r_G (x-axis). The background colors indicate the range of relative effectiveness compared to the oracle results, which are the scores of the original mBERT with continual pretraining applied. Each point in the figures represents the average score across all languages for the given configurations.[6]

[6] Due to space constraint, Section 4 only provides visualization of all results, where numerical scores can be found in

Semantically similar subwords could share the same embeddings to a large degree. In classification tasks (the left four sub-figures), applying SG alone (the blue dashed lines) can already preserve over 90% effectiveness on the downstream tasks with 10% of the original vocabulary size. After applying continual pretraining (the blue solid lines), the same level of effectiveness (> 90%) can be preserved with only 5% of the original vocabulary size. The embedding task (the rightmost sub-figure) is comparatively more sensitive to the semantic grouping, yet still maintains over 85% effectiveness with 20% of the original vocabulary size after the continual pretraining. The different behaviors suggest that classification tasks may require only coarse semantic representations to make predictions, while the embedding tasks require more fine-grained lexical representations to produce reasonable sentence- or passage-level representations.

Comparison with First-k. Figure 3 also shows the results of First-k (the red lines) described in Section 2.1. This is to compare SG with mLMs that have the same number of word embedding entries by adopting a smaller vocabulary size. As the figure shows, while First-k could still share similar effectiveness with SG at $r_G = 40\%$, removing more subwords from the vocabulary deteriorates the effectiveness drastically on all downstream tasks, regardless of whether continual pretraining is applied. This pair of results suggests that while simply reducing the number of subwords has a detrimental effect on the mLMs capacity, which echoes previous findings (Conneau et al., 2020; Liang et al., 2023; Ali et al., 2024; Tao et al., 2024), not all subwords require a unique representation. In other words, the effect of the number of the word embedding entries should be disentangled from the size of the vocabulary.

Enhance the semantic similarity via post-hoc operations. In the above results, the semantic grouping is applied on the off-the-shelf mLMs, exploiting the spatial structure of the untouched word embedding. *Could the similarities of tokens under the same semantic concepts be further enhanced by post-hoc operations on the mLMs?* We show that this direction is possible and promising, using the cross-lingual subword alignment (CLSA) operation as an example.

Figure 4 compares the results of applying SG on mBERT with and without CLSA, where the orange lines are consistently higher than or similar to the blue ones regardless of whether the model has been continually pretrained. On certain datasets, e.g., XNLI and MIRACL, CLSA brings visible improvement at all grouping ratios, with more significant improvements at lower grouping ratios on the other datasets. This shows that the semantic similarity among the subwords has room to be improved by post-hoc operations.[7]

4.2 Eliminating the Confounding Effect of Embedding Parameters

While applying the SG, the number of the word embedding parameters is also reduced by the same degree. Do the changes on effectiveness truly stem from the semantic grouping, or any form of word embedding parameters reduction have the same effect? To address the concern, we apply SG on word embeddings with reduced embedding dimension via the dimension reduction (DR) operation, to see whether the effectiveness diminishes as the embedding parameters are reduced. We first show that SG maintains the same level of effectiveness on the reduced dimensions. Moreover, while the embedding dimension has reached its limit, SG could further push down the overall word embedding parameters to a level that could not be achieved by reducing the word embedding dimension alone.

SG maintains the same level of effectiveness with reduced embedding dimensions. Figure 5 shows the results of SG with three word embedding dimensions $d \in \{768, 128, 32\}$, with both CLSA and continual pretraining applied. While DR reduces the effectiveness of the mLM with full vocabulary, the slope of the curve flattens as the dimension reduces, indicating less relative effectiveness drop on the downstream tasks at lower embedding dimension d. Specifically, the SG results on embedding dimension $d = 32$ is on par and even outperform the results on the initial dimension $d = 768$ as vocabulary size reaches 5%. These results show that the effect of SG is largely independent from the embedding parameters.

Ap. G, Table 5; We also investigate the impact on individual language in Ap. C. While the overall trend per language is similar, we do not observe a consistent impact over languages across different benchmarks.

[7]Note that CLSA is used as a proof-of-concept instead of a general solution: As mentioned in Section 2.2, CLSA is applied to words that are tokenized into single subword, which limits the coverage on the entire vocabulary intrinsically. Thus instead of promoting the CLSA method itself, we use its success in enhancing the semantic similarity as a proof-of-concept that further post-hoc operations on the word embedding can be beneficial and worth further exploration.

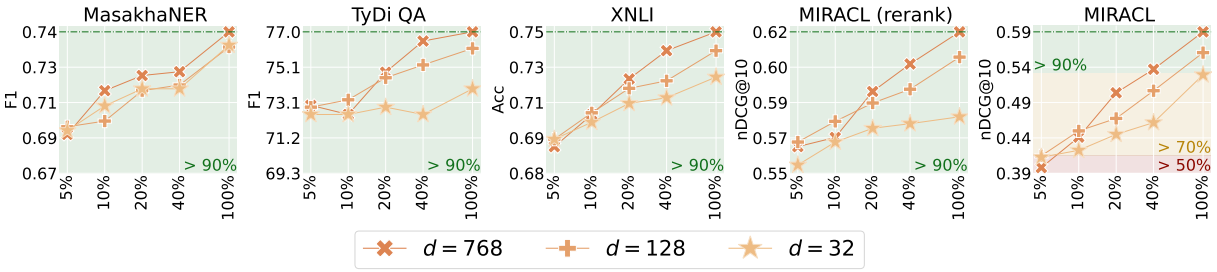

Figure 5: Results of semantic grouping with word embedding dimension $d \in \{768, 128, 32\}$, with CLSA and continual pretraining. Background colors design is identical to Figure 3. The relative performance of all classification tasks are higher than 90%, thus with full green background.

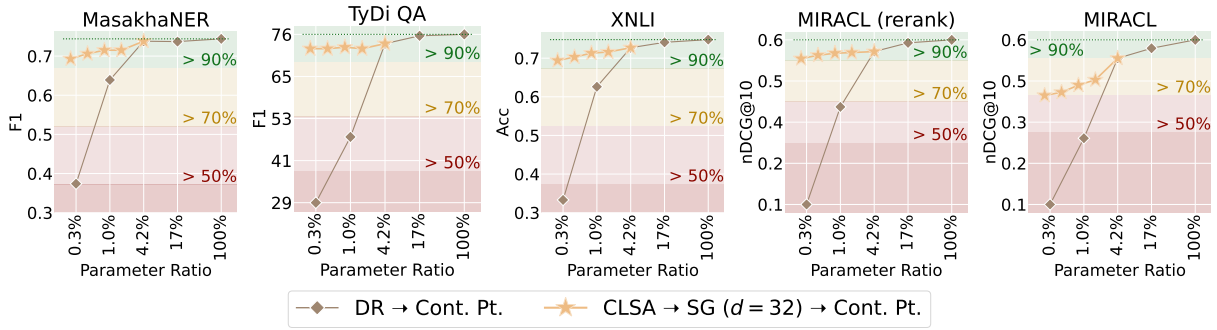

Figure 6: Comparison the effectiveness of Dimension Reduction (DR) and semantic Grouping (SG) at the same level of word embedding parameters size. **x-axis**: the ratio of the word embedding size after DR or SG in log scale. The brown lines: DR with dimension $d \in \{2, 8, 32, 128, 768\}$ and full vocabulary, corresponding to embedding parameters ratio of {0.3%, 1.0%, 4.2%, 17%, 100%}. The yellow lines: SG with dimension $d = 32$ and vocabulary grouping ratio $r_G \in \{5\%, 10\%, 20\%, 40\%\}$, corresponding to embedding parameters ratio of {0.21%, 0.42%, 0.84%, 1.7%, 4.2%}. Background colors design is identical to Figure 3.

SG achieves lower number of embedding parameters beyond dimension reduction. Apart from the above observation, Figure 6 unites the results of SG at embedding dimension $d = 32$ and DR from the perspective of the total number of word embedding parameters (x-axis). The brown lines show the DR results of reducing the embedding dimension d from 768 to $\{2, 8, 32, 128\}$ with the vocabulary untouched, yielding to embedding parameters ratio of {0.3%, 1.0%, 4.2%, 17%}. With DR, the effectiveness on all downstream tasks is largely preserved until dimension $d = 32$ and drop sharply once the dimension falls beneath it, indicating that saving the embedding parameters via dimension reduction has reach its limit.

On the other hand, the yellow lines show the models starting from word embedding with dimension $d = 32$ (thus 4.2% embedding parameters) and then applied SG with grouping ratio r_G from 40% to 5%,[8] yielding to embedding parameters ratio from 4.2% to 0.21%. The effectiveness differences between the two lines are clear: while further reducing the embedding parameters by similar scale, SG could largely preserve the downstream effectiveness whereas DR fails miserably. This strongly informs that SG is complementary to DR on their effect towards the word embedding parameters, and that it may provide a new perspective on understanding the necessary parameters in the word embeddings.

4.3 Backbones

Experiments above are all based on mBERT (Devlin et al., 2019), how would the findings generalize to mLMs with different tokenization algorithms, vocabulary size, model size, and pretraining corpora? We select three additional mLMs to address the above concern: XLM-R (base), XLM-R (large) (Conneau et al., 2020), and XLM-V (base) (Liang et al., 2023), which all deploy ULM (Kudo and Richardson, 2018) to construct the vocabulary, and pretrained on CC100, but differ from each other in terms of model sizes, tokenization pre-processing, vocabulary allocation, and vocabulary size.

[8] Identical to the yellow lines in Figure 5.

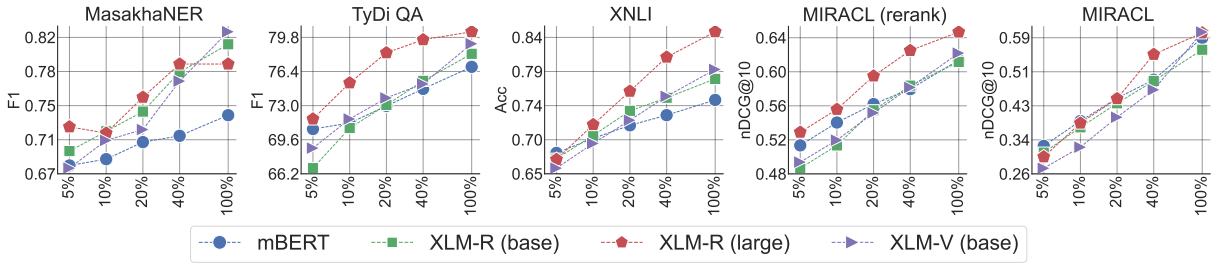

Figure 7: Results of SG on multiple backbones, all with CLSA applied and no continual pretraining. All models show similar trend regarding semantically grouped subwords.

Results of SG are shown in Figure 7, where background colors are omitted as different backbones do not share the same oracle results. Instead, we compare the results of the other backbones with mBERT (the blue lines) to see whether they follow similar trend as more subwords are semantically grouped. Overall, the slopes of the four curves are similar on all five benchmarks, indicating that the findings are likely to generalize over different multilingual LMs with various tokenization processes and model sizes.

4.4 Semantic Similarities Categories among the Grouped Subwords

This section provides insights into the semantic similarities among grouped subwords. Figure 8 illustrates examples of eight semantic groups, which are based on the model with embedding dimension $d = 768$ and the application of CLSA.[9]

Among the eight displayed groups examples displayed, six of them are selected from the Swadesh list (Swadesh, 1952), a widely used compilation of basic concepts across languages. The keywords are selected to cover various parts of speech, including pronouns, nouns, adjectives, prepositions, and verbs. The remaining two groups are labeled as "NUMBERS" and "PUNCTUATIONS", which exhibit strong clustering upon manual inspection. Overall, we identify several patterns of semantic similarities among the grouped subwords:

1. Semantically identical or similar words across languages: *mother vs. mẹ, they vs.* 그들은 .[10]
2. Semantically related words from the same language: *at vs. on, heavy vs. hard.*

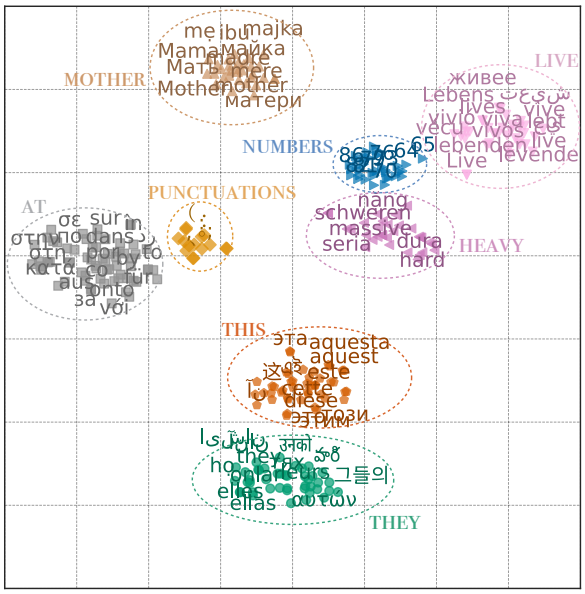

Figure 8: Eight semantic tokens formed by grouped subwords on mBERT word embedding with dimension $d = 768$ and CLSA applied. The labels outside the circles are either the keyword of the cluster, or "NUMBERS"/ "PUNCTUATIONS" if the cluster is a collection of numbers or punctuations.

3. Semantically related words across languages: *heavy vs. intenso, who vs. που*.[11]
4. Numbers in similar range: *the cluster numbers ranging from 60 to 100.*
5. Punctuations: *., ;!*

We also notice that some words are not grouped based on desired semantics: The current methods are limited in the polysemous situation and controlling the desired semantic per group: For example, in the group of "I", while the desired semantic is the first person singular pronouns in different languages, the group mainly includes the single letter such as "A", "B", "C" on mLMs with $d = 768$, as shown by Figure 13 in Ap. H, suggesting that future work is necessary to better handle such cases.

[9] Note that only parts of the subwords are displayed per group for space limit. For the full list of subwords, please refer to Ap. H and Figure 13.

[10] *mẹ*: "mother" in Vietnamese; *Cette*: "this" in French; 그들은 : "they" in Korean.

[11] *intenso*: "intense" in Spanish; *που*: "where" in Greek.

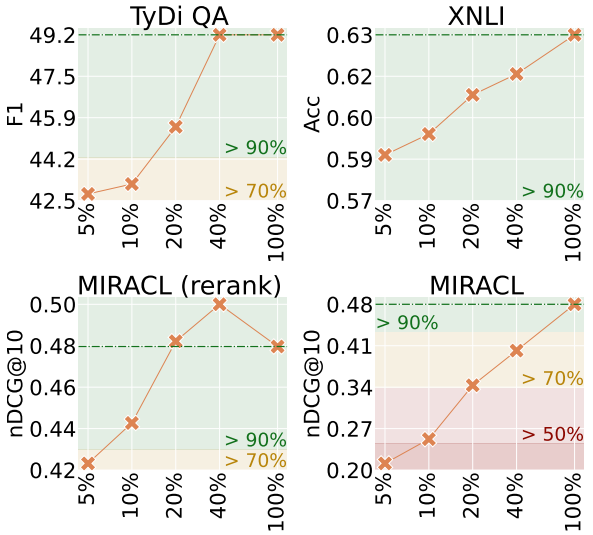

Figure 9: Results when transferring from English to other languages on four benchmarks, with CLSA and continual pretraining applied. MasakhaNER is skipped as English training data is not available. Background colors design is identical to Figure 3. The dashed green line denotes the zero-shot performance without semantic grouping.

4.5 Cross-lingual Transfer

Lastly, we investigate how the semantic grouping affects the cross-lingual transferability from English to other languages. We evaluate on four of the above tasks, where MasakhaNER is skipped in this section as English training data is not available.

Surprisingly, on two classification tasks: TyDiQA and MIRACL (rerank), the zero-shot results at grouping ratio $r_G \in \{40\%, 20\%\}$ are on par or even better than the oracle results, where subwords and embeddings are not semantically grouped. That is, fine-grained semantics show similar cross-lingual transferability with word-based understanding in some circumstances, suggesting that they may serve as the anchors for cross-lingual transferring in the current language models.

On the other hand, cross-lingual transfer on embedding tasks is more challenging especially with small group ratio, i.e., coarse-grained semantic tokens. While the zero-shot results on MIRACL at $r_G \in \{20\%, 40\%\}$ remains in the reasonable range — achieving more than 70% of the relative zero-shot effectiveness — the relative zero-shot results falls under 50% when r_G further drops. This mirrors the in-domain results from previous sections that embedding tasks require more fine-grained semantic information to form effective sentence or paragraph representations.

5 Related Works

Semantic Similarities in Language Models. The semantics latent knowledge in word embeddings have been leveraged to adapt the pretrained multilingual LMs to unseen languages or scripts (Pfeiffer et al., 2021; Wang et al., 2022; Liu et al., 2023). While it has been shown that the shared semantics assist the language transfer, it is unclear how much the semantics alone could achieve compared to the full-fledged models. Another line of works focuses on the isomorphic representations of *contextualized* word embeddings. Li et al. (2024) find that vision and language models shows isomorphic representations, and Peng and Søgaard (2024) reveal similar conclusion on the contextualized representations of multilingual large language models. Two additional works also study the contextualized concept encoded in the language models: Sajjad et al. (2022) analyze how latent concepts are encoded in representations and how they align with human-defined concepts. Shani et al. (2023) analyze how large language models encode the TYPEOF relations of concepts, and propose a model-agnostic, proof-of-concept method to shift the model to a concept-level understanding.

Bilingual Lexicon Induction and Word Alignment. The Cross-lingual Subword Alignment (CLSA) operation is related to the task of *Bilingual Lexicon Induction (BLI)* and *Word Alignment*. BLI aims to induce the equivalent translation in language L_2 given a word in language L_1 (Artetxe et al., 2016; Conneau et al., 2017; Wang et al., 2020; Shi et al., 2021). CLSA is similar to BLI in terms of focusing on uncontextualized cross-lingual words pairs, but different in that it aims to align the word pairs in the embedding space and targets on not only bilingual but also multilingual words. Word alignment aims to find bilingual word pairs in parallel sentences (Cao et al., 2020; Jalili Sabet et al., 2020). While names are similar, the goals are different: In addition to above differences with BLI, CLSA focuses on uncontextualized cross-lingual word pairs and does not involve parallel sentences.

Word Sense Clustering. The Semantic Grouping (SG) operation is related to the task of word sense clustering. Although both share similar objectives, word sense clustering primarily relies on corpus statistics (Snow et al., 2007), whereas the SG operation is based on the similarity of word embeddings.

Parameter Redundancy and Model Compression. Dalvi et al. (2020) study the layer and neuron redundancy on BERT and XLNet, and many works propose to compress the overall model size via pruning (Gordon et al., 2020; Ashkboos et al., 2024; Yang et al., 2024), knowledge distillation (Turc et al., 2019; Sanh et al., 2019), and quantization (Shen et al., 2020). From the perspective of model compression, our work provides a new view on the word embeddings redundancy from the perspective of shared semantics among subwords.

6 Conclusion

Inspired by how human understand text based on semantic concepts rather than superficial forms, this work studies the degree to which current multilingual language models understand based on subword-level semantic concepts. We find that the general shared semantics could get the models a long way in understanding languages and making predictions, especially for classification tasks. Additional experiments show that the observations generalize across mLMs with different tokenization algorithm, vocabulary size, model size, and pre-training corpora. Inspections on the grouped subwords suggest that they exhibit multiple patterns of semantic similarities, including synonyms and word translations in many languages.

Not only the subword-level semantics is prominent in in-domain language understanding, in some cases, it also serves as anchors of cross-lingual transfer and thus potentially a promising direction of bridging the understanding of different languages. We hope that this work sheds light on understanding the multilingual vocabulary and word embeddings from the semantic perspective, and spurs further research on shared semantic information at the subword level across languages.

Limitations

The Scope of Semantics. This work only discusses the application scenario where pragmatics are not heavily involved. Other situations such as poetry, humour, sentiment analysis intuitively would require not only the semantic meanings, but also exquisite understanding of the words nuances, yet out of the scope of this work. Similarly, the work is probably not applicable to figurative language such as metaphor, irony, etc.

Word-level Semantics Only. One of the major limitation of this work is not consider the phrase-level semantics in the study.

Encoder-only Models. As a natural limitation of the semantical grouping method itself, it is not straightforward to extend the method to decoder-only models since it forbids predicting explicit subword at each decoding step. Thus only the encoder-only models and tasks are evaluated in this work. Further design and exploration would be required to apply the method to decoder-only models.

Results on Embedding Tasks. Results show that the embedding tasks are more sensitive to the semantical grouping compared to the classification tasks. More questions could be raised from the phenomenon: does it make embedding task a better evaluation metrics for the semantical grouping, or that simply the embedding task require more fine-grained understanding of the subword information? How much could embeddings benefit if the semantical grouping algorithm could be improved? We believe that these are important questions to further understand effectiveness and *limitation* of this direction. Limited by the paper capacity, we leave them for future exploration.

Acknowledgement

We greatly thank Jiarui Xu, Xinyu Shi, Xueguang Ma, Raphael Tang, Sanket Vaibhav Mehta, and the anonymous reviewers for their valuable discussions and feedback.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga

Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Buschhoff, Charvi Jain, Alexander Weber, Lena Jurkschat, Hammam Abdelwahab, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Samuel Weinbach, Rafet Sifa, Stefan Kesselheim, and Nicolas Flores-Herr. 2024. Tokenizer choice for LLM training: Negligible or crucial? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3907–3924, Mexico City, Mexico. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas. Association for Computational Linguistics.

Saleh Ashkboos, Maximilian L. Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. 2024. SliceGPT: Compress large language models by deleting rows and columns. In *The Twelfth International Conference on Learning Representations*.

Alan Bundy and Lincoln Wallen. 1984. *Semantic Primitives*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *ArXiv*, abs/1710.04087.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Fahim Dalvi, Hassan Sajjad, Nadir Durrani, and Yonatan Belinkov. 2020. Analyzing redundancy in pretrained transformer models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4908–4926, Online. Association for Computational Linguistics.

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.

Mattia Fumagalli, Roberta Ferrario, et al. 2019. Representation of concepts in ai: Towards a teleological explanation. In *JOWO*.

Liane Gabora, Eleanor Rosch, and Diederik Aerts. 2008. Toward an ecological theory of concepts. *Ecological Psychology*, 20(1):84–116.

Peter Gardenfors. 2014. *The geometry of meaning: Semantics based on conceptual spaces*. MIT press.

Cliff Goddard and Anna Wierzbicka. 2013. *Words and Meanings: Lexical Semantics Across Domains, Languages, and Cultures*. Oxford University Press.

Mitchell Gordon, Kevin Duh, and Nicholas Andrews. 2020. Compressing BERT: Studying the effects of weight pruning on transfer learning. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 143–155, Online. Association for Computational Linguistics.

Ray Jackendoff. 1988. Conceptual semantics. *Meaning and mental representations*, 1:81–97.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for

open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Jiaang Li, Yova Kementchedjhieva, Constanza Fierro, and Anders Søgaard. 2024. Do Vision and Language Models Share Concepts? A Vector Space Alignment Study. *Transactions of the Association for Computational Linguistics*, 12:1232–1249.

Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. XLM-V: Overcoming the vocabulary bottleneck in multilingual masked language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13142–13152, Singapore. Association for Computational Linguistics.

Johann-Mattis List, Michael Cysouw, and Robert Forkel. 2016. Concepticon: A resource for the linking of concept lists. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2393–2400, Portorož, Slovenia. European Language Resources Association (ELRA).

Yihong Liu, Haotian Ye, Leonie Weissweiler, Renhao Pei, and Hinrich Schuetze. 2023. Crosslingual transfer learning for low-resource languages based on multilingual colexification graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8376–8401, Singapore. Association for Computational Linguistics.

Gregory Murphy. 2004. *The big book of concepts*. MIT press.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *ArXiv*, abs/1901.04085.

Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy J. Lin. 2019. Multi-stage document ranking with BERT. *ArXiv*, abs/1910.14424.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748.

Qiwei Peng and Anders Søgaard. 2024. Concept space alignment in multilingual LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5511–5526, Miami, Florida, USA. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021. UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hassan Sajjad, Nadir Durrani, Fahim Dalvi, Firoj Alam, Abdul Khan, and Jia Xu. 2022. Analyzing encoded concepts in transformer language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3082–3101, Seattle, United States. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Chen Shani, Jilles Vreeken, and Dafna Shahaf. 2023. Towards concept-aware large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13158–13170, Singapore. Association for Computational Linguistics.

Sheng Shen, Zhen Dong, Jiayu Ye, Linjian Ma, Zhewei Yao, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. 2020. Q-BERT: Hessian based ultra low precision quantization of bert. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8815–8821.

Haoyue Shi, Luke Zettlemoyer, and Sida I. Wang. 2021. Bilingual lexicon induction via unsupervised bitext construction and word alignment. 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 813–826, Online. Association for Computational Linguistics.

Rion Snow, Sushant Prakash, Daniel Jurafsky, and Andrew Y. Ng. 2007. Learning to merge word senses. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1005–1014, Prague, Czech Republic. Association for Computational Linguistics.

Morris Swadesh. 1952. Lexico-statistical dating of prehistoric ethnic contacts. *Proceedings of the American Philosophical Society*, 96:452–463.

Chaofan Tao, Qian Liu, Longxu Dou, Niklas Muennighoff, Zhongwei Wan, Ping Luo, Min Lin, and Ngai Wong. 2024. Scaling laws with vocabulary: Larger models deserve larger vocabularies.

Bill Thompson, Seán G Roberts, and Gary Lupyan. 2020. Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour*, 4(10):1029–1038.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2022. Expanding pretrained models to thousands more languages via lexicon-based adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 863–877, Dublin, Ireland. Association for Computational Linguistics.

Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime G. Carbonell. 2020. Cross-lingual alignment vs joint training: A comparative study and a simple unified framework. In *ICLR*. OpenReview.net.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Yifei Yang, Zouying Cao, and Hai Zhao. 2024. Laco: Large language model pruning via layer collapse. *ArXiv*, abs/2402.11187.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

	mNER	TyDiQA	XNLI	MIRACL	
				rerank	retrieval
epochs	50	3	2	5	40
warmup ratio	–	–	–	0.1	0.1
batch size	64	128	64	32	256
learning rate	5e-05	5e-05	5e-06	5e-06	1e-05
adam β_1	0.9	0.9	0.9	0.9	0.9
adam $beta\_2$	0.99	0.99	0.99	0.99	0.99

Table 2: Downstream task traning configurations. **mNER: MasakhaNER**

A Training and Evaluation Configurations

Continual Pretraining. All continual pretraining in this work share the same hyperparameters. Language models are trained on the MLM objective for 25,000 steps, with a batch size 1024 and a learning rate of $1e-4$. We randomly masked 15% of the tokens. The hyperparameters were chosen following the initial pretraining configurations of Devlin et al. (2019).

Downstream Task. For each task, we use all the available training data provided by each dataset. The training configurations are provided in Table 2.

B Alignment Datasets

Figure 11 compares the impact of different word alignment datasets on the downstream tasks. All experiments are followed by $r_G = 5\%$ semantic grouping and continual pretraining. On all downstream datasets, we compare the results of using four groups of alignment data:

1. MUSE
2. MUSE and Round-Trip
3. MUSE and PanLex
4. MUSE, PanLex, Colex, and Concepticon

where Round-Trip are the pair of word that are the nearest neighbors to each other in the embedding space, serving as a regularization in the CLSA procedure. We found that scenario 4 gives the best overall results, and thus use it as our default configuration.

C Impact on Individual Languages

This section explores whether the languages are consistently affected across tasks by the semantic grouping. To this end, we compare the effectiveness drop on the overlapping languages of each pair of benchmarks, and compute their Pearson correlation coefficient. Five pairs of benchmarks are selected for analysis, which fall under 3 groups:

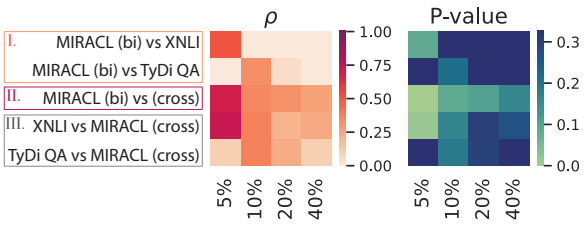

Figure 10: Pearson correlation between the relative performance drop per language between a pair of benchmarks. Each row indicates a pair of benchmarks, where (bi) refers to retrieval (which uses bi-encoder) and (cross) refers to reranking (which uses cross-encoder). Each column indicates a semantic grouping rate r_G. *Left*: Pearson correlation coefficient ρ; *Right*: corresponding one-tail p-values.

I. **different task types and data sources**:
   (a) MIRACL (retrieval) vs. XNLI
   (b) MIRACL (rerank) vs. TyDiQA
II. **different tasks types, same data source**: MIRACL (retrieval) vs. MIRACL (rerank)
III. **same task type, different data sources**:
   (a) MIRACL (rerank) vs. XNLI
   (b) MIRACL (rerank) vs. TyDiQA

The benchmark selection is mainly under the consideration of the number of overlapping languages: MasakhaNER have no overlapping languages with the other datasets, and TyDiQA and XNLI only have 3 overlapping languages.

Figure 10 shows Pearson correlation coefficient ρ (left) and the corresponding p-values (right) in two heatmaps. In two heatmaps, higher saturation indicates higher ρ or smaller p-values, respectively, which together indicates stronger correlations in higher confidence. Each row corresponds to a pair of benchmarks, and each column corresponds to a semantic grouping ratio r_G. The three blocks in the figure corresponds to the three groups defined above from top to bottom.

Overall, we observe consistent trend on the two heatmaps, where the top-2 rows are smaller in the coefficient ρ (lighter color in the left heatmap) and larger in p-values (darker color in the right heatmap), and that the bottom-3 rows are larger in the coefficient ρ and smaller in p-values. This indicates that languages are affected similarly by the semantic grouping when the benchmarks pair share either the same data source (group 2) or the task type (group 3). In contrast, the impact on the languages is less consistent across benchmarks that

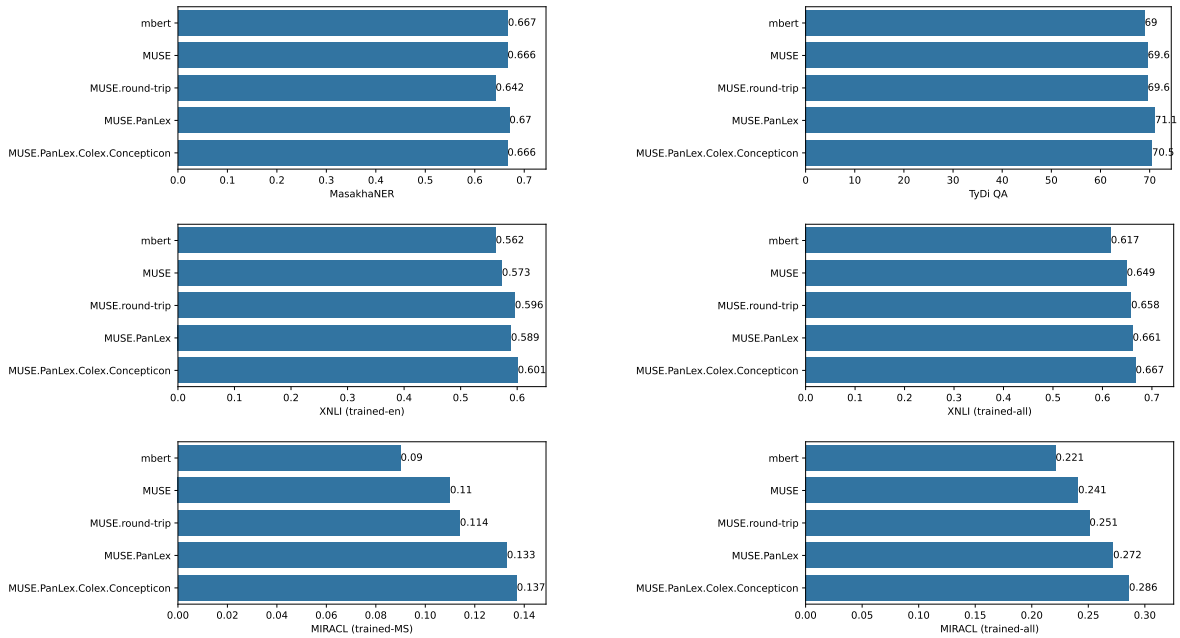

Figure 11: Comparison of alignment source dataset on downstream tasks. Each bar represents one alignment dataset(s), where "mbert" is the baseline result when there is no alignment applied. The x-axis is the average score on all languages per task.

d	r_G	# Para (M)	GPU Mem (G)	steps / sec
768	5.0%	90	19.1	5.1
	10.0%	95	39.8	2.6
	20.0%	104	51.5	2.1
	40.0%	123	74.9	2.1
	100.0%	178	74.6	1.9
128	5.0%	86	19.1	5.2
	100.0%	101	74.6	1.9
32	5.0%	86	19.1	5.3
	100.0%	89	74.6	1.9

Table 3: Efficiency statistics of mBERT under SG and DR, collected on a 80G A100 GPU during pretraining with batch size 128 per device. (green: best; red: worse; yellow: neutral)

shares neither the task type nor the data source (group 1).

D Discussion on Memory and Efficiency

This section discusses the effect of semantic grouping on the model size, and its memory usage and training speed during the continual pretraining. Table 3 shows above statistics of mBERT with $r_G \in \{5\%, 10\%, 20\%, 40\%, 100\%\}$ with word embedding dimension $d = 768$, and $r_G \in \{5\%, 100\%\}$ with $d \in \{128, 32\}$.[12]

[12]Statistics of LMs with $r_G \in \{10\%, 20\%, 40\%\}$ and $d \in \{128, 32\}$ are similar to $d = 768$, thus skipped for simplicity.

Model Size. The model size is affected linearly with the vocabulary size or the word embedding dimension. As the word embedding initially takes over half of the total model parameters in mBERT, grouping the subwords to 5% of the vocabulary size brings visible savings on the overall model size, from 178M to 90M.

Memory. We found that memory usage during the pretraining could be largely saved via reduced vocabulary size, but not the word embedding dimension. We explain it by that the memory usage during the pretraining is bottlenecked by the activations, especially the final token-level logit matrix, whose size is solely determined by the vocabulary size but not the word embedding dimension. As a result, the memory savings from compact vocabulary is prominent, from 74.6G to 19.1G when reducing the vocabulary size from 100% to 5%, while saving the word embedding dimension barely changes the memory usage.

Training Speed. The trend of training speed is similar to the memory usage, where saving the word embedding dimension has negligible effect on the training speed while saving the vocabulary size has significant impact, from 1.9 to 5.1 steps per second.

Method	r_G	nDCG@10
Oracle	100%	0.452
Grouping on MUSE (zh)	99.2%	0.357
Grouping on MUSE (5L)	92.1%	0.248
Grouping on MUSE (all)	86.2%	0.193
K-Means	40%	0.304

Table 4: Results on MIRACL (zh), comparing Grouping via Bilingual Lexicons vs K-Means. Models are fine-tuned on MS MARCO, without CLSA or continual pretraining.

E Ablation on Distance Metric

We compare the results using cosine versus Euclidean distance in Figure 12, where the results show that grouping on Euclidean distance greatly underperformance cosine distance especially at higher dimension ($d = 768$). We interpret this as that the vector norm is an undesired feature when pursuing the semantic similarity between the subwords, which amplifies the distance between semantic similar subword at high dimension.

F Grouping via Bilingual Lexicons

As an intuitive alternative to grouping via K-Means, we explored to group the subwords via the ground-truth bilingual lexicons in the preliminary experiments, finding that it has limited coverage on the subwords, thus grouping ratios, and also underperforming the K-Means-based grouping. See Table 4 for the results.

G Numerical Results of Experiments in Figure 2–6

Table 5 presents the numerical results of experiments in Figure 3, 4, 5, 6, and 7.

H More Inspection Examples

Figure 13 shows more examples of the grouped subwords additional to Figure 8.

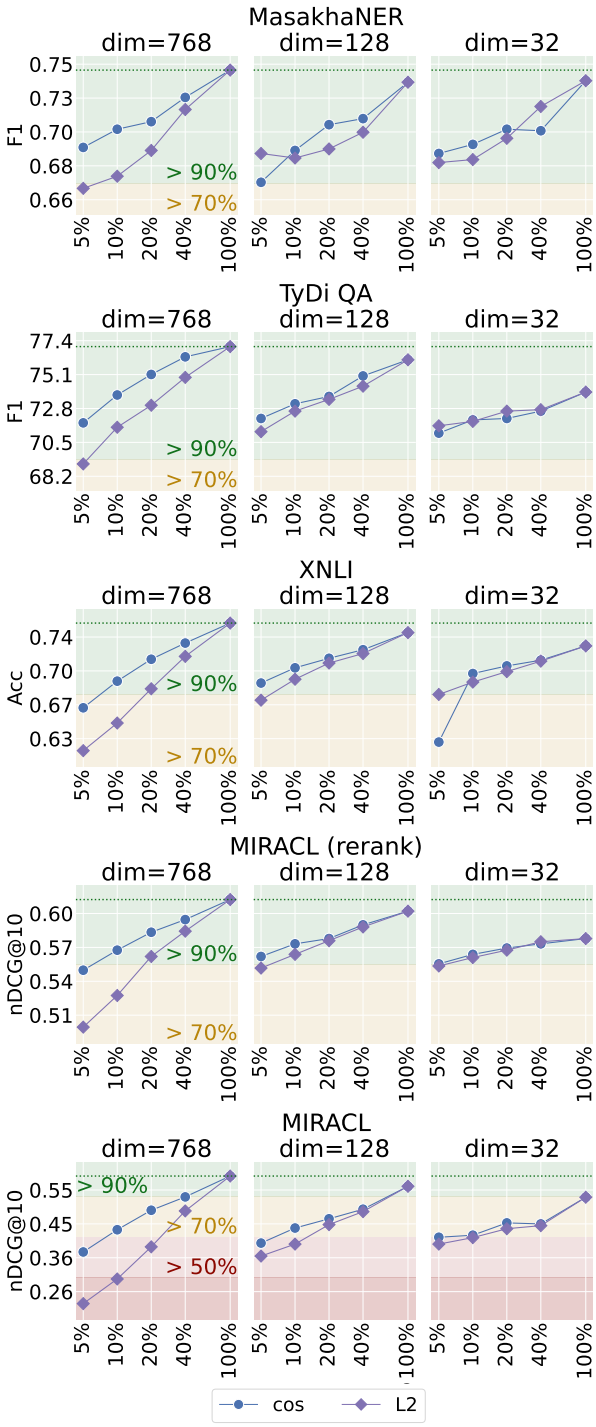

Figure 12: Comparison of L2 versus cosine distance when using K-Means to group the subwords.

d	r_G	w/ CLSA?	w/ PT?	MasakhaNER		TyDiQA		XNLI		MIRACL (cross-)		MIRACL (bi-)	
				F1	perc	F1	perc	Acc	perc	nDCG@10	perc	nDCG@10	perc
768	100%	–	×	0.735	98.7%	76.9	99.9%	0.750	99.9%	0.613	99.5%	0.586	98.7%
768	100%	–	✓	0.745	100.0%	77.0	100.0%	0.751	100.0%	0.616	100.0%	0.594	100.0%
						Figure 3							
768	40%	×	×	0.704	94.5%	74.6	96.9%	0.725	96.5%	0.584	94.8%	0.493	83.0%
768	20%	×	×	0.698	93.7%	72.6	94.3%	0.702	93.5%	0.553	89.8%	0.436	73.4%
768	10%	×	×	0.680	91.3%	71.3	92.6%	0.678	90.3%	0.523	84.9%	0.372	62.6%
768	5%	×	×	0.659	88.5%	70.2	91.2%	0.650	86.6%	0.495	80.4%	0.306	51.5%
768	40%	×	✓	0.727	97.6%	76.3	99.1%	0.730	97.2%	0.597	96.9%	0.533	89.7%
768	20%	×	✓	0.711	95.4%	75.1	97.5%	0.713	94.9%	0.585	95.0%	0.494	83.2%
768	10%	×	✓	0.706	94.8%	73.7	95.7%	0.690	91.9%	0.568	92.2%	0.437	73.6%
768	5%	×	✓	0.694	93.2%	71.8	93.2%	0.662	88.1%	0.549	89.1%	0.372	62.6%
						Figure 4							
768	40%	✓	×	0.712	95.6%	74.7	97.0%	0.729	97.1%	0.580	94.2%	0.487	82.0%
768	20%	✓	×	0.705	94.6%	73.0	94.8%	0.715	95.2%	0.563	91.4%	0.443	74.6%
768	10%	✓	×	0.686	92.1%	71.3	92.6%	0.697	92.8%	0.542	88.0%	0.389	65.5%
768	5%	✓	×	0.679	91.1%	70.7	91.8%	0.677	90.1%	0.516	83.8%	0.330	55.6%
768	40%	✓	✓	0.724	97.2%	76.5	99.4%	0.741	98.7%	0.602	97.7%	0.540	90.9%
768	20%	✓	✓	0.722	96.9%	74.8	97.1%	0.726	96.7%	0.590	95.8%	0.506	85.2%
768	10%	✓	✓	0.714	95.8%	74.1	96.2%	0.707	94.1%	0.570	92.5%	0.442	74.4%
768	5%	✓	✓	0.691	92.8%	73.0	94.8%	0.690	91.9%	0.566	91.9%	0.398	67.0%
						Figure 5							
128	40%	✓	✓	0.717	96.2%	75.2	97.7%	0.725	96.5%	0.591	95.9%	0.509	85.7%
128	20%	✓	✓	0.714	95.8%	74.5	96.8%	0.721	96.0%	0.585	95.0%	0.469	79.0%
128	10%	✓	✓	0.698	93.7%	73.3	95.2%	0.708	94.3%	0.577	93.7%	0.451	75.9%
128	5%	✓	✓	0.695	93.3%	72.9	94.7%	0.694	92.4%	0.568	92.2%	0.415	69.9%
32	40%	✓	✓	0.712	95.6%	74.7	97.0%	0.729	97.1%	0.580	94.2%	0.487	82.0%
32	20%	✓	✓	0.705	94.6%	73.0	94.8%	0.715	95.2%	0.563	91.4%	0.443	74.6%
32	10%	✓	✓	0.686	92.1%	71.3	92.6%	0.697	92.8%	0.542	88.0%	0.389	65.5%
32	5%	✓	✓	0.679	91.1%	70.7	91.8%	0.677	90.1%	0.516	83.8%	0.330	55.6%
						Figure 6							
128	100%	–	✓	0.737	98.9%	76.1	98.8%	0.741	98.7%	0.605	98.2%	0.564	94.9%
32	100%	–	✓	0.735	98.7%	76.9	99.9%	0.750	99.9%	0.613	99.5%	0.586	98.7%
8	100%	–	✓	0.639	85.8%	47.8	62.1%	0.626	83.4%	0.414	67.2%	0.276	46.5%
2	100%	–	✓	0.374	50.2%	29.4	38.2%	0.333	44.3%	0.123	20.0%	0.064	10.8%
						Figure 7: XLM-R base							
768	100%	✓	×	0.814	100.0%	78.2	100.0%	0.779	100.0%	0.611	100.0%	0.558	100.0%
768	40%	✓	×	0.783	96.2%	75.5	96.5%	0.753	96.7%	0.584	95.6%	0.484	86.7%
768	20%	✓	×	0.739	90.8%	73.1	93.5%	0.735	94.4%	0.556	91.0%	0.431	77.2%
768	10%	✓	×	0.717	88.1%	70.8	90.5%	0.701	90.0%	0.516	84.5%	0.374	67.0%
768	5%	✓	×	0.695	85.4%	66.8	85.4%	0.669	85.9%	0.490	80.2%	0.314	56.3%
						Figure 7: XLM-R large							
1024	100%	✓	×	0.792	100.0%	80.4	100.0%	0.844	100.0%	0.645	100.0%	0.598	100.0%
1024	40%	✓	×	0.792	100.0%	79.6	99.0%	0.809	95.9%	0.624	96.7%	0.547	91.5%
1024	20%	✓	×	0.755	95.3%	78.3	97.4%	0.762	90.3%	0.595	92.2%	0.442	73.9%
1024	10%	✓	×	0.715	90.3%	75.3	93.7%	0.716	84.8%	0.557	86.4%	0.384	64.2%
1024	5%	✓	×	0.722	91.2%	71.7	89.2%	0.668	79.1%	0.531	82.3%	0.304	50.8%
						Figure 7: XLM-V base							
768	100%	✓	×	0.828	100.0%	79.2	100.0%	0.792	100.0%	0.621	100.0%	0.600	100.0%
768	40%	✓	×	0.773	93.4%	75.2	94.9%	0.755	95.3%	0.582	93.7%	0.463	77.2%
768	20%	✓	×	0.719	86.8%	73.8	93.2%	0.722	91.2%	0.553	89.0%	0.398	66.3%
768	10%	✓	×	0.707	85.4%	71.7	90.5%	0.690	87.1%	0.522	84.1%	0.327	54.5%
768	5%	✓	×	0.676	81.6%	68.8	86.9%	0.656	82.8%	0.497	80.0%	0.277	46.2%

Table 5: Numerical results of experiments in all figures, where each number is an averaged result of all languages per benchmark. We skip the per-language score due to the space limit. d: word embedding dimension; r_G: grouping ratio; **"cross-"**: cross-encoder **"bi-"**: bi-encoder **perc**: the relative performance to the oracle results (i.e., the second row for mBERT, the corresponding 100% rows for the other backbones) The background colors indicate the relative performance to the oracle results: green: >90%, yellow: 70%–90%, red: 50%–70% and the dark red color means <50%. Better viewed in colors.

Keywords	dim=768	dim=128	dim=32
I	A, B, C, CH, E, H, I, M, N, O, Q, S, V, X, a, d, f, h, j, m, n, p, pm, t, ös, A, O, A, O, м, o, अ, অ, আ, এ, ఒ, ఆ	##इ, I, ich, mu, sì, yo, yá, бы, بی ,میں ,من, तब, तो, ही, ত, তখন, তা, ఈ, 予, 俺, 汝, 私, 저	I, Ich, Ja, ich, lord, me, mich, minut, mir, mis, mo, my, més, mí, self, sí, tôi, 们, 俺, 分, 別, 我, 来, 汝, 私
they	They,chúng,ellas,elles,ellos,eux,heidän,hulle,họ,ihnen,ils,leur,leurs,loro,mereka,niiden,niitä,onlar,onlara,onları,onların,their,them,they,αυτών,τους,им,их,они,тях,ایشان,آنها,آنان,उनकी,उनके,उनको,वे,এদের,তাদের,তারা,వారి,그들은,그들의	Ihre,Leur,Sus,Their,chúng,deres,elles,ellos,euren,eux,heidän,ih,ihnen,ihre,ihrem,ihren,ihrer,ihres,ils,its,leur,leurs,loro,niiden,niitä,oni,onlar,onlara,onları,onların,su,sus,suya,suyos,their,them,they,τους,ги,ими,их,они,свои,тях,هم,ایشان,ان,آنها,آنان,उनकी,उनको,उन्हें,ये,এদের,তাদের,వారి,వారు,그녀의,그들은,그들의	Elles,Lui,They,chúng,dier,elles,estén,eux,,هم##hänelle,hänellä,ihnen,ils,leur,niitä,noi,onlar,onlara,onları,them,they,οποίες,τους,ги,ими,их,उनकी,उन्हें,তাদের,ఇవి,,هم,ایشان,آنها,آنان,них,они,тях,వారు,これに,그들은
this	aquest, aquesta, cette, diese, diesem, dieser, dieses, esta, este, esto, eta, ini, này, questa, questo, this, tämä, tämän, tätä, ésta, éste, тази, това, този, эта, этим, это, этот, این, آن, هذه, هذا, ذلك, এই, এটি, ఇది, ఈ, この, その, 之, 其, 当, 此, 这, 這	##この, ##その, Esta, This, diese, este, this, tämän, tätä, тази, тези, това, този, эта, эти, этой, этот, эту, ఈ, この, 斯, 此, 该, 这	Cette, Diese, Dieser, Esta, Estas, Meine, This, ces, diese, dieser, dieses, estas, este, esto, estos, ini, this, tämä, tämän, tätä, Эта, Эти, тази, тези, това, този, эти, этот, эту, وهذه, ఈ, この, これらの, 这
who	koji, mà, que, qui, that, who, yang, što, που, което, които, който, която, кој, что, што, що, که, جو, אשר, जो, কে, या, だと, 하는	Who, habían, jotka, kdo, ki, koga, quién, who, Кто, Что, кого, кое, кои, кто, що, که, যিনি, 誰, 谁	cual, cuales, joita, joka, jota, jotka, lesquels, mitä, mà, que, qui, quien, quienes, welche, welcher, welches, which, who, yakni, yang, οποία, οποίο, οποίος, который, التي, الذي, جو, والتي, والذي, که, जिसे, जे, जो, だと
heavy	Heavy, ağır, berat, ciddi, difficile, difficult, difícil, dura, dure, duro, greu, hard, heavy, khó, leve, massive, nhẹ, nặng, pesante, schwer, schwere, schweren, seria, serious, seriously, severe, teško, zwaar, теж, 艱, 重, 难	Heavy, ağır, berat, deras, gravemente, heavy, leve, massive, nặng, pek, pesante, schwer, schwere, schweren, seriously, starke, teško, большим, теж, گرم, مشكل, شديد, 沉, 重, 큰	Greatest, Heavy, agir, ağır, deras, forte, giảm, heavy, intensa, intense, intenso, massive, nặng, pesante, plein, profonde, schwere, schweren, starke, большим, شديد, ả, ố, 重, 커, 큰
fire	Feuer, Smoke, dust, feu, fire, fires, fogo, fuego, fuoco, humo, incendie, incendio, lửa, smoke, tobacco, النار, آتش, 拂, 炎, 烟, 煙, 熏, 燻, 菸, 雾, 霧	Brand, Feuer, Fire, Fuego, ag, api, fai, feu, fire, fuego, incendio, llama, lửa, огонь, النار, آتش, 火, 炎, 焰	##selen, ##サス, Fire, Fuego, Riot, bão, earthquake, fire, fuego, hurricane, incendio, lửa, neste, storm, terremoto, trận, tsunami, النار, 炎, 焰
mother	Mama, Mother, Mutter, ibu, ibunya, madre, majka, mama, mater, moeder, mother, mãe, mère, mẹ, nurse, μητέρα, Мать, майка, матери, мати, мать, мајка, مادر, ام, أم, मा, 媽, 嬷, 嬤, 母, 어머니	Mama, Mother, Mutter, grandmother, ibunya, madre, majka, mama, mother, mãe, mère, nang, tante, μητέρα, Мать, майка, матери, мать, мајка, مادر, 亲, 妈, 婆, 媽, 嬷, 嬤, 母, 어머니	Father, Maker, Mama, Mother, Robin, mother, Мать, баща, майка, матери, мать, مادر, 娃, 母, 父, 어머니
live	Lebens, Live, alive, hidup, leben, lebenden, lebt, lebte, lebten, levde, levende, live, lived, lives, living, sống, viva, vivant, vive, viven, vivent, vivió, vivo, vivos, vivre, vécu, žive, živi, жив, живее, живи, زنده, حي, تعيش, ライブ	Lebens, Live, alive, asui, elää, habita, hidup, leben, lebenden, lebt, lebte, lebten, live, lived, living, rege, sağ, sống, vital, viva, vivant, vive, viven, vivent, vivere, vivien, vivir, vivió, vivo, vivos, vivre, wonen, Žije, žije, žive, živi, жив, живе, живее, живеят, живи, жил, жили, существование, يعيش, حي, تعيش, חיים, ライブ, 寿, 活	elää, errichten, existieren, existir, ganado, hidup, leben, lebten, live, living, pequeños, reside, resides, residing, sống, vivant, vive, viven, vivent, vivir, vivo, vivos, vivre, yaşayan, život, живеят, живи, существование, تعيش, يعيش, 住
at	at,au,auf,aus,by,dans,en,for,für,in,into,na,nach,on,onto,over,por,på,sobre,sur,to,trên,tại,upon,vào,với,with,à,în,από,για,εν,επί,κατά,με,σε,στα,στη,στην,στις,στο,στον,στους,в,во,във,за,из,к,към,на,над,по,с,со,у,در,برای,با,ใน,ở,で,に,を,在,에	##ที่,aikaa,an,at,kepada,nơi,terhadap,untuk,έτος,κατά,στη,στις,στο,στον,στους,близ,بر,ओर,সাল,ఏడాది,వద్ద,సంవత్సరం,จาก,ณ,ที่,เมื่อ,แห่ง,ใน,デ,在,에,에서	at,atas,dalam,dans,dessus,in,inside,ninu,trên,tại,untuk,upon,varten,κατά,προς,σε,στα,στη,στην,στις,στο,στον,στους,близ,в,във,на,свыше,согласно,بر,จาก,ณ,เมื่อ,ใน,にかけて,まで,在,에,에서,위에,으로
80	63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 91, 92, 93, 94, 95, 96, 97, 98, 99	65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 91, 92, 93, 94, 95, 96, 97, 98, 99	66, 67, 68, 69, 70, 71, 72, 73, 74, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 91, 92, 93, 94, 95, 96, 97, 98, 99
:	; ,: ,, ,) ,。 ,、 ,؛ ,، ,; ,: ,/ ,. ,- , , ,)	; ,: ,, ,) ,、 ,؛ ,، ,; ,: , , ,)	##ः, ##ঃ, :, ;, Her, Hän, ؛, ！, :, ;

Figure 13: More examples of the grouped subwords on mBERT with CLSA