Tokenization and Morphology in Multilingual Language Models: A Comparative Analysis of mT5 and ByT5

Thao Anh Dang
Utrecht University
Radboud University
t.t.a.dang@uu.nl

Limor Raviv
Max Planck Institute
for Psycholinguistics
limor.raviv@mpi.nl

Lukas Galke
University of Southern Denmark
galke@imada.sdu.dk

Abstract

Morphology is a crucial factor for multilingual language modeling as it poses direct challenges for tokenization. Here, we seek to understand how tokenization influences the morphological knowledge encoded in multilingual language models. Specifically, we capture the impact of tokenization by contrasting a minimal pair of multilingual language models: mT5 and ByT5. The two models share the same architecture, training objective, and training data and only differ in their tokenization strategies: subword tokenization vs. character-level tokenization. Probing the morphological knowledge encoded in these models on four tasks and 17 languages, our analyses show that the models learn the morphological systems of some languages better than others and that morphological information is encoded in the middle and late layers. Finally, we show that languages with more irregularities benefit more from having a higher share of the pre-training data.

1 Introduction

Tokenization, the process of segmenting a text into individual units, plays a special role in language modeling as it is disconnected from the otherwise end-to-end training procedure (Xue et al., 2021, 2022; Sennrich et al., 2016). Languages differ in their morphological structure (Goldman and Tsarfaty, 2022; Dryer and Haspelmath, 2013; Evans and Levinson, 2009; Ackerman and Malouf, 2013) and it has been shown that morphologically more complex languages are harder to acquire by humans (DeKeyser, 2005; Raviv et al., 2021; Kempe and Brooks, 2008) and deep neural network models (Galke et al., 2024; Park et al., 2021; Mielke et al., 2019; Cotterell et al., 2018). Here, we seek to understand to what extent different tokenization strategies influence the ability of multilingual language models to capture morphological knowledge in different languages (See Figure 1).

Ideally, a language model would be equally proficient in a variety of languages (Lample and Conneau, 2019; Conneau et al., 2020; Ruder et al., 2019). Understanding the influence of tokenization is crucial in the context of multilingual language modeling (Xue et al., 2021, 2022; Warstadt et al., 2020), as it is challenging to find a set of tokens that is equally good for modeling all the languages in the world. Beyond the proportions of languages in the training data, it is important to take into account the morphological structure of the different languages (Anh et al., 2024; Galke et al., 2024; Cotterell et al., 2018). Importantly this needs to be already considered when selecting the data for learning the tokenizer – even before language model pre-training – as this influences what subword structures end up as the tokens to be processed by the language model. The issue of tokenization and, in related matter, how to mix different languages, are particularly relevant in the current era of large language models (Touvron et al., 2023; Brown et al., 2020; Bubeck et al., 2023; Wei et al., 2022), when aiming for similar performance on a diverse range of languages (Le Scao et al., 2023).

With models such as ByT5 (Xue et al., 2022), a character-level language model based on the T5 architecture (Raffel et al., 2020), it has been shown that tokenizer-free language models yield commensurate downstream performance with their tokenizer-based counterparts (Xue et al., 2022; Edman et al., 2024), such as mT5 (Xue et al., 2021), another T5-based model that trained on the exact same data as ByT5. Specifically, in machine translation, Edman et al. (2024) have found that character-level ByT5 yields similar performance as mT5 when allowing more training to recover word-level structures. Yet, the interplay of morphology and tokenization is so far poorly understood.

Here, we seek to dissect the root of these findings through analyzing the effect of the tokenization strategy (character-level vs. subword-level) on the

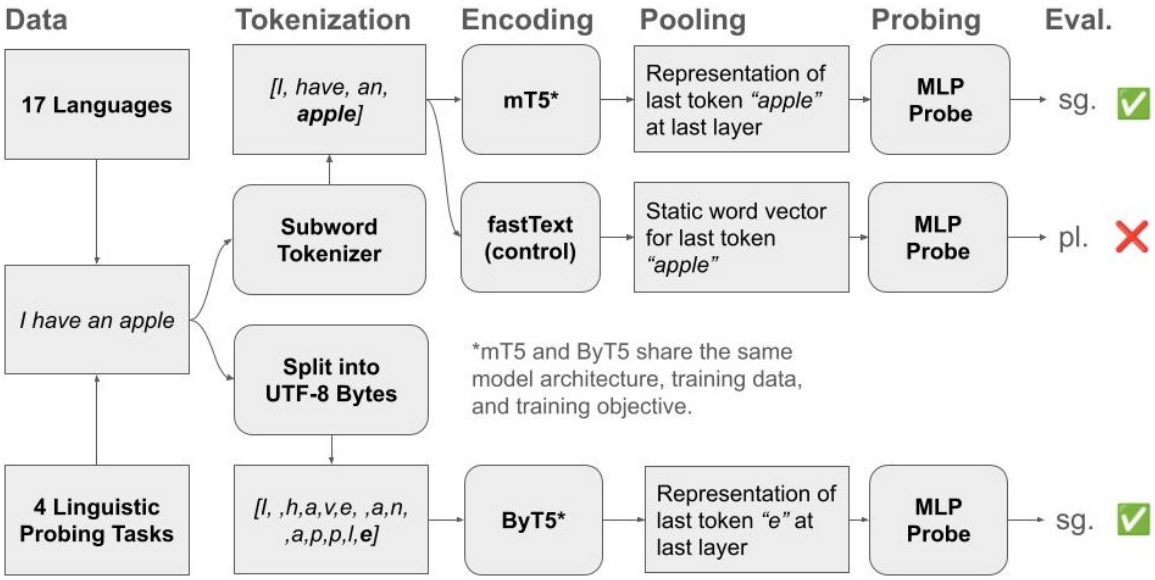

Figure 1: Overview of our experimental procedure

morphological knowledge encoded in contextualized representations of multilingual language models. Specifically, we contrast ByT5 as a character-level multilingual language model and mT5 as a subword-level multilingual language model, with both models sharing the same architecture and being trained on the same data. We use well-established structural probing techniques to capture the amount of linguistic information encoded in the contextualized word representations of the language models (Rogers et al., 2021; Manning et al., 2020; Belinkov et al., 2020; Tenney et al., 2019). Focusing on morphology, we probe the contextualized representations of ByT5 and mT5 for morphological knowledge on 17 languages from a large-scale multilingual dataset (Acs et al., 2023). In addition, we use the non-contextualized fastText model (Bojanowski et al., 2017) as a control to understand to what extent the context is important for multilingual language models reflecting morphosyntactic structures in their representations. We further explore to what extent the captured morphological knowledge depends on the share that the languages have in the pre-training data, various linguistic factors, such as the type of task (number, tense, case, gender), and the language's degree of morphological complexity, as quantified by the degree of irregularity (Wu et al., 2019) and type-to-token ratio (TTR) (Bentz et al., 2015).

By systematically contrasting tokenizer-free ByT5 and subword-tokenized mT5, two pre-trained multilingual language models based on the T5 architecture, and trained on the same data, we find:

- Multilingual language models learn the morphological systems of some languages better than others.
- Morphological knowledge representation improves over transformer layers.
- Both subword- and byte-level models display approximate the same level morphological knowledge encoded in the activations after a small number of transformer layers.
- A language's degree of irregularity plays a substantial role for capturing morphological knowledge, suggesting that more irregular languages would benefit from a higher proportion of training data.

2 Background and Related Work

Morphology Morphology concerns how meaningful word units can be combined to express a range of grammatical information (Bloomfield, 1933). Languages differ greatly in their morphological systems and degrees of morphological complexity (Dryer and Haspelmath, 2013; Lupyan and Dale, 2010; Bloomfield, 1933). It has been shown that morphologically more complex languages are harder to acquire by humans (DeKeyser, 2005; Raviv et al., 2021; Kempe and Brooks, 2008) and deep neural networks (Galke et al., 2024; Cotterell et al., 2018). Some studies specifically investigate the link between morphological complexity and the challenges in language modeling (Cotterell et al., 2018; Mielke et al., 2019; Park et al., 2021; Galke

et al., 2024; Anh et al., 2024). However, the findings have been mixed. As most of these studies focus more on the overall learnability (Park et al., 2021; Gerz et al., 2018), there may be confounds other than morphological complexity, which we isolate here with a fine-grained analysis.

Linguistic Probing Linguistic probing can be categorized into behavioral probing and structural probing (Madsen et al., 2022). Behavioral probing aims to understand how a language model behaves in a specific or new setting (Hupkes et al., 2023). Structural probing instead seeks to localize where linguistic abilities are encoded (Rogers et al., 2021). It was suggested that morphological knowledge is mainly encoded at the lower layers (Peters et al., 2018; Belinkov et al., 2020). However, most studies focus on analyzing monolingual language models, such as BERT (Devlin et al., 2019), or neural machine translation systems trained on pairs of languages (Acs et al., 2023; Belinkov et al., 2020; Bisazza and Tump, 2018; Edmiston, 2020), while studies on multilingual language models are rare.

Tokenization Tokenization is a crucial step in language modeling, especially for multilingual language models. The dominant tokenization methods for language models are based on the Byte-Pair Encoding (BPE) algorithm (Sennrich et al., 2016), and rely on the data mix to ensure coverage of different languages (Le Scao et al., 2023). Starting with single characters, BPE iteratively merges tokens based on co-occurrence statistics, allowing for both subword and multi-word tokens. Comparing BPE with other subword tokenizers, Ali et al. (2024) found that the preference for tokenization methods differs across languages. While BPE works better for Germanic languages, such as German and English, Unigram (Kudo, 2018) is more well-suited for Romance languages, such as Spanish. The importance of tokenization in multilingual language modeling is evident (Hofmann et al., 2021; Toporkov and Agerri, 2024), as many studies propose new tokenization algorithms to make language models capture the morphological structure of the different languages (Goldman and Tsarfaty, 2022).

Character-level Language Models A line of research investigates character-level language models to circumvent the issues of multi-lingual tokenization (Fleshman and Van Durme, 2023; Clark et al., 2022; Xue et al., 2022; Gao et al., 2020; Chung et al., 2016; Lee et al., 2017; Kim et al., 2016). In machine translation, Lee et al. (2017) found that character-level tokenizers perform as well as or better than models based on sub-word tokenization. They highlighted that character-level tokenization offers better translation quality in multilingual and low-resource settings because of the shared vocabulary. Edman et al. (2024) argued that character-level models are better at learning information that operates at a low level of granularity, such as morphology. Comparing translation capability between multilingual language models with standard tokenization (mT5) and character-level model (ByT5), they found that ByT5 outperforms mT5 in several aspects. First, ByT5 produces higher-quality translations than mT5, even in the case of low-resource languages. ByT5 is also better in handling rare and similar words. In addition, low-resource languages may benefit from shared vocabulary, namely the set of characters (Gao et al., 2020). However, it has also been suggested that the effect varies across languages (Ali et al., 2024), which in-turn motivates the present study.

Summary Overall, earlier work on probing the morphological knowledge of language models is mainly conducted on machine translation and recurrent neural networks (Belinkov et al., 2020; Vylomova et al., 2017). More recently, the focus has been extended to Transformer-based language models (Acs et al., 2023; Edmiston, 2020; Wu and Dredze, 2020), yet the majority of studies investigated BERT and its variants (Rogers et al., 2021). Previous studies provide mixed findings about the morphological knowledge of language models and only very few multilingual studies. Here, we complement the literature by analyzing how captured morphological knowledge is influenced by the tokenization strategy, the languages' proportions in the model's training data, the language's morphological complexity, and the task type.

3 Models and Data

We compare two pre-trained language models: mT5 and ByT5. The two models share the same architecture and are trained on the same data with the same training objective (masked span prediction). The key difference between mT5 and ByT5 is their tokenization strategy: mT5 uses a standard subword tokenization strategy, whereas ByT5 operates on character level. Thus, we can investigate the effect of tokenization.

mT5 The mT5 model (Xue et al., 2021) is a multilingual version of T5, an encoder-decoder language model (Raffel et al., 2020). It is trained on the mC4 corpus (Xue et al., 2021), which consists of text in 101 languages compiled from the Common Crawl web scrape. Like T5, mT5 employs a masked span prediction training objective and the SentencePiece tokenizer (Kudo and Richardson, 2018), a variant of the BPE tokenizer by Sennrich et al. (2016). The vocabulary comprises approximately 250,000 subwords, covering 104 languages (Xue et al., 2021) while sharing (sub-)words between languages.

ByT5 ByT5 (Xue et al., 2022) is a tokenizer-free variant of mT5, inheriting most of the properties of mT5, using the same T5 model architecture, and the same masked span prediction training objective. The only crucial difference between them is the tokenization method: While mT5 uses a standard tokenizer, ByT5 operates on single-character tokens, or more precise: UTF-8 encoded bytes. Another difference is that ByT5's encoder stack consists of three times more layers than the decoder to process a larger number of bytes. For the masked span prediction objective, ByT5 uses a span of 20 bytes. The similarity in architecture and sizes of mT5 and ByT5 enables our comparison of how the tokenization strategy impacts the morphological knowledge encoded in the learned representations.

Dataset We use the multilingual morphological probing dataset by Acs et al. (2023). The dataset consists of 247 probing tasks, available in 42 languages and 10 language families. It is built upon the Universal Dependencies tree bank and covers both frequent and infrequent words. In each language, each task includes a training set of 2000 examples, a development test of 200 examples, and a test set of 200 examples. We selected 16 out of all 42 languages which had at least two tasks available. However, we also included Arabic despite having only one task available to better cover the Semitic language family. In total, our considered dataset consists of 17 languages and 43 morphological probing tasks, covering number, case, gender, and tense – focusing on nominal and verbal inflection.

4 Methodology

Feature Extraction For training, we extracted the contextualized embeddings of the words in the training set for each task in each language and each model. Both mT5 and ByT5 are available in different sizes. We chose to test the `mT5-base` model. We froze the weights and extracted the hidden states for the entire sentence before extracting the word embedding corresponding to the target. Since we also aim to look at how much morphological knowledge is learned at each layer, we extracted the word embedding at each hidden layer of the network, including the input embedding layer. Each word representation is associated with a label, which corresponds to the respective morphological feature from the task. We trained separate probes for each language and each task in each layer of the model.

Probing Classifiers Previous studies often use two architectures for probing classifiers, namely linear classifiers (Hupkes et al., 2018; Belinkov et al., 2020) and multilayer perceptrons (MLPs) (Lin et al., 2019; Conneau et al., 2018; Adi et al., 2017; Ettinger et al., 2018; Zhang and Bowman, 2018). Both types of probes have received convincing arguments. Linear probes capture information that is linearly separable in the representations (Liu et al., 2019; Belinkov et al., 2020), whereas MLP probes can additionally capture nonlinear patterns in the representations (Hewitt and Liang, 2019). Studies show that linear classifiers and MLPs produce similar accuracy (Conneau et al., 2018; Belinkov et al., 2017a; Qian et al., 2016). We have considered both types of probes for this study (see Appendix C). For the main results, we employ MLP probes.

Subword Pooling In both mT5 and ByT5, words are segmented into either subword units or characters. As such, when passing through the hidden layers, each subword or character has its own embedding. There are several methods to then approximate the embedding for an entire word: The first method is to take the weighted **average** of the embeddings of all components. The second way is to consider the embedding of the **last** subword or character as the representation for the entire word. Both methods have limitations. Averaging the token embeddings may cancel out some information, while the last embedding may not contain all the information about the entire word. Belinkov et al. (2020) compared both methods and found that using the embeddings of the last token produced higher accuracy scores. We have considered both options (see Appendix D) and can confirm that last-token pooling leads to higher probing accuracy. For the main results, we employ last-token pooling.

Evaluation and Control We aim to probe the morphological knowledge of a range of typologically diverse languages. While the dataset of Acs et al. (2023) supports 42 languages, in some languages, there is a large gap between the number of tasks in each language. While Russian has 12 tasks, Polish and Armenian have only one task. To ensure a fair comparison of the learned morphological representations, we selected languages with at least two tasks along with Arabic to cover the Semitic language family. We focused exclusively on the morphological properties of words, excluding agreement tasks. In total, we ran 43 morphological probing tasks for 17 languages. Appendix E provides the details of the morphological properties of the 17 considered languages, their morphological complexity scores, and their proportion in the ByT5/mT5 training data. Appendix F provides a detailed description of the types of probing tasks, covering number, case, gender, and tense.

As non-contextualized control, we employ fastText word embeddings (Bojanowski et al., 2017) which are available in 157 languages. We probed fastText embeddings using the exact same procedure and evaluation metric. We obtained the static word embeddings without any pooling. Contrasting the contextualized representations of mT5 and ByT5 with fastText allows us to quantify to what extend the contextualized models make use the sentence context to tackle the morphological tasks.

5 Results

Overall Probing Accuracy We first looked at the overall probing performance of mT5, ByT5, and fastText as well as the differences between the two Transformer-based models compared to the fastText baseline. For all analyses except for the layer-wise analysis, we used the probing accuracy of the last hidden layer. To obtain the overall performance of mT5 and ByT5, we averaged over all languages and tasks, resulting in a single accuracy score for each model (see Table 1). Full results for each task and language are provided in Appendix G.

On the surface, it appears that mT5 and ByT5 have comparable performance and both models outperformed fastText. ByT5 slightly surpassed mT5, yet this difference is very small. This finding is different from that of Belinkov et al. (2017a), who found that character-level tokenizers are better than subword tokenizers in representing morphology.

Model	Mean Probing Accuracy
mT5-base	82.57
ByT5-base	**82.86**
fastText (baseline)	77.52

Table 1: Probing accuracy of mT5, ByT5 and fastText, averaged over languages and tasks

Table 2 shows the difference between accuracy scores of mT5 and ByT5, grouped by language. It can be seen that accuracy scores are not equal across both languages and tasks. Comparing mT5 and ByT5, it seems that they perform on par with each other in most languages. However, mT5 scores higher in Turkish and ByT5 achieves much higher on Hindi tasks. The difference between contextualized language models and non-contextualized fastText also tells to what extent contextual information affects morphological abilities. From the results, we observed that for most languages, mT5 and ByT5 achieved considerably higher probing accuracy than the baseline. The largest difference is observed in the case of Basque ($> 50\%$). This implies that contextual word embeddings capture morphological knowledge better than static embeddings for these languages. In contrast, the results are lower than the non-contextual baseline in French and Russian. Contextual information seems to make accessing morphological information more difficult in these two languages.

Considering the differences between both language models and the fastText baseline, as shown in Table 2, it can be seen that they generally outperformed the baseline yet perform substantially worse than baseline in French and Russian. Averaging over tasks, ByT5 yields the highest probing accuracy in 7 languages, while mT5 yields the highest probing accuracy on 6 tasks out of 17 languages.

Table 3 shows the probing accuracy scores for each language, averaged over tasks. It appears that there are some differences in accuracy across languages, with the hardest language being Russian for mT5 and Arabic for ByT5. The accuracy scores of some languages are higher than the others. Unsurprisingly, the models perform best on the English language, followed by Hebrew, Portuguese, and Romanian. The models learn the morphological systems of German, French, Estonian, and Latvian moderately well. Arabic and Russian achieved lowest accuracy scores. However, Arabic results should be interpreted with caution as there

Language	Family	Model		
		mT5	ByT5	fastText
English	Germanic	98.00	**98.50**	97.75
Dutch	Germanic	**93.50**	92.75	82.25
German	Germanic	68.46	**72.40**	68.87
French	Romance	71.93	77.53	**92.95**
Spanish	Romance	93.83	**94.00**	71.16
Portuguese	Romance	95.00	**96.25**	88.16
Romanian	Romance	94.75	**95.25**	92.25
Hebrew	Semitic	97.50	**98.00**	92.49
Arabic	Semitic	**66.17**	49.25	37.81
Russian	Slavic	61.26	57.43	**79.20**
Czech	Slavic	**88.79**	87.79	78.81
Hindi	Indic	67.83	**87.50**	58.02
Urdu	Indic	**88.50**	84.50	74.33
Turkish	Turkic	**94.78**	83.69	78.28
Latvian	Baltic	**77.14**	72.01	73.76
Estonian	Uralic	**91.08**	90.03	82.94
Basque	Basque	91.69	**91.91**	52.60

Table 2: Probing accuracy of mT5, ByT5, and fastText by languages, language families, averaged over tasks. The best score per language is marked in bold font.

Task	mT5	ByT5	fastText
Number	**93.75**	93.56	88.15
Tense	80.30	72.44	**80.89**
Gender	76.85	**85.16**	77.61
Case	**71.53**	69.16	55.49

Table 3: Probing accuracy of mT5, ByT5, and fastText, averaged over languages and tasks

is only one task available (i.e., case).

Layer-wise Analysis Figure 2 illustrates the difference in probing accuracy between languages and between tasks for mT5 and ByT5. It can be seen that there are some degrees of variation between layers. In languages that show high overall performance, namely English, Dutch, Portuguese, Spanish, Basque, and Hebrew, probing accuracy shows very little improvement over layers. In other languages, accuracy increases, reaches its peak at the middle and slightly decreases at late layers in other languages. This finding is partly consistent with Acs et al. (2023), Edmiston (2020), and Hewitt et al. (2021), who also reported best performance in the middle to late layers. However, we further show that this is not true for all languages. There are cases where morphological knowledge is successfully learned in the early layer and carried on throughout the network. Our results contrast with Belinkov et al. (2020) and Peters et al. (2018), who found that morphological information is best encoded in the first layer of the model, and then has the tendency to decrease over time.

Comparing mT5 and ByT5, there are a few noticeable differences. In the plots for ByT5, accuracy scores of each language and each task improves considerably after the embedding layer. This trend is less visible in mT5, although performance does improve over layers. Morphology is better learned in the embedding layer of mT5 than that of ByT5. This may imply that character-level language models need more layers to capture morphological patterns of languages.

Effect of Task Type To investigate whether morphological features are learned differently by mT5 and ByT5, we averaged the scores for each task across all languages, resulting in a single score for each task (see Table 3). The results strongly suggest that each morphological feature is encoded differently. Interestingly, mT5 and ByT5 show different patterns. Both models perform equally well at number and worst at case. However, tense is learned better than gender by mT5 while the opposite is true for ByT5. Comparing both language models with the baseline, they surpass the baseline in all tasks except for tense, where ByT5 performs considerably worse than fastText.

Case seems to be the hardest task for both models. Besides the case task often having more possible classes than other features, case is also more context-dependent than other features. Case marking is used to indicate the syntactic function of the word in the sentence. As such, one word may have different cases in different contexts and thus is inflected distinctively. Gender is also relatively difficult, especially for mT5. However, looking at individual languages (see Table 2), the mean score of mT5 is affected by Hindi and Latvian, whose scores are exceptionally lower than the baseline (less than 25%). Except for those two languages, mT5 and ByT5 perform equally well.

Morphological Complexity and Training Data We explore the effects of two types of morphological complexity, namely TTR (Bentz et al., 2015) and the degree of irregularity (Wu et al., 2019) on probing accuracy. Complexity values per language can be found in Appendix E. We hypothesize that probing accuracy is influenced by the proportion of the respective language in the training data, which then also modulates the effect of a language's mor-

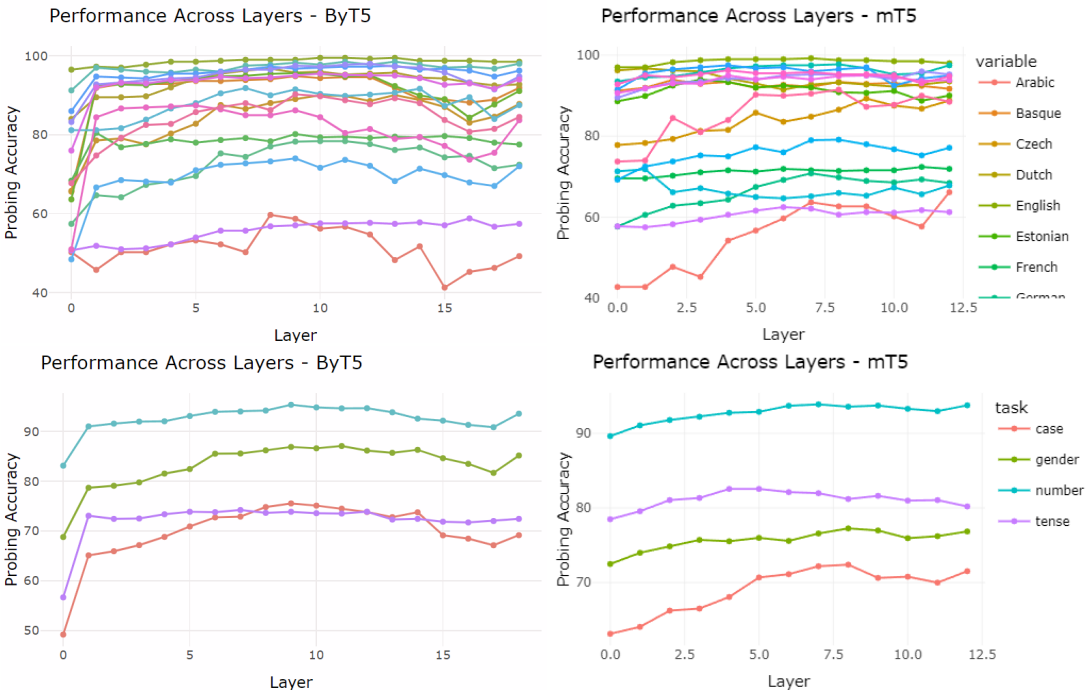

Figure 2: Probing accuracy of ByT5 (left) and mT5 (right) across layers grouped by languages and tasks. Each line represents a language (top) or a task (bottom). Each data point is the accuracy scores at each layer of each language.

phological complexity. To understand the effect of these factors and their interaction, we fitted a generalized mixed effect logistic regression model predicting accuracy from the two morphological complexity measures and the proportion of training data. There are 200 data points for each task in each language. The analysis was conducted with the R-package `lme4`. All variables were scaled and centered. Random intercepts were added for language and task.

Our results show significant positive effects of training data size and irregularity on probing accuracy, as well as their interaction. In more detail, there is a strong effect of the amount of training data and a language's degree of irregularity on probing accuracy (training data: $\beta = 2.11$, $SE = .55$, $p < .001$, irregularity: $\beta = 2.83$, $SE = .42$, $p < .001$). The effects of training data and irregularity were highly correlated (0.831). In addition, there is an interaction effect between the language's irregularity and its training data in the model ($\beta = 2.03$, $SE = .31$, $p < .001$). The effect of training data size on probing accuracy is stronger when there is more irregularity in the language. This means that high irregularity in the morphological systems amplifies the impact of training data on the morphological abilities of language models. Detailed results of the statistical models can be found in Appendix H.

6 Discussion

Languages' morphology is learned differently Our probing results in mT5 and ByT5 show that the morphological knowledge of some languages is better represented than the others. Some languages (e.g., English, Dutch) achieved nearly perfect accuracy in probing tasks (higher than 90%). However, both mT5 and ByT5 performed worse at German, French, Russian, and Arabic tasks. These results to some extent contradict Edmiston (2020), who found comparable performance in all languages. Acs et al. (2023) also did not observe performance differences across languages, but only between part-of-speech and morphological features.

Differences across tasks We observed that some tasks are more difficult for the language models: Number is the easiest task whereas case is the hardest one. This difference can be partly attributed to the higher number of possible categories but also to the context-dependent nature of case. This is supported by the non-contextualized baseline results for case, which are substantially lower than both mT5 and ByT5. These findings are in agreement with earlier findings by Bisazza and Tump (2018) and Edmiston (2020). Why tense and also case features are particularly challenging to find in the representations of character-based language models is an interesting question for future research.

Character-level models are on par with subword level models in representing morphology We found that both models yield highly similar average performance on the probing tasks. Remarkably, despite being tokenized at the byte level, ByT5 is able to reconstruct morphological knowledge already in the first transformer block, arriving at a similar level as mT5. Belinkov et al. (2020) and Vylomova et al. (2017) tested machine translation systems and found character-level tokenizers to surpass BPE in learning morphology. However, through investigating large-scale language models, we found no advantage of character-level tokenizer in encoding morphology. Instead, we found that the preferable model differs between individual languages, e.g., subword tokenization works better for Turkish, whereas character-level tokenization benefits Hindi morphology. fastText performed nearly as well as mT5 and ByT5. We attribute this to the word-level nature of morphological features. Moreover, we used separated fastText models for each language, instead of generic multilingual ones.

Morphology is best represented in middle to late layers Our findings show that morphological knowledge generally improves over layers in both language models. There are languages which have high performance across layers. Yet, ByT5 shows greater improvement after the embedding layer than mT5. We also observed that morphological information is best encoded in the middle to late layers of the models in some languages. This finding supports Acs et al. (2023), Hewitt et al. (2021), and Edmiston (2020). Our findings differ from Belinkov et al. (2017b); Tenney et al. (2019); Peters et al. (2018), who found that morphology is a low-level feature and is encoded along with word identity in the first layer of the network.

Morphological irregularity amplifies the effect of training data Considering the relationship between morphological knowledge encoded in language models and the languages' morphological complexity, our analysis reveals effects of morphological irregularity and training data sizes on the performance of probing classifiers, in a way that the effect of irregularity is mediated by training data. When a language is highly irregular, a larger share of the training data is beneficial to fully capture its morphological system. Previous studies on the effect of training data sizes show that its effect is not present at the representation level yet at the downstream level (Warstadt et al., 2020; Zhang and Bowman, 2018). We have shown here that the importance of the relative training data size in multilingual language modeling can be already found when probing for morphological knowledge. Considering the interplay with morphological complexity, Mielke et al. (2019) and Gerz et al. (2018) correlated modeling difficulty with morphological counting complexity (Sagot, 2013), vocabulary sizes of languages, and dependency length – and found vocabulary size to be the most important factor. Our study complements those findings by establishing that the degree of irregularity plays a substantial role for what morphological properties are captured by a language model – and that this factor amplifies the effect of the language's share of the pre-training data.

It may seem unexpected that a higher degree of irregularity has a positive effect on probing accuracy. However, a possible explanation is that irregular forms are better memorized *because* they appear more often in the training data, given the correlation of irregularity with frequency (Wu et al., 2019). This could also be linked to the word predictability advantage of Zipfian distributions that has been shown to aid word segmentation in humans (Lavi-Rotbain and Arnon, 2022) – which we deem an interesting direction for future work.

Limitations and ethical considerations can be found in Appendices A and B, respectively.

7 Conclusion

We have analyzed the effects of tokenization, training data proportions, and linguistic factors on morphological knowledge encoded in the parameters of pre-trained multilingual language models. Through analyzing 17 languages and 4 morphological tasks, we have shown that the morphological knowledge encoded in multilingual language models differs across languages, despite the global average scores being similar. Beyond differences across languages, we also found differences across tasks, showing that tense and case are particularly hard to find in the representations of character-based language models. To further understand what exactly influences those difference, we have analyzed the effect of morphological complexity in relation to the language's proportion in the language model's pre-training data. We found that the degree of irregularity plays a significant role and amplifies the effect of training data, suggesting that more irregular languages benefit from a having a higher share in the data mix used for pre-training.

References

Farrell Ackerman and Robert Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language*, 89(3):429–464.

Judit Acs, Endre Hamerlik, Roy Schwartz, Noah A Smith, and Andras Kornai. 2023. Morphosyntactic probing of multilingual bert models. *Natural Language Engineering*, pages 1–40.

Judit Ács, Ákos Kádár, and Andras Kornai. 2021. Subword pooling makes a difference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2284–2295, Online. Association for Computational Linguistics.

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *International Conference on Learning Representations*.

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.

Dang Anh, Limor Raviv, and Lukas Galke. 2024. Morphology matters: Probing the cross-linguistic morphological generalization abilities of large language models through a Wug test. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 177–188, Bangkok, Thailand. Association for Computational Linguistics.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017a. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2020. On the linguistic representational power of neural machine translation models. *Computational Linguistics*, 46(1):1–52.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017b. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Christian Bentz, Annemarie Verkerk, Douwe Kiela, Felix Hill, and Paula Buttery. 2015. Adaptive communication: Languages with more non-native speakers tend to have fewer word forms. *PloS one*, 10(6):e0128254.

Arianna Bisazza and Clara Tump. 2018. The lazy encoder: A fine-grained analysis of the role of morphology in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2871–2876, Brussels, Belgium. Association for Computational Linguistics.

Leonard Bloomfield. 1933. *Language*. H. Holt.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. 2016. A character-level decoder without explicit segmentation for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1693–1703, Berlin, Germany. Association for Computational Linguistics.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

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, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing

sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Ryan Cotterell, Sabrina J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 536–541, New Orleans, Louisiana. Association for Computational Linguistics.

Robert M DeKeyser. 2005. What makes learning second-language grammar difficult? A review of issues. *Language learning*, 55.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186.

Matthew S Dryer and Martin Haspelmath. 2013. WALS Online (v2020. 3). *Zenodo*, 7385533.

Lukas Edman, Gabriele Sarti, Antonio Toral, Gertjan van Noord, and Arianna Bisazza. 2024. Are character-level translations worth the wait? Comparing ByT5 and mT5 for machine translation. *Transactions of the Association for Computational Linguistics*, 12:392–410.

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

Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. 2018. Assessing composition in sentence vector representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1790–1801, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Nicholas Evans and Stephen C Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and brain sciences*, 32(5):429–448.

William Fleshman and Benjamin Van Durme. 2023. Toucan: Token-aware character level language modeling. *arXiv preprint arXiv:2311.08620*.

Lukas Galke, Yoav Ram, and Limor Raviv. 2024. Deep neural networks and humans both benefit from compositional language structure. *Nature Communications*, 15(10816).

Yingqiang Gao, Nikola I. Nikolov, Yuhuang Hu, and Richard H.R. Hahnloser. 2020. Character-level translation with self-attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1591–1604, Online. Association for Computational Linguistics.

Daniela Gerz, Ivan Vulić, Edoardo Ponti, Jason Naradowsky, Roi Reichart, and Anna Korhonen. 2018. Language modeling for morphologically rich languages: Character-aware modeling for word-level prediction. *Transactions of the Association for Computational Linguistics*, 6:451–465.

Omer Goldman and Reut Tsarfaty. 2022. Morphology without borders: Clause-level morphology. *Transactions of the Association for Computational Linguistics*, 10:1455–1472.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher Manning. 2021. Conditional probing: Measuring usable information beyond a baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1626–1639, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre is not superb: Derivational morphology improves BERT's interpretation of complex words. 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 3594–3608, Online. Association for Computational Linguistics.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, Dennis Ulmer, Florian Schottmann, Khuyagbaatar Batsuren, Kaiser Sun, Koustuv Sinha, Leila Khalatbari, Maria Ryskina, Rita Frieske, Ryan Cotterell, and Zhijing Jin. 2023. A taxonomy and review of generalization research in NLP. *Nature Machine Intelligence*, 5(10):1161–1174.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Vera Kempe and Patricia J Brooks. 2008. Second language learning of complex inflectional systems. *Language Learning*, 58(4):703–746.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander Rush. 2016. Character-aware neural language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple

subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

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.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Ori Lavi-Rotbain and Inbal Arnon. 2022. The learnability consequences of zipfian distributions in language. *Cognition*, 223:105038.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside bert's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of NAACL-HLT*, pages 1073–1094.

Gary Lupyan and Rick Dale. 2010. Language Structure Is Partly Determined by Social Structure. *PLoS ONE*, 5(1):e8559.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022. Post-hoc interpretability for neural nlp: A survey. *ACM Computing Surveys*, 55(8):1–42.

Christopher D Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

Sabrina J Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What kind of language is hard to language-model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989.

Hyunji Hayley Park, Katherine J Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. 2021. Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276.

Matthew E Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509.

Peng Qian, Xipeng Qiu, and Xuan-Jing Huang. 2016. Investigating language universal and specific properties in word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1478–1488.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Limor Raviv, Marianne de Heer Kloots, and Antje Meyer. 2021. What makes a language easy to learn? a preregistered study on how systematic structure and community size affect language learnability. *Cognition*, 210:104620.

Limor Raviv, Louise R Peckre, and Cedric Boeckx. 2022. What is simple is actually quite complex: A critical note on terminology in the domain of language and communication. *Journal of Comparative Psychology*, 136(4):215.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Benoît Sagot. 2013. Comparing complexity measures. In *Computational approaches to morphological complexity*.

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

Anders Søgaard, Anders Johannsen, Barbara Plank, Dirk Hovy, and Hector Martínez Alonso. 2014. What's in a p-value in nlp? In *Proceedings of the eighteenth conference on computational natural language learning*, pages 1–10.

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

Olia Toporkov and Rodrigo Agerri. 2024. On the role of morphological information for contextual lemmatization. *Computational Linguistics*, pages 1–35.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ekaterina Vylomova, Trevor Cohn, Xuanli He, and Gholamreza Haffari. 2017. Word representation models for morphologically rich languages in neural machine translation. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 103–108.

Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel Bowman. 2020. 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.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Shijie Wu, Ryan Cotterell, and Timothy O'Donnell. 2019. Morphological irregularity correlates with frequency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5117–5126.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Kelly Zhang and Samuel Bowman. 2018. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 359–361.

A Limitations

Several limitations should be taken into account: First, our experiments do not take orthographic transparency into account, as all models and complexity measures are based on written language. Second, we ran each experiment only once due to the high volume of experiments. Third, while we aimed to cover as many typologically different languages as possible, the dataset we use supports mostly Indo-European languages, such that 12 out of our 17 considered languages are Indo-European. Moreover, the current study only focuses on inflectional morphology due to the lack of datasets for probing derivational morphology. Lastly, p-values should be treated with care when the sample size (here: 17,200) is large (Søgaard et al., 2014).

B Ethical Considerations

We emphasize that morphological complexity of languages bears no implication on their quality – having more complexity does not make one language better than another (see Raviv et al., 2022).

C Effect of Probing Architecture

Previous studies on probing linguistic features have had a debate over which type of probe is sufficient to extract the relevant knowledge, but insufficient to learn the knowledge itself (Belinkov, 2022). Here, we also compare the accuracy scores of MLP probes and linear probes for mT5 and ByT5, as shown in Table 4.

Model	**Probing Architecture**	
	MLP	**Linear**
mT5-base	82.57	80.37
ByT5-base	82.86	80.61

Table 4: Probing accuracy of mT5 and ByT5 when using MLP and linear classifiers, averaged over languages and tasks

We observed no considerable differences in accuracy scores across types of probes. For both types of probes, the mean accuracy scores are all around

80%. This pattern is most inline with Acs et al. (2023), suggesting that linear classifiers are as effective as non-linear ones in extracting morphological knowledge of multilingual LLMs. Moreover, the observation that linear probes perform equally well as MLPs implies that morphology is a relatively simple feature that can be learned early and straightforwardly by the models.

D Effect of Pooling Methods

We compare the probing accuracy when using these two pooling methods, namely the **average** and **last** method. Table 5 shows the results of when using these two methods for mT5 and ByT5.

Model	Subword Pooling Method	
	last	**average**
mT5-base	82.57	75.28
ByT5-base	82.86	76.40

Table 5: Probing accuracy of mT5 and ByT5 when using different subword pooling methods. The results were averaged over languages and tasks and the reported probes are MLPs

Comparing the accuracy scores of the two pooling methods, it can be seen that the *last* method achieved considerably higher accuracy scores than the *average* method. The difference is approximately 6-7 points. This also holds for both models. It seems that the representational information and/or the morphological content of a word is mostly encoded in its last token. Previous comparisons have also reported similar results (Acs et al., 2023; Ács et al., 2021; Belinkov et al., 2020).

E Details of the Considered Languages

E.1 Morphological Properties

Table 6 shows the morphological properties of the considered languages.

E.2 Proportion of mT5/ByT5's training data and morphological complexity of the language

Table 7 shows the language's proportion of training data and their morphological complexity scores: TTR and Irregularity.

Language	Tr. Data	TTR	Irregularity
English	5.67%	-0.460	-5.94
German	3.05%	-0.010	-6.28
Dutch	1.98%	-0.390	-6.68
French	2.89%	-0.340	-4.16
Romanian	1.58%	-0.420	-3.40
Spanish	3.09%	0.001	-8.81
Portuguese	2.36%	0.038	-9.11
Turkish	1.93%	1.550	-5.96
Czech	1.72%	0.430	-5.63
Russian	3.71%	0.870	-7.74
Hebrew	1.06%	2.020	-1.78
Arabic	1.66%	1.630	-0.06
Hindi	1.21%	-0.300	-2.10
Estonian	0.89%	1.760	-2.79
Latvian	0.87%	0.770	-7.90
Urdu	0.61%	-0.450	9.20
Basque	0.57%	1.310	19.86

Table 7: Percentages of training data of mT5 and Byt5 from Xue et al. (2021), irregularity scores from Wu et al. (2019), and TTR scores from Bentz et al. (2015) for each investigated language. For irregularity, higher scores mean being more morphologically irregular. In contrast, higher TTRs mean higher complexity.

F Considered Probing Tasks

Here, we provides an overview the morphological properties that we investigate in the study, namely number, tense, case, and gender, and how they may vary between languages. (Acs et al., 2023)

Number In many languages, especially inflected languages, nouns are marked as either singular or plural (Bloomfield, 1933). An exception is Latvian, which includes singular, plural, and partitive nouns. Plurality is usually expressed by adding certain endings to the nouns, and sometimes include changing their vowels. These endings are determined in different ways across languages. For instance, in some Indo-European languages such as Spanish, the plural form of nouns is affected by their gender. In this task, the LLMs have to predict whether the target word is a plural or singular noun.

Tense Most languages mark tenses (Bloomfield, 1933). In some languages, tenses are indicated by inflecting verbs. In other languages, for example, Estonian, adjectives can also express tense. In certain languages, tense can interact with other morphological features, namely mood and aspect. Inflection patterns for tense are usually dependent

on the ending, conjugation pattern of the verbs, and whether they are regular or irregular. In Spanish and French, it is also dependent on the subject pronouns. In Hindi, verbs indicating tense must agree with gender and number of the subject.

Case A case system is a grammatical category used in many languages to mark the relationship between a noun or pronoun and other words in a sentence. Case marking is typically indicated through inflection. The number of cases varies across languages. Cases are often marked with inflection. In some languages, case often affects how articles, pronouns, and adjectives should be inflected. Previous probing studies show that case is often one of the most challenging morphological categories to be learned by language models (Edmiston, 2020; Bisazza and Tump, 2018; Acs et al., 2023).

Gender Most Indo-European languages mark genders in nouns and often require agreement with in other part-of-speech in the sentence, such as verbs and adjectives Bloomfield (1933). Gender systems exhibit substantial diversity in their number of genders, assignment rules. Some languages (e.g., Basque) do not have a gender system. Romance languages have a binary gender system. On the other hand, Germanic and Slavic languages often have more than two genders. For example, Dutch nouns are either common or neuter while German nouns can be masculine, feminine, or neuter. Spanish masculine nouns end in *"-o"* while feminine nouns end in *"a"*. There is also a certain degree of irregularity.

G Extended Results

Table 8 shows the full breakdown of probing accuracy per task and per language.

H Detailed Results of the Statistical Analysis

We provide the formulation and results of the statistical models. We define our model as the combination of two interaction terms between `irregularity` and `TD` and between `TTR` and `TD`. The syntax is as follows. The results can be found in Table 9.

```
accuracy ~ irregularity*TD + TTR*TD +
(1|language) + (1|task)
```

Fixed Effects				
Variable	Estimate	SE	*t*-value	*p*-value
(Intercept)	2.71845	0.61012	4.456	<.001 ***
Training data (TD)	2.11411	0.55564	3.805	**<.001** ***
Irregularity (I)	2.83197	0.42242	6.704	**<.001** ***
TTR (T)	-0.62728	0.57127	-1.098	0.272179
TD*I	2.03621	0.31730	6.417	**<.001** ***
TD*T	-0.62728	0.57127	-1.098	0.272179
Random Effect				
Group	Name	Variance	Std.Dev.	
language	(Intercept)	3.0920	1.7584	
task	(Intercept)	0.3571	0.5976	

Table 9: Results of linear mixed effect regression with *probing accuracy* of mT5 as outcome variable and *language* and *task* as random effects. Fixed effects are irregularity (I) and TTR (T), training data (TD) and their two-way interactions.

Language	Genus	Family	POS	Number	Tense	Case	Gender
English	Germanic	Indo-European	N	2	–	–	–
English	Germanic	Indo-European	V	–	2	–	–
German	Germanic	Indo-European	N	2	–	4	3
German	Germanic	Indo-European	V	–	2	–	–
Dutch	Germanic	Indo-European	N	2	–	–	2
French	Romance	Indo-European	N	2	–	–	2
French	Romance	Indo-European	V	–	4	–	–
Spanish	Romance	Indo-European	N	2	–	–	2
Spanish	Romance	Indo-European	V	–	4	–	–
Portuguese	Romance	Indo-European	N	2	–	–	2
Romanian	Romance	Indo-European	N	2	–	–	2
Turkish	Turkic	Altaic	N	2	–	7	–
Russian	Slavic	Indo-European	N	2	–	6	3
Russian	Slavic	Indo-European	V	–	3	–	–
Czech	Slavic	Indo-European	N	2	–	–	3
Hebrew	Semitic	Afro-Asiatic	N	2	–	–	2
Hindi	Indic	Indo-European	N	2	–	2	2
Urdu	Indic	Indo-European	N	2	–	2	–
Urdu	Indic	Indo-European	V	2	–	–	–
Basque	Basque	Basque	N	2	–	11	–
Estonian	Finnic	Uralic	N	2	–	18	–
Estonian	Finnic	Uralic	V	–	–	–	–
Latvian	Baltic	Indo-European	N	3	–	5	3
Latvian	Baltic	Indo-European	V	–	3	–	–
Arabic	Semitic	Afro-Asiatic	N	–	–	2	–

Table 6: List of studied languages, their genera and families, along with the number of possible classes within the studied dataset per morphological properties of 17 investigated languages (N = noun; V = verb)

No.	**Language**	**Task**	**mT5**	**ByT5**	**fastText**
1	Arabic	case	66.17	49.25	37.81
2	Basque	case	91.39	92.55	17.22
3	Basque	number	92.00	89.42	87.99
4	Czech	gender	81.09	78.87	72.13
5	Czech	number	96.50	90.05	85.50
6	Dutch	gender	90.00	88.39	72.50
7	Dutch	number	97.00	98.24	92.00
8	English	number	97.50	98.55	98.50
9	English	tense	98.50	98.53	97.00
10	Estonian	case	88.10	86.84	62.85
11	Estonian	number	91.50	92.47	89.49
12	Estonian	tense	90.50	93.84	96.49
13	French	gender	95.00	90.39	92.00
14	French	number	99.50	98.24	95.49
15	French	tense	21.29	46.30	91.08
16	German	case	65.00	40.42	28.00
17	German	gender	29.85	74.42	76.00
18	German	number	92.00	89.00	84.50
19	German	tense	87.00	85.95	87.00
20	Hebrew	gender	95.50	95.05	89.49
21	Hebrew	number	99.50	98.82	95.49
22	Hindi	case	13.00	81.58	63.49
23	Hindi	gender	95.50	90.74	49.00
24	Hindi	number	95.00	90.95	61.57
25	Latvian	case	89.50	98.74	84.50
26	Latvian	gender	68.00	32.61	64.49
27	Latvian	number	68.50	70.39	63.49
28	Latvian	tense	82.59	84.50	82.58
29	Portuguese	gender	95.00	93.66	97.00
30	Portuguese	number	95.00	97.55	97.50
31	Romanian	gender	93.50	93.68	91.00
32	Romanian	number	97.00	95.84	93.50
33	Russian	case	48.04	36.02	82.55
34	Russian	gender	8.46	82.98	50.74
35	Russian	number	96.00	93.39	91.50
36	Russian	tense	92.54	9.59	92.03
37	Spanish	gender	93.50	94.66	97.00
38	Spanish	number	99.00	97.84	97.50
39	Spanish	tense	89.00	86.47	31.00
40	Turkish	case	95.57	72.88	61.57
41	Turkish	number	94.00	89.32	94.99
42	Urdu	case	87.00	77.37	61.50
43	Urdu	number	90.00	90.87	81.49

Table 8: Accuracy scores of each task in each language from mT5, ByT5, and fastText