

# Morphological Knowledge in Multilingual Large Language Models:

A Comparative Analysis of mT5 and ByT5

by

Thi Thao Anh Dang

S1097513

MA degree programme in  
Linguistics and Communication Sciences (research)

July 25<sup>th</sup>, Nijmegen

Supervisor(s): Dr. Lukas Galke  
Assessor: Dr. Andreas Liesenfeld

**Radboud University**



# Abstract

Since the revolution of large language models (LLMs) in natural language processing, researchers have been interested in how they encode and process linguistic knowledge (Belinkov, Gehrmann, & Pavlick, 2020; Rogers et al., 2021; Tenney, Das, & Pavlick, 2019). In this thesis, we investigate the morphological properties of 17 languages in mT5 and ByT5. We used probing classifiers to extract the amount of morphological knowledge encoded in contextualized word representations. We first compared languages with respect to their morphological content in LLMs. We also analyzed the morphological learnability of each hidden layer of the models. We then examined the effect of morphological complexity and training resources on the morphological representations of LLMs. By comparing mT5 and ByT5, two LLMs trained with the same architecture, objective, and training data yet different tokenizers, we are able to explore the impact of tokenization on morphological learning. Our analyses show that (1) LLMs learn the morphological systems of some languages better than others, (2) morphological information is encoded in the middle and late layers of the network, (3) morphology is learned earlier by subword-level models, yet at the end subword-level and character-level models obtain comparable morphological capabilities, and (4) morphological irregularities and training sizes effect the morphological capabilities of LLMs, such that irregular languages require more training data. Our findings also provide methodological insights for probing studies. We found that the last subword or character of a word encodes most of its morphological information. In addition, we found that both linear classifiers and MLPs are able to extract morphological information of words.

# Acknowledgements

Upon completing my thesis, I would like to give my deepest gratitude to the people that helped me during this intense Research Master program. First and foremost, I would like to thank my supervisor, Dr. Lukas Galke, for his great support during my internship at the Max Planck Institute for Psycholinguistics. I encountered many difficulties during the time I wrote this thesis, yet he always provided me with the most detailed explanations, corrected my code, and commented on my manuscript. I learned so many research skills from him, especially writing and coding. I also want to thank Dr. Limor Raviv and all members of the LEADs lab for offering me many research opportunities. It is my honor to be a part of such a warm and welcoming research environment.

Secondly, I would like to thank Prof. Stefan Frank for supervising my internship at the Centre for Language Studies. My internship there was delayed for quite a long time because of my limited coding skill. Yet he was always patient and supportive. Thanks to his insightful supervision, I was able to finish the internship while enhancing my coding skills.

I could not go through this 2-years journey in the Netherlands without emotional support from my precious friends and family. I would like to thank my supportive parents, brother, and sister for always being there for me. I thank all of my friends in the Netherlands, especially Huy, Truc, and Phuoc, for spending their so much of their time hanging out with me. My time in the Netherlands was so enjoyable with you aside.

Despite living thousands of kilometers away, my Vietnamese friends were always there when I needed someone to talk to. I thank Gia Han for listening to me talking about my daily problems from Australia, thank Duong Tram and Bao Tram for visiting me in the Netherlands. Last but not least, I thank Tin for always being by my side no matter how annoying I can be.

# Contents

Abstract	<b>i</b>
Acknowledgements	<b>ii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation	1
1.2 The Current Study	3
1.2.1 Research Gaps	3
1.2.2 Research Questions	5
1.3 Structure of the Thesis	6
<b>2 Background and Related Work</b>	<b>7</b>
2.1 Probing the Linguistic Knowledge of LLMs	7
2.1.1 Behavioral Probing	8
2.1.2 Structural Probing	9
2.2 Morphological Knowledge of Large Language Models	10
2.3 Morphological Complexity	13
2.4 Tokenization Methods	15
2.4.1 Sub-word Tokenization	15
2.4.2 Character-level Tokenization	16
2.4.3 Impact of Tokenization Method on LLMs' Performance Across Languages	17
2.4.4 Summary	18
<b>3 Data and Methods</b>	<b>19</b>
3.1 Large Language Models	19
3.1.1 T5	19
3.1.2 mT5	20

3.1.3	ByT5 . . . . .	21
3.2	Probing Tasks . . . . .	21
3.2.1	Number . . . . .	22
3.2.2	Tense . . . . .	22
3.2.3	Case . . . . .	22
3.2.4	Gender . . . . .	23
3.3	Considered Languages . . . . .	23
3.4	Dataset . . . . .	24
3.5	Feature Extractions . . . . .	24
3.6	Probing Classifiers . . . . .	25
3.7	Subword Pooling . . . . .	26
3.8	Controls and Evaluation . . . . .	26
<b>4</b>	<b>Results</b>	<b>30</b>
4.1	Overall Probing Accuracy . . . . .	30
4.2	Layer-Wise Analysis . . . . .	33
4.3	Effect of Model Configuration . . . . .	34
4.3.1	Probing Architecture . . . . .	34
4.3.2	Subword Pooling Methods . . . . .	35
4.4	Effect Of Linguistic Factors . . . . .	36
4.4.1	Task . . . . .	36
4.4.2	Morphological Complexity and Training Data . . . . .	37
<b>5</b>	<b>General Discussion and Conclusions</b>	<b>40</b>
	List of Figures	44
	List of Tables	46
	Bibliography	47
	Appendix	57

## Chapter 1

# Introduction

### 1.1 Motivation

Large language models (LLMs) such as Llama (Touvron et al., 2023), GPT-3 (Radford et al., 2019), and T5 (Raffel et al., 2020) have attracted a lot of attention from researchers and the public for their abilities to perform human-like tasks. At the moment, most LLMs are trained on Transformer architecture by Vaswani et al. (2017). A lot of LLMs are multilingual. For instance, mBERT and mT5, the multilingual expansions of BERT and T5, support more than 100 languages (Devlin et al., 2018; Xue et al., 2021). These LLMs are excellent at performing natural language processing (NLP) tasks (A. Wang et al., 2018; Wei et al., 2022). Examples of these downstream tasks include text summarization (L. Wang et al., 2023), machine translation (Edman et al., 2024), text summarization (Ahmed & Devanbu, 2022), and sentiment analysis (Roumeliotis et al., 2024). They are able to understand and generate text in multiple languages with excellent lexical and grammatical accuracy (Conneau et al., 2020; Lample & Conneau, 2019; Xue et al., 2021). LLMs are trained on massive amounts of text data using machine learning algorithms. These algorithms learn the relationship between words through autoregressive training, in which they learn to estimate the probability of the next word in context. Notably, LLMs are trained with the objective of predicting the most probable token that comes next in the sequence. In order to achieve such impressive language proficiency, they have to be equipped with some formal linguistic knowledge, namely syntactic, semantic, and morphological knowledge. However, while it is evidential that LLMs perform well on downstream NLP tasks, it is unclear *how* they acquire the linguistic knowledge that is deemed to be the prerequisite for those tasks to be successful. Witnessing the impressive linguistic ability of LLMs, researchers sought to understand what patterns the models have

learned. One growing research line is to discover what kind of linguistic information is learned in the hidden activations of LLMs. Examples of these activations are *contextualized word embeddings* and *sentence embeddings*.

A large body of research has been conducted on analyzing the linguistic knowledge of neural language models, these include Long Short-Term Memory (LSTMs) (Linzen et al., 2016), neural machine translation (NMT) systems (Belinkov, 2022; Belinkov, Durrani, et al., 2017; Vylomova et al., 2017), and LLMs, with the most studied one being BERT (Rogers et al., 2021). The linguistic features being studied ranged from syntax (Hupkes et al., 2018; Linzen et al., 2016; McCoy et al., 2019), semantic (Ettinger, 2020; Tenney, Xia, et al., 2019), and morphology (Acs et al., 2023; Belinkov, Durrani, et al., 2017, 2020; Edmiston, 2020; Vylomova et al., 2017). In general, it has been shown that language models understand some aspects of the structural nature of human language, although this does not necessarily mean that their capabilities are the same as us humans. A classic example is the study by Linzen et al. (2016), which show that LSTMs can learn subject-verb agreement. Although this area of research is huge, the current focus is on English. It is less known whether LLMs can produce high-quality representations for low-resource languages (Adams et al., 2017; Gandhe et al., 2014; Ruder et al., 2019). There is little work on the linguistic capabilities of other languages. Given that most LLMs nowadays are multilingual, it is important to understand how linguistic knowledge of LLMs is manifested across languages. This is an intriguing question considering the differences in orthography, typology, and complexity between languages.

This study focuses on analyzing the morphological knowledge, the ability to learn how grammatical and semantic information is embedded in word segments, of LLMs. Morphology is particularly interesting when studying the linguistic knowledge of multilingual LLMs because each language has a unique morphological system and some languages are morphologically-richer than others. This means they may have larger sets of morphological features and/or have more irregularities (Ackerman & Malouf, 2013). Such morphological complexity has been suggested to affect the multilingual LLMs' performance, such that morphologically-complex languages are more difficult for them to learn (Cotterell et al., 2018; Galke et al., 2023; Mielke et al., 2019; Park et al., 2021).

Deep learning systems, including LLMs, transform text into trillions of numbers for processing, ultimately enabling them to interpret and generate human language. Words are processed and encoded differently across systems due to recent advances in tokenization, the task of segmenting text into tokens in order for the systems to process at ease. Currently,

there are two predominant methods for tokenization, namely subword-level tokenization (Kudo & Richardson, 2018; Sennrich et al., 2016) and character-level tokenization (J. H. Clark et al., 2022; Fleshman & Van Durme, 2023; Xue et al., 2022). It is unclear how the choice of tokenization methods affect how morphology is learned and how this effect is manifested across languages. In addition, LLMs achieved such impressive capabilities because of the massive amount of training samples. LLMs are trained on a very large amount of text, with English being the dominant language. The amount of training data undoubtedly has an impact on the performance of LLMs (Warstadt et al., 2020). However, it is unknown how it influences their morphological knowledge, especially when considering the potential effect of morphological complexity.

On the basis of these observations, this work seeks to understand whether morphological properties are captured differently across languages and how technical factors, such as tokenization method, amount of training data, and linguistic factors, such as morphological complexity, may create such differences. To do this, we probed the morphological knowledge of mT5 and its token-free variant, ByT5, in 17 morphologically different languages. mT5 and ByT5 are trained with different tokenization methods (i.e., sub-word level and character-level), yet they are both trained using the T5 architecture and on the same multilingual dataset. The probing procedure includes: (1) extracting contextual word embeddings from LLMs' encoder, (2) training probed classifiers on those embeddings, and the word's morphological features, and (3) evaluating classification accuracy as a proxy for how well the word representations encode the morphological features.

## 1.2 The Current Study

This section presents the gaps in the body of work in analyzing the morphological knowledge of multilingual LLMs and formulates the research questions which would fill in the gaps.

### 1.2.1 Research Gaps

Over the past few years, much research has been done to discover whether linguistic representations indeed exists and how it is manifested in the hidden activation of language models (Manning et al., 2020; Rogers et al., 2021; Tenney, Das, & Pavlick, 2019). Despite this large body of literature on linguistic probing, there is still much to be done. This section presents the current gaps in the literature.

**Scarcity of work on morphology** The main focus of current probing studies is on syntax and semantic knowledge. Not much is known about the morphological knowledge of LLMs, especially in the multilingual context. Notably, there are even fewer studies on derivational morphology. It has been found that morphological knowledge is encoded at the lower levels in the network (Belinkov, Durrani, et al., 2020; Peters et al., 2018). However, most studies focus on analyzing monolingual LLMs, such as BERT and ELMo, or NMT systems trained on pairs of languages (Belinkov, Durrani, et al., 2020; Bisazza & Tump, 2018), while studies on multilingual LLMs are rare (Belinkov, Durrani, et al., 2020; Edmiston, 2020). In addition, very few studies have systematically investigated the relationship between the amount of training data on the morphological knowledge of language models.

**Strong focus on BERT and NMT systems** Most studies about the linguistic knowledge of neural language models investigate NMT systems, which are often trained on a source and a target language (Belinkov, Durrani, et al., 2017, 2020; Vylomova et al., 2017). A few studies have investigated BERT (Acs et al., 2023; Edmiston, 2020). It is unclear whether the findings generalize to other LLMs, such as mT5 and ByT5.

**Inconsistent Effect of Tokenization Methods** The effect of tokenization on the performance of LLMs has long been established. Studies investigating such effects strongly focused on NMT systems (Belinkov, Durrani, et al., 2020), which are limited in the number of modeled languages. Many of these studies have found that character-level models have better morphological knowledge because they can handle out-of-vocabulary words more effectively. 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., 2023). The importance of morphology in multilingual language modeling is evident (Hofmann et al., 2021; Toporkov & Agerri, 2024), such that many studies propose tokenization algorithms to make language models capture the morphological structure of languages (Goldman & Tsarfaty, 2022). Since LLMs are all-in-one systems and available in many languages, more studies on a larger number of languages need to be conducted in order to gain more insight into the broader effect of tokenization methods, the only step that is disconnected from the end-to-end learning systems.

**Mixed findings about the effect of morphological complexity** There are some studies that investigate the link between morphological complexity and the language modeling difficulty of LLMs (Cotterell et al., 2018; Dang et al., 2024; Galke et al., 2023;

Mielke et al., 2019; Park et al., 2021). However, the findings have been mixed. Since most of these studies focus more on the overall learnability (e.g., surprisal, Park et al., 2021, perplexity, Gerz et al., 2018) rather than the specific morphological knowledge, there can be other confounds that affects learnability other than morphological complexity.

## 1.2.2 Research Questions

The current study aims to discover the morphological representations of multilingual LLMs, specifically mT5 and ByT5. The following questions are of interest:

- (1) How do multilingual LLMS represent morphological knowledge differently across languages, and how is this difference affected by morphological complexity?
- (2) At which layers of LLMs is morphological knowledge captured?
- (3) How does tokenization methods (i.e., subword-level and character-level tokenizers) affect morphological representations of LLMs?
- (4) How does the amount of training data affect the morphological knowledge learned by LLMs?

To answer these questions, we used structural probing techniques to probe the embeddings of mT5 (Xue et al., 2021) and ByT5 (Xue et al., 2022) for their morphological knowledge in multiple languages and for multiple morphological features. There are two reasons behind this model choice. First, these two LLMs are multilingual, open-source, and have transparent information about the amount of training data for all supported languages (Xue et al., 2021). Second, they are trained using the same training data, architecture, and training objectives. The only difference is that mT5 uses sub-word tokenization methods (SentencePiece), while ByT5 is trained on raw unicode characters. This interesting difference allows us to investigate the effect of tokenization methods while controlling for possible confounds (e.g., difference in model architecture and training objectives). To our knowledge, this is the first study to probe the morphological representations of these two LLMs.

To probe the morphological knowledge of mT5 and ByT5, we trained probing classifiers on contextualized word representations and their corresponding morphological labels. The accuracy scores of the classifiers are considered an indication of the morphological knowledge encoded in the word representations. we then analyzed the patterns of morphological knowledge across layers, across languages, and across tokenization methods. We also aim to provide insights into the relationship between morphological complexity, training data

sizes, and morphological knowledge of LLMs. Drawing upon previous studies (Dang et al., 2024; Galke et al., 2023), we expect that the degree of irregularity of the morphological system to negatively affect LLMs' morphological representations.

### 1.3 Structure of the Thesis

The rest of the thesis is structured as follows.

**Chapter 2** aims to provide background on probing the linguistics knowledge of language models and to formulate the research questions.

**Chapter 3** first introduces the LLMs and dataset and presents the probing procedure in detail. It also contains information about baselines and evaluation of probing accuracy.

**Chapter 4** reports a series of analyses on the accuracy of the probing tasks, comparing probing results of different types of probes and pooling methods.

**Chapter 5** discusses and summarizes the findings as well as provides insights for the research questions.

## Chapter 2

# Background and Related Work

In Chapter 1, we provided the context for the study by introducing the research line on probing linguistic knowledge of LLMs as well as present the current gaps in probing the morphological knowledge of LLMs and formulate the research questions for the study. In this chapter, we set out the backdrop for the investigation by introducing the LLMs (i.e., T5, mT5, and ByT5) and tokenization strategies. We then describe various approaches to probing the morphological abilities of LLMs and the current findings.

## 2.1 Probing the Linguistic Knowledge of LLMs

Transformer-based LLMs are simple, powerful, and efficient (Brown et al., 2020; Zhao et al., 2023). However, as discussed in the *Introduction*, their internal workings are not transparent because of the end-to-end training procedure and high number of learned parameters. While a large amount of research successfully improves the performance of these systems in various benchmarks (Bubeck et al., 2023; Wei et al., 2022), there is limited knowledge about what patterns they have learned from the massive amount of training data. As such, LLMs are often called “black-box” and interpreting their internal working become one of the most important topics in NLP research. This research endeavor is called *interpretability* (Belinkov, Gehrmann, & Pavlick, 2020; Madsen et al., 2022). One of the most intriguing characteristics of LLMs is their language capabilities, as they are not explicitly taught yet able to learn linguistic patterns the in training text. Their internal linguistic structures have become one of the most important topics in this research line. Several approaches have been proposed to investigate the linguistic properties of LLMs. They are presented in the subsections below.

### 2.1.1 Behavioral Probing

LLMs are trained on fixed-sized datasets. How capable are they in generalizing their knowledge to unseen data? Many studies address this question by prompting the LLMs to tackle new tasks that are reflective of specific knowledge. This can be simple linguistic structures such as subject-verb agreement (Linzen et al., 2016), word order, and morphological inflection (Dang et al., 2024; L. Liu & Hulden, 2022; Weissweiler et al., 2023). Linzen et al. (2016) test whether LSTMs are able to learn syntactic dependencies using the number prediction task. Given an incomplete sentence (e.g., *The keys to the cabinet \_\_*), the LSTMs have to predict the number of the following verbs as either *singular* or *plural*. They tested LSTMs with various training objectives (e.g., classification, language modeling).

A more complex probing approach is made possible because of the ability of some LLMs to follow explicit instruction and reason accordingly (Wei et al., 2022). As such, these behavioral probing experiments are usually run on strongly capable models such as GPT-3.5 and GPT-4, aiming to elicit the exact information of interest (Belinkov, Gehrmann, & Pavlick, 2020). For example, Weissweiler et al. (2023) use a Wug test-like experimental paradigm by Berko (1958) to investigate to what extent ChatGPT can perform morphological inflection on nonce words in multiple languages. They prompted GPT-3.5 with an instruction such as: “*Fill in the blank with the correct past tense of the word ‘wug’. Give your response in one word. They wug all the time. In fact, they just \_\_ yesterday.*” (Weissweiler et al., 2023). GPT-3.5 responds with a suitable form of the nonce word. How well their answers are compared with human baselines reflects their morphological knowledge.

Behavioral probing is not only used to test what is learned by LLMs but also to see if the models learn what it is supposed to learn. McCoy et al. (2019) tested language models on their natural language inference (NLI) ability. They posited that the reason behind LLMs’ high performance on various linguistic benchmarks may not be because they understand syntax, but because they employ certain statistical heuristics. They designed a heuristic analysis task, in which the model receives a *premise* and a *hypothesis*. Based on these, it predicts whether the meaning of the premise entails that of the hypothesis. McCoy et al. (2019) provided an example of the task as followed. If the sentence “*The judge was paid by the actor.*” is the premise then it entails the hypothesis “*The actor paid the judge.*”. McCoy et al. (2019) hypothesized that the models may employ the *lexical*

*overlap heuristic*, which means that the premise entails the hypothesis if all words in the hypothesis sentence are contained in the premise sentence. If the model adopts this heuristic, then it would predict that "The actor was paid by the judge." entails "The actor paid the judge.". They found that BERT and other language models perform poorly in heuristic analysis task while achieve very high accuracy scores on standard NLI tasks.

### 2.1.2 Structural Probing

While behavior probing offers a way to qualitatively understand whether LLMs acquire certain knowledge, it is not possible to know how LLMs learn internally. Recently, NLP researchers have put forth an approach called *structural probing* (Madsen et al., 2022). The proposal is that linguistic knowledge is implicitly stored in token representations, which are the embeddings. Given a classifier trained on a set of embeddings and some labels of a linguistic feature (e.g., sentence length), if the model can successfully predict the features of words in the test set, it can be inferred that these embeddings capture this feature. This method is often called differently in literature. It can be *diagnostic classification* (Hupkes et al., 2018), *probing tasks* (Conneau et al., 2018; Rogers et al., 2021), or *auxiliary prediction tasks* (Adi et al., 2016). It is often a minimal linear classifier or a multilayer perceptron (MLP).

Probing classifiers transform the task of evaluating the learned information of language models into a simple multi-class classification problem. This approach has been used extensively in addressing the question of how much linguistic knowledge is captured by hidden state activations of the network. To illustrate, consider one of the first study that use this approach to probe and evaluate the linguistic information captured by the classic word embedding algorithms. Using linear classifiers as probing models, Köhn (2015) evaluates word embeddings of several embedding algorithms (e.g., CBOW, skip-gram) (Mikolov et al., 2013) on morphosyntactic knowledge in multiple languages. It was found that these embeddings achieve high and consistent accuracy in morphosyntactic tasks across languages, which implies that morphology is learned equally well across languages. However, it is not without uncertainty and limitations. Belinkov (2022) provided a comprehensive review and evaluation of the potentials and shortcomings of this approach. First, it can be difficult to interpret the performance of the classifiers. If the classifiers yield high accuracy, it can be that the representations successfully capture the linguistic property of interest. It can also be that it is actually the classifier that learns about the

property rather than the LLM (Hewitt & Liang, 2019), especially in the case when the probing classifier is complex (Hupkes et al., 2018).

Due to its simplicity and agnostics, this probing approach is receiving a lot of attention from NLP scientists. More extensions have been discovered, especially in examining the linguistic content of multiple languages and how it interacts between languages. For example, it is shown that a single probe can predict multiple morphological features (Shapiro et al., 2021). Stanczak et al. (2022) trained probing classifiers combining data from the same morphological categories, yet in two different languages. They found that there is shared information and the amount of information correlates with the degree of typological similarity of the languages.

More recently, researchers have started to look at the linguistic knowledge embedded in the attention pattern of Transformers (Chen et al., 2023; K. Clark et al., 2019; Voita et al., 2019). K. Clark et al. (2019) probed BERT’s attention head for multiple linguistic features (e.g., dependency syntax and coreference resolution). Under the fact that attention heads decide how words are related to each other, they consider each head as a classifier that receives a token as an input and output the token that it pays most attention to.

## 2.2 Morphological Knowledge of Large Language Models

Morphological knowledge concerns how meaningful word units can be combined to express a range of grammatical information. Morphology is divided into two subtypes. **Inflectional morphology** carries the grammatical information about the relationship between the word, such as number, tense, case, and mood, and its surrounding. For example, when one wants to mention two apples, one needs to add the suffix *-s* to form plurality. On the other hand, **derivational morphology** is about the change in part-of-speech and often also the meaning of the word. The noun *development* is formed by adding the suffix *-ment* to the noun *develop*. Adding the prefix *un-* to a certain word reverses its meaning. For example, *interesting* and *uninteresting* have opposite meaning.

Languages in the world exhibit a wide range of morphological systems, which refer to the ways words are formed and modified to convey different meanings (Dryer & Haspelmath, 2013). These systems can be broadly categorized into several types Bloomfield (1933): Synthetic languages, such as English, Dutch, and German, form words by combining morphemes with the stems. On the other hand, analytic languages, such as Chinese and Vietnamese, have little to no inflection and rely heavily on word order and auxiliary words

to convey grammatical relationships. Agglutinative languages, like Turkish, Finnish, and Japanese, are a type of synthetic language that primarily uses discrete affixes (prefixes and suffixes) to modify word meanings. For example, in Turkish, the word "*ev-ler-im-iz-den*" means "*from our houses,*" with each affix contributing a specific meaning: "*ev*" (house), "*-ler*" (plural), "*-im*" (my), "*-iz*" (our), "*-den*" (from). Fusional languages, such as Russian, Arabic, and Latin, also combine morphemes, but the affixes often fuse together, making it difficult to separate the individual meanings of each morpheme. For instance, in Latin, the word "*amāvērunt*" (they loved) contains the stem "*amā-*" (love) and the ending "*-vērunt,*" which encodes both the past tense and the third-person plural subject. In addition, morphological systems may vary per language families. Germanic languages like English, Dutch, German have little inflectional morphology than Slavic languages (e.g., Russian and Czech) and Indic languages (e.g, Hindi and Urdu). In order to acquire morphological knowledge, LLMs have to recognize that words are formed using multiple segments and certain segments carry grammatical information.

Not much work has been done on probing the morphological knowledge of LLMs. On the behavioral level, it has been shown that LLMs has yet to reach human performance on the Wug test (Dang et al., 2024; Weissweiler et al., 2023). Some structural probing studies have also been conducted. Vylomova et al. (2017) used probing classifiers to test character-level and word-level NMT systems trained with LSTMs. They built NMT systems in several language pairs. All of which have rich and varying morphological systems (e.g., Russian, Arabic, and French). Their results reveal one of the first systematic observations about the morphological knowledge of LLMs. First, they found that character-level models are better at capturing morphological knowledge than word-level models. Through a layer-wise analysis, they showed that the lower layers learn morphology better, while higher layers better capture the semantic properties of words.

Edmiston (2020) reported a different observation. It was found that the probing accuracy is stable across layers for English, Spanish, and French. However, for German and Russian, performance is poor at the lower layers and improves at the higher layers. In contrast with the above-mentioned studies, (Acs et al., 2023) showed that morphological knowledge is indeed encoded in the higher layers of the network. One consistent finding of all of the previous studies is that, on average, the concatenated representations of all layers always achieve the highest accuracy scores. This conclusively shows that morphological knowledge may be unevenly distributed across layers.

Belinkov, Durrani, et al. (2020) used structural probing techniques to investigate the

representations of syntactic, semantic, and morphological knowledge in neural machine translation models. They analyzed the morphological knowledge of NMT systems in different source-target language pairs and found that the lower layers captured the most morphological knowledge, while syntactic and semantic information is learned in the higher layers of these models. Moreover, the highest classification accuracy is achieved when concatenating all intermediate layers. This shows that while most of the morphological features are learned in the first layers, there are a few other features that are learned at the higher levels. Notably, this pattern holds across NMT systems trained in different language pairs. Furthermore, they found that the probing task accuracy is lower for morphologically rich languages. They also compared the linguistic representations of character-level models and subword-level models, and found that morphological information is better captured in models trained on characters.

Mikhailov et al. (2021) probed the morphological knowledge of several BERT-based and XML-based models and their POS – fine-tuned versions in 4 European languages using a range of probing techniques. In addition to morphosyntactic probing tasks, they also used fine-tuning and perturbation techniques, which tests how the knowledge is affected after removing certain lexical or grammatical elements around the target words. They found that the distribution of knowledge is similar across languages and also across models. Specifically, the accuracy scores are highest in the middle layers. A more fine-grained neuron-level analysis shows that there is indeed a difference in how LLMs learn morphological systems between languages. For morphologically complex languages such as Russian and German, the LLMs needs more neurons to learn. They also observed that while English morphological knowledge is distributed evenly across layers, that knowledge of other languages is mostly in the middle or deeper layers of the network. These results are in contrast with Belinkov, Durrani, et al. (2020) and Vylomova et al. (2017) who found that low-level features like morphology are learned in the lower layers.

Methodologically similar to our work, Acs et al. (2023) conducted a large-scale study on 248 morphological probing tasks across 42 languages. They investigated two multilingual LLMs in the BERT family, namely mBERT and XLM-RoBERTa. In addition, they performed perturbation experiments, in which they mask either the target word or the adjacent words. Their main finding is that morphology is actually encoded in the embeddings and that the amount of knowledge varies across morphological features rather than across languages. In other words, they found no difference in morphological content between typologically-diverse languages. Through a layer-wise analysis, they also found

that morphological knowledge is learned better at the deeper layer, consistent with Hewitt and Liang (2019), and that probing performance is not different across layers. Their perturbation results show that morphological knowledge is mainly encoded in the word and is slightly affected by the left context.

There are also studies that suggest that there are shared morphological representations across languages (Stanczak et al., 2022). Instead of probing at the embedding level, they looked deeper into the neurons, aiming to access how information is structured in subsets of neurons in the embeddings. Using this method, they ran probing experiments on 43 languages on mBERT and XLM-R-base and found that some groups of neurons encode universal morphosyntactic information. The degree of shared knowledge is correlated with the similarity in language typology. This finding thus provides indirect support for the idea that language typology affects how morphology is learned by LLMs.

Overall, earlier work on probing the morphological representation of language models is conducted on NMT systems and LTSMs Belinkov, Durrani, et al. (2020) and Vylomova et al. (2017). More recently, the focus has been extended to Transformer-based LLMs (Acs et al., 2023; Edmiston, 2020; Wu & Dredze, 2020), yet the majority of studies investigated BERT and its variants. To summarize, previous studies provide mixed generalizations about the morphological content of LLMs regarding whether this knowledge is learned differently across languages and where it is best learned in the hidden layers of the models.

### 2.3 Morphological Complexity

Languages in the world vary in typological features. One of the most established ones is morphology. Some languages have a lot of morphological distinctions and forms (e.g., German, Russian) while others only have one word form for all grammatical expressions (e.g., Chinese, Vietnamese). On the other side of the coin, some languages have deterministic morphological rules, such that the inflected form of a word can be easily ruled out from the inflected form of another word. However, the morphological system of some languages can have many irregular forms, meaning that the inflected forms are not formed based on explicit rules. It has long been established that languages have different degrees of complexity (Dryer & Haspelmath, 2013). In linguistic literature, by conducting experiments on artificial languages, it has been shown that morphologically complex languages are harder to acquire (DeKeyser, 2005; Kempe & Brooks, 2008; Raviv et al., 2021). Extending to deep learning models, Galke et al. (2023) found that LLMs and recurrent neural networks

learn languages better if they exhibit regular structures. Cotterell et al. (2018) found the effect of the number of morphological distinctions of language modeling.

A long-standing question is what determines a language’s degree of morphological complexity. Is a language more complex than others if it has more word forms? Ackerman and Malouf (2013) propose measuring morphological complexity along two dimensions. Enumerative (E-) complexity is the number of morphological distinctions in a language morphological system. Languages that have more word forms have higher e-complexity. On the other hand, integrative (I-) complexity is the degree of irregularity of the morphological paradigm. Irregularity can be defined as the degree to which the inflection process is not determined by explicit rules. For example, the English past tense is formed by adding the suffix "-ed" or -d to the infinitive verb. However, there are some verbs that do not obey this general rule. The past form of *go* and *lead* is *went* is *led*, respectively. Cotterell et al. (2019) proposed the relationship between I-complexity and E-complexity, which they call *the low entropy conjecture*. This hypothesis posits that if a language’s morphological system consists of many inflected forms per lexeme, it cannot have a high degree of irregularity. In other words, a morphological system can be either high in E-complexity or high in I-complexity, but not both. Given this hypothesis and previous studies on the effect of morphological complexity on the learnability of languages (Cotterell et al., 2018; Dang et al., 2024; Galke et al., 2023), we predict that if the complexity of a morphological system affect how it is learned by LLMs, languages with higher I-complexity (more irregular) are harder to learn by LLMs. In this study, we consider the I-complexity score for each language computed by Wu et al. (2019). For ease of following, we refer to I-complexity as *irregularity*. Additionally, in computational linguistics, the most basic common measure for morphological complexity has been type-token ratio (TTR), which is the total number of word types divided by the number of word tokens in the same text. In this study, in addition to irregularity, we also test the effect of TTR on the morphological representations of LLMs. We used the TTR scores quantified by Bentz et al. (2015). TTR has been criticized for being affect by corpus size (Durán et al., 2004; Tweedie & Baayen, 1998). Bentz et al. (2015) computed TTR scores using 3 parallel corpora, namely the Universal Declaration of Human Rights Corpus, the Parallel Bible Corpus, and the Europarl Parallel Corpus. As such, their TTR scores may avoid this limitation. See Table 2.1 for all considered languages and their morphological complexity measures, along with the sizes of training data in for each language in mT5 and ByT5.

Language	Training Data	TTR	Irregularity
English	5.67%	-0.46	-5.94
German	3.05%	-0.01	-6.28
Dutch	1.98%	-0.39	-6.68
French	2.89%	-0.34	-4.16
Romanian	1.58%	-0.42	-3.40
Spanish	3.09%	0.001	-8.81
Portuguese	2.36%	0.038	-9.11
Turkish	1.93%	1.55	-5.96
Czech	1.72%	0.43	-5.63
Russian	3.71%	0.87	-7.74
Hebrew	1.06%	2.02	-1.78
Arabic	1.66%	1.63	-0.06
Hindi	1.21%	-0.3	-2.10
Estonian	0.89 %	1.76	-2.79
Latvian	0.87%	0.77	-7.9
Urdu	0.61%	-0.45	9.20
Basque	0.57%	1.31	19.86

*Table 2.1* 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.

## 2.4 Tokenization Methods

This section provides an overview of tokenization methods as well as their advantages and disadvantages on the performance of multilingual LLMs.

### 2.4.1 Sub-word Tokenization

One of the preprocessing steps for LLMs is tokenization, the task of segmenting text into smaller units. The straightforward approach is simply to split words by whitespace. However, much work in NMT systems shows that this method has some shortcomings. One of those is the inability to encode and translate rare and out-of-vocabulary words because of the fixed and limited vocabulary sizes (Luong et al., 2014). In addition, the word-

level tokenization method requires separate vocabulary for each language, hindering the possibility to build multilingual NMT systems (Johnson et al., 2017; Kudo & Richardson, 2018). To address these challenges, Sennrich et al. (2016) proposed a simple yet elegant tokenization method that divides text into subwords, called Byte-Pair Encoding (BPE). BPE works by merging the most frequent sequences of characters into a larger unit within a word and building a vocabulary consisting of meaningful subwords. When NMT systems encounter an unknown word, they are able to tokenize the word into known subwords and translate those meaningful subwords instead of the entire word. It is important to note that while subword tokenization is morphologically-inspired, it is not a morphological segmentation tool like Morfessor (Smit et al., 2014). A subword does not necessarily correspond to a morpheme. For example, consider *tiktoken* - a BPE tokenizer of GPT-3.5 and GPT-4 (available at <https://platform.openai.com/tokenizer>), it tokenizes the word *interchangeable* into inter-, change and -able, which is morphologically correct. However, it splits the word *indication* into ind- and -ication.

There are some variants based on the original BPE algorithm. The SentencePiece tokenizer by (Kudo & Richardson, 2018) enables subword vocabulary to be built on raw sentences instead of single word like in BPE. This tokenizer is particularly effective for languages without word boundary, such as Chinese and Japanese. These sub-word tokenization methods have gained a lot of popularity. They are used not only in training NMT systems but also in pretraining the most state-of-the-art LLMs, such as BERT (Devlin et al., 2018), GPT-3 (Radford et al., 2019), and LLaMA (Touvron et al., 2023). Many LLMs are now pretrained using subword tokenizers.

### 2.4.2 Character-level Tokenization

While subword tokenization has become a standard practice in training LLMs, it is not without limitations. The large vocabulary size resulting sub-word tokenization may increase computational complexity and training efficiency of the models (Gao et al., 2020). Recent research has proposed to use raw character to train NMT systems and LLMs (Chung et al., 2016; J. H. Clark et al., 2022; Fleshman & Van Durme, 2023; Gao et al., 2020; Kim et al., 2016; Lee et al., 2017; Xue et al., 2022). Character-level tokenization can be efficient in training multilingual LLMs because many languages can be modeled using the same set of character (Gao et al., 2020). Research on MT has shown that character-level systems outperform word-level and subword-level system in translation quality (Chung et al., 2016;

Edman et al., 2024).

Tokenizer	Tokenized Sentence
Morfessor	Professor admit@@ s to shoot@@ ing his girl@@ friend
BPE (BERT)	Professor admits to sho@@ oting his gir@@ l@@ friend
SentencePiece (mT5)	Professor admit @@s to shooting his girlfriend
Character-level (ByT5)	P r o f e s s o r _ a d m i t s _ t o s h o o t i n g _ h i s _ g i r l f r i e n d

Table 2.2 An Example, taken from Belinkov, Durrani, et al. (2020) of how Morfessor (Smit et al., 2014), BPE (Sennrich et al., 2016), SentencePiece (Kudo & Richardson, 2018), and byte-level tokenizer (Xue et al., 2022) tokenize the sentence "Professor admits to shooting his girlfriend".

### 2.4.3 Impact of Tokenization Method on LLMs' Performance Across Languages

Tokenization is undoubtedly one of the most crucial steps in building LLMs, especially for language modeling objectives. It has a huge impact on how words are processed and represented. Tokenization methods therefore can have a huge impact on the performance of the models. Ali et al. (2023) compared the impact of BPE and Unigram tokenizers on downstream tasks and found that the preference for tokenization methods differs across languages. While BPE works better for Germanic languages, such as German and English, Unigram is more well-suited for Romance languages, such as Spanish. Lee et al. (2017) evaluated both tokenization methods in machine translation tasks and found that character-level tokenizers perform as well as or better than sub-word tokenizers. They highlighted that character-level tokenizer offers better translation quality in multilingual and low-resourced setting 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 sub-word level LLMs (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. It is also better in handling rare and similar words. We will discuss the effect of tokenization methods on the morphological knowledge of LLMs in the later section.

#### 2.4.4 Summary

In this chapter, we provide background on probing morphological knowledge of LLMs. we introduce LLMs (i.e., mT5 and ByT5), probing methods, and tokenization methods. A survey of previous studies on morphological probing shows mixed findings on how LLMs encode morphological system of multiple languages. We aim to investigate how morphology is represented across morphological systems, features, and tokenization methods. In the next chapter, we describe in detail how we extracted hidden representations and trained and evaluated probing classifiers.

## Chapter 3

# Data and Methods

### 3.1 Large Language Models

This section provides an overview of the LLMs used in the study, namely mT5 and ByT5. A concise introduction of T5 (Raffel et al., 2020), the foundation for building mT5 and ByT5. We chose mT5 and ByT5 for our morphological analysis due to their similar architectures and training data. The primary differences between them lie in their tokenization methods and network depth, allowing us to specifically examine the impact of tokenization methods on their morphological representations.

#### 3.1.1 T5

LLMs have revolutionized the field of natural language processing, enabling remarkable performance on a wide range of tasks (Brown et al., 2020; Zhao et al., 2023). However, most existing models are fine-tuned for specific downstream tasks. This limits their versatility and requires separate models for different tasks. To address this limitation, Raffel et al. (2020) introduced the Text-to-Text Transfer Transformer (T5), a transformer-based model that reframes all NLP problems as a text-to-text problem. This allows a single model to be trained on multiple tasks simultaneously. By representing inputs and outputs as text sequences, T5 can handle a wide range of NLP tasks, such as translation, summarization, and question answering within a unified architecture. This approach eliminates the need for task-specific architecture and enables efficient transfer learning across tasks. T5 is pre-trained on the Colossal Clean Crawled Corpus (C4), a massive corpus of unlabeled English text data extracted from the Common Crawl corpus. It is trained with masked language modeling objectives, where one or some consecutive tokens of the input is masked

and the LLMs have to predict the masked part.

### 3.1.2 mT5

Model Size	Param	Enc.	Dec.	Embed. Dim.
mT5-small	300M	8	8	512
mT5-base	582M	12	12	768
mT5-large	1.23B	24	24	1024
mt5-XL	3.74B	24	24	2048
mT5-XXL	12.9B	24	24	4096
ByT5-small	300M	12	4	1472
ByT5-base	582M	18	6	1536
ByT5-large	1.23B	36	12	1536
Byt5-XL	3.74B	36	12	2560
ByT5-XXL	12.9B	36	12	4672

*Table 3.1* Configuration of mT5 and ByT5, from Xue et al. (2022) (Param = Number of parameters, Enc. = Number of encoder layers; Dec. = Number of decoder layers; Embed. Dim. = Embedding dimensions)

The mT5 model is a multilingual version of T5, a generative text-to-text LLMs (Raffel et al., 2020). It is trained on the mC4 corpus (Xue et al., 2021). This corpus consists of text in 101 languages, compiled from the Common Crawl web scrape. Both mT5 and T5 are comparable both in terms of model architecture and training objective. mT5 is an encoder-decoder LLM built upon the Transformer architecture. It is available in 5 sizes, ranging from 300M parameters (mT5-small) 13B parameters (mT5-XXL). Both T5 and mT5 are trained using the SentencePiece tokenizer (Kudo & Richardson, 2018), a variant of the BPE tokenizer by Sennrich et al. (2016). It has the vocabulary size of approximately 250,000 subwords, covering 104 languages (Xue et al., 2021). This means that there are shared subwords between languages. Similar to T5, mT5 is pre-trained using a masked span prediction objective. The masking span of mT5 is 3 subwords.

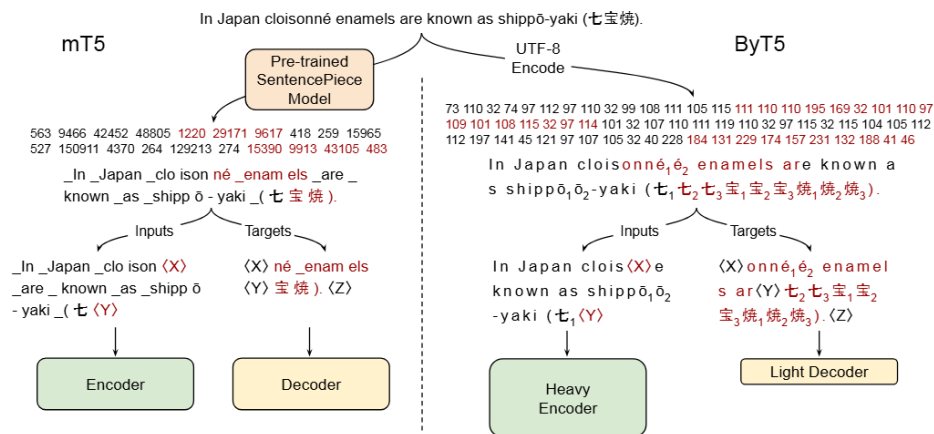


Figure 3.1 The difference in tokenization method of ByT5 and mT5 from Xue et al. (2022)

### 3.1.3 ByT5

The token-free variant of mT5, ByT5, inherits most of the property of mT5, including model architecture and training procedure. The training objective is also masked language modeling. However, ByT5 uses a span of 20 bytes. The only crucial difference between them is the tokenization method. While mT5 is trained using a sub-word tokenizer, ByT5 operates directly on raw text (bytes). It encodes text as a sequence of UTF-8 bytes. See Figure 3.1 for the difference in the tokenization step between mT5 and ByT5. Another important difference between mT5 and byT5 is the structure of the encoder and decoder stacks. In mT5, the encoder and decoder always have the same number of hidden layers. However, the encoders of byT5 have three times more layers than the decoders. This choice is made because of the need to process a large number of bytes (Xue et al., 2022). The similarity in architecture and sizes of mT5 and byT5 leads to an interesting comparison of how the tokenization method impacts how morphological knowledge is learned and how the difference in number of parameters affects it. Details about the hyperparameters of mT5 and ByT5 are provided in Table 3.1.

## 3.2 Probing Tasks

This section 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.

### 3.2.1 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. In English, they are dependent on whether the nouns end with a consonant or vowel. However, 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.

### 3.2.2 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. The tense probing task requires LLMs to identify the tenses marked in the verbs or adjectives. For some Indo-European languages, they are either past or present. Spanish and French have four different word forms for present, past, imperfect, and future tense.

### 3.2.3 Case

A case system is a grammatical category used in many languages, especially Indo-European 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. Arabic has only 3 cases, while Finnish has 16 cases. Cases are often marked with inflection. In some languages, case often affects how articles, pronouns, and adjectives should be inflected. Consider the case system in German. There are 4 cases. Nominative cases are used to indicate the subject of the sentence. In the sentence *Der Mann isst* (The man eats), the case affects the article. If the noun is in accusative case, which marks the direct object, the article for a masculine noun such as *Mann* should be

*den* (e.g., Ich sehe den Mann). In German, the case of noun, along with its gender, also decides how adjectives are inflected. When modifying a masculine noun in nominative case, the adjective *klein* (small) should become *kleine* while in other cases (i.e., dative, genitive, and accusative), it should be *kleinen*. Previous probing studies show that case is often one of the most challenging morphological categories to be learned by language models (Acs et al., 2023; Bisazza & Tump, 2018; Edmiston, 2020).

### 3.2.4 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 significant diversity in their number of genders, assignment rules. Some languages (e.g., Basque) do not have a gender system. Romance languages such as French, Spanish, and Portuguese have a binary gender system (masculine and feminine). 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. The gender carries little semantic information and is often characterized by its ending. Spanish masculine nouns end in "-o" while feminine nouns end in "a". There is also a certain degree of irregularity. Understanding the gender system means that LLMs can classify nouns into the correct predetermined gender.

## 3.3 Considered Languages

We aim to probe the morphological knowledge of a range of topologically different languages. While the chosen probing tasks by 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 systematic comparison of morphological representations, we selected languages with more than two tasks, resulting in a balanced set. The languages chosen include English, German, Dutch, French, Spanish, Portuguese, Romanian, Turkish, Czech, Hebrew, Arabic, Hindi, Urdu, Russian, Estonian, Latvian, and Basque, totaling 17 languages. Additionally, we focused exclusively on the morphological properties of words, excluding agreement phenomena. For example, while adjectives may be inflected to agree with the number of their head nouns, such as plural nouns requiring plural adjectives, this aspect was not tested in our

study. See Table 3.2 for the tasks that are available for each considered language and the numbers of possible values for each task.

### 3.4 Dataset

In this study, 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. In each language, each task includes a training set, a development test, and a test set. The dataset includes both frequent and infrequent words (Acs et al., 2023). All the training sets consist of 2000 samples. The development data contains 200 samples. The testing data contains of 200 samples. There are no overlapping samples between the training, testing, and development set. Each sample is a triple of a sentence, a target word, the index of the target word, and a morphological tag. The tag sets differ across morphological features. For example, for number, the tags are either `singular` or `plural`. For tense, they can be `past` or `present`. They also differ in the number of tags. For instance, in some languages, there are 3 different possible genders, namely `masculine`, `feminine`, or `neural`. Based on the currently-selected 17 languages, we ran a total of 43 morphological probing tasks, covering number, case, gender, and tense. We focused on nominal and verbal inflection. Probing tasks on number, gender, and case tasks are for nouns while tense tasks are for verbs.

### 3.5 Feature Extractions

As the first step in probing the morphological knowledge of mT5 and ByT5, we extracted the contextual 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 `base` mode. After freezing the weights of the encoder part, we then extracted the hidden states for the entire sentence. Sentence representations were then used to obtain contextual representations of the target word. Since we also aim to look at how much morphological knowledge is learned at each time step in the training process, we extracted the word representation at each hidden layer of the network, including one embedding layer and 12 Transformer layers for mT5 or 18 Transformer layers for ByT5. Each word representation is associated with a label, which is its corresponding morphological

feature. The models are available to be downloaded from the Hugging Face library at <https://huggingface.co/models>. The entire experiment was run on two NVIDIA Tesla V100 GPUs on Ponyland server available at Radboud University. The total number of experiments is 5547. They were run over the course of 20 days.

### 3.6 Probing Classifiers

Previous studies often use two architectures for probing classifiers, namely linear classifiers (Belinkov, Durrani, et al., 2020; Hupkes et al., 2018) and multilayer perceptrons (MLPs) (Adi et al., 2016; Conneau et al., 2018; Ettinger et al., 2018; Lin et al., 2019; K. Zhang & Bowman, 2018). Both types of probes have received convincing arguments. Linear classifiers indicate the amount of information that can be straightforwardly detected in the representations, thus providing a faithful indication of their linguistic knowledge (Belinkov, Durrani, et al., 2020; N. F. Liu et al., 2019). While it is argued that while MLPs may produce higher accuracy due to their ability to remember to the training data (Hewitt & Liang, 2019), some studies show that linear classifiers and MLPs produce consistent accuracy (Belinkov, Durrani, et al., 2017; Conneau et al., 2018; Qian et al., 2016). Given the debate, we used both types of classifiers in the study and compared their performance in the probing tasks.

To build probing classifiers, we first extracted the contextual embeddings from mT5 and ByT5 for 2000 target words in the training data and used these embeddings to train the probing classifiers. These classifiers were then used to predict the morphological label for words in the test set. Probing classifiers are simple linear classifiers with one hidden layer. MLPs have one hidden layer with 50 neurons. The nonlinearity function is Sigmoid. A dropout layer with a dropout probability of 0.2 was added between the hidden and the output layer to control for overfitting. The number of classes depends on the number of morphological features in the specific task. For example, the probing classifier for English number system will be a binary classifier with two possible values (i.e., singular or plural) while for the German gender system, it will be a multi-class classifier because there are 3 possible genders for each noun (i.e., masculine, feminine, and neutral). We used the Adam optimizer (Kingma & Ba, 2014) and cross entropy loss as the loss function. The learning rate is 0.001.  $\beta_1$  is set to 0.9 and  $\beta_2$  is 0.999. We chose the aforementioned learning rate after piloting the experiment with different learning rates and select the best-performing one. Probing classifiers were trained using early stopping criterion on

development losses with a stopping window of 25 epochs. The maximum number of epochs is 200 and the batch size is always 128. For each task in each language, we extracted the word embeddings of each layer and trained a separate probing classifier for each layer. In other words, each layer of the investigated LLMs were made to do the same task. After training and validating the classifiers, we had them predicting the morphological features for words in the testing set and report the accuracy scores, following standard practice in linguistic probing (Acs et al., 2023; Belinkov, Durrani, et al., 2020; Vylomova et al., 2017).

### 3.7 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 is assigned a unique embedding. Literature has suggested a couple of methods to approximate the word embedding. 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, Durrani, et al. (2020) compared both methods and found that using the embeddings of the last token produced higher accuracy scores. Since there has not been an established consensus on which pooling method best captures the encoded information, we chose to compare both methods. This may provide more insights into how morphological information is distributed across tokens or characters.

### 3.8 Controls and Evaluation

One of the challenges of using the structural probing approach is to correctly interpret the accuracy of the probing classifiers. As discussed previously, a high probing accuracy score may not imply that the model representation successfully encodes the linguistic feature (Belinkov, 2022; Hewitt & Liang, 2019). It may be the case that the probes accidentally learn the tasks rather than the representations themselves (Hewitt & Liang, 2019). Research has proposed some approaches and metrics to evaluate how the performance of the probing classifiers indicate the information learned by the model. Hewitt and Liang (2019) suggested

a method to assess the degree to which the probing classifiers reflect the model embeddings. The idea is that if the high accuracy of the probing classifiers is due to the information learned by the embeddings of the model, then they should be less accurate when classifying embeddings associated with random labels. Such classifiers are then considered *selective*. They propose *control task*. The procedure is as follows. Assuming that the task has three labels. In the control task, there are also three random labels that may not be the correct morphological feature of the target word, the embeddings extracted from the LLMs will be assigned with one of the random labels. Control tasks serve as a measure of how much information is learned from the representations compared with the probing classifiers themselves. If the accuracy scores of classifiers trained on random labels are comparable to those trained on the correct morphological labels, it means that they simply remember the training data. In contrast, if training on true labels significantly improves probing accuracy, some morphological information should be present and extractable by the classifiers. As such, the results of the control tasks are considerably worse than the results of probing tasks, it can be inferred that the representations of LLMs capture morphological knowledge to a certain extent. This approach is prevalent. However, we cannot use it for the following reasons. The goal of control tasks is to estimate how much information is encoded in the trained representations compared to how much the classifier remembers the training data. The condition for doing control tasks is that some parts of training data have to appear again in the test data. As such, we may know whether the classifier remembers the randomly-assigned label of the target token. In the dataset that we are using, the target words in the training, development, and testing sets are not repeated. Therefore, it is impossible to know if the classifier extracts relevant information from the training data or it accidentally learns the task.

As an alternative, we used fastText (Bojanowski et al., 2017), a static word embedding algorithm. There are pretrained fastText embeddings for words in 157 languages. We probed fastText embedding using the exact procedure described in Section 3.5 and 3.6, namely extracting the embedding of the target words and training probing classifiers to predict their morphological features. The crucial difference is that with mT5 and ByT5, we extracted the sentence representations and used pooling methods to get the contextual word embeddings. Here, we obtained the static word embeddings that are always constant regardless of context. fastText word vectors have an embedding dimension of 300. We ran a total of 43 probing tasks for 17 considered languages. After training and evaluating the probing classifiers on the training and development samples, we computed accuracy

scores based on the performance on the testing samples. We then compare mT5 and ByT5 performance against it. If probing classifiers trained on mT5 and ByT5 representations surpass their fastText counterpart, it would imply that their accuracy scores are reflective of their morphological representations.

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

Table 3.2 Morphological properties of 12 investigated languages (N = noun; V = verb)

## Chapter 4

# Results

In Chapter 3, we present in detail how we probe the morphological capabilities of mT5, ByT5, and the fastText baseline. We trained probing classifiers for each task in each language. For each classifiers, we trained contextual word representations and the training and development data and evaluated their accuracy in predicting morphological tags in the testing data. In this section, we present analyses on accuracy scores. We compare accuracy scores across languages, tasks, models, probing classifiers, and pooling methods and statistically investigate the effect of morphological complexity and training data on the morphological capabilities of mT5 and ByT5.

### 4.1 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 (the last two columns). Recall that we ran probing tasks on individual layers of both models. Since the hidden state of the last hidden layer shows how morphological knowledge is learned after going through the encoding process, for all analyses except for the layer-wise analysis, we used the probing accuracy of the last hidden layer. Table 4.1 shows the accuracy scores of the probing classifiers trained on the last hidden state of each probing task performed by mT5, ByT5, and the fastText baseline, grouped by languages and language families. 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. We refer readers to the Appendix for a full report of accuracy scores for each task in each language.

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

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

Table 4.1 Probing accuracy of mT5, ByT5 and fastText, averaged over languages and tasks

very small. This finding is different from that of Belinkov, Durrani, et al. (2017), who found that character-level tokenizer is better than subword tokenizers in representing morphology. Considering the differences between both LLMs and the fastText baseline, as shown in Figure 4.1, it can be seen that they generally outperformed the baseline yet perform much worse than baseline in French and Russian. In addition, mT5 achieved lower probing accuracy than fastText in Arabic and Hindi. ByT5 performed worse than fastText in German and Urdu.

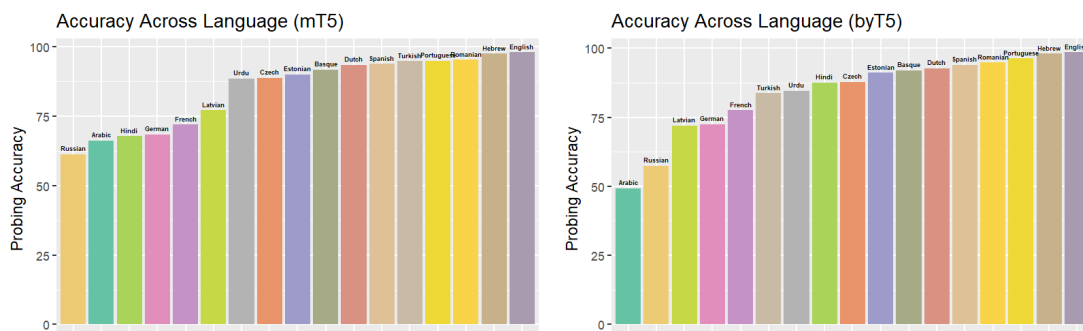


Figure 4.1 Probing accuracy of mT5 (left) and ByT5 (right) across languages

Figure 4.2 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. There are a couple of different results. mT5 scores higher in Turkish and ByT5 achieve much higher on Hindi tasks.

The difference between both LLMs and fastText also tells how contextual information affects morphological abilities. mT5 and ByT5 use contextual word representation, while fastText embeddings are static and independent of the context. 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 baseline in French and Russian. Contextual information seems to significantly hurt their performance in these two languages.

Language	Family	Model			mT5-fastText	ByT5-fastText
		mT5	ByT5	fastText		
English	Germanic	98	98.5	97.75	0.25	0.75
Dutch	Germanic	<b>93.5</b>	<b>92.75</b>	82.25	11.25	10.5
German	Germanic	68.46	72.40	68.87	-0.41	3.53
French	Romance	71.93	77.53	92.95	<b>-21.02</b>	<b>-15.42</b>
Spanish	Romance	93.83	94	71.16	22.67	22.84
Portuguese	Romance	95	96.25	88.16	6.84	8.09
Romanian	Romance	94.75	95.25	92.25	2.5	3
Hebrew	Semitic	97.5	98	92.49	5.01	5.51
Arabic	Semitic	66.17	49.25	37.81	28.36	11.44
Russian	Slavic	61.26	57.43	79.20	<b>-17.94</b>	<b>-21.77</b>
Czech	Slavic	88.79	87.79	78.81	9.98	8.98
Hindi	Indic	<b>67.83</b>	<b>87.5</b>	58.02	9.81	29.48
Urdu	Indic	88.5	84.5	74.33	14.17	10.17
Turkish	Turkic	<b>94.78</b>	<b>83.69</b>	78.28	16.5	5.41
Latvian	Baltic	77.14	72.01	73.76	3.38	-1.75
Estonian	Uralic	91.08	90.03	82.94	8.14	7.09
Basque	Basque	91.69	91.91	52.60	39.09	52.82

Table 4.2 Probing accuracy of mT5, ByT5, and fastText by languages, language families, averaged over tasks. Scores in bold indicate considerable differences between models.

Table 4.2 shows the probing accuracy scores for each language, averaged over tasks. The last two columns show the difference between each model and fastText. Negative number means that the model performs worse than baseline. 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 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 is only one task available (i.e., case).

## 4.2 Layer-Wise Analysis

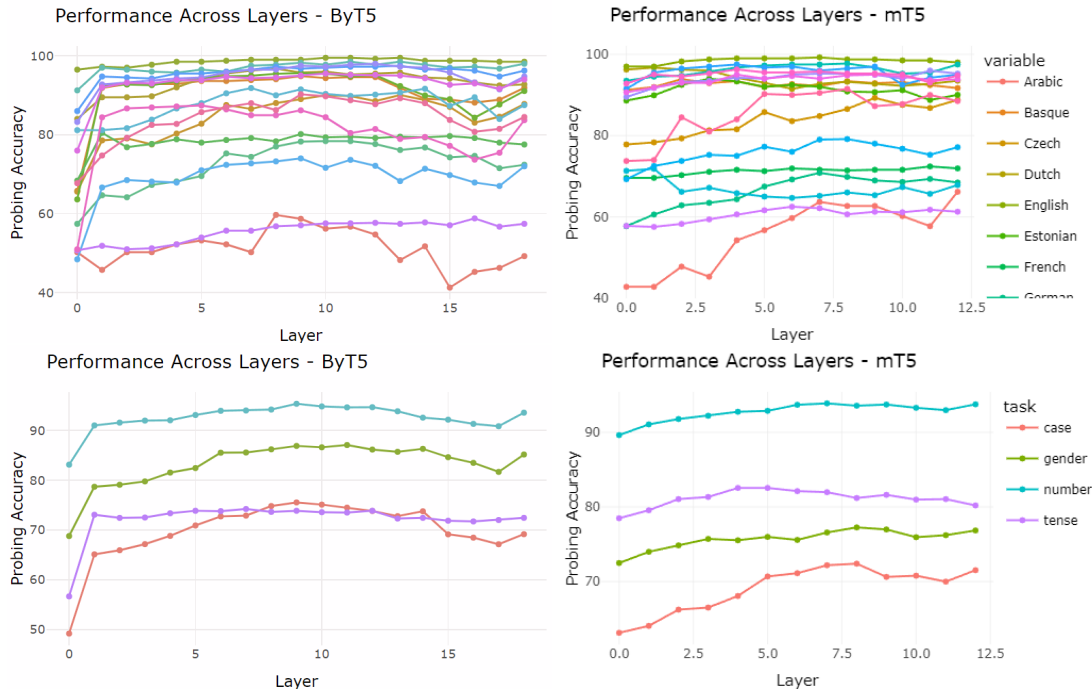


Figure 4.2 Probing accuracy of mT5 (right) and ByT5 (left) across layers grouped by languages and tasks. Each line represents a language (top) or a task (bottom). Each data point is the accuracy scores at each layer of each language.

Previous studies suggest that different layers of LLMs may encode different types of linguistic information. For example, Peters et al. (2018) and Tenney, Das, and Pavlick (2019) found that lower layers capture local information such as morphology, while syntax and semantic knowledge is encoded in the higher layers. To investigate how morphological knowledge is distributed across layers of mT5 and ByT5, we compared the probing accuracy scores of both models for individual tasks. Figure 4.2 illustrates the difference in probing accuracy between languages and between tasks for mT5 and ByT5.

Another pattern is that there are some degrees of variation between layers. In some high-performing languages, 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, Durrani, 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 most pronounced in Turkish (see the pink line in the top left plot of Figure 4.2) where the gap between the embedding layer and the first layer is approximately 33%. For French (green), we also see a huge improvement throughout layers before it peaks at layer 10. The only outlier is Arabic, yet an upward trend throughout layers until layer 8 is also observed. 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 LLMs need more layers to capture morphological patterns of languages. At task level, all tasks show the same pattern and it holds across models.

## 4.3 Effect of Model Configuration

### 4.3.1 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). Hewitt and Liang (2019) found that MLP probes tend to remember more about the training data, thus have low *selectivity* (see Section 3.8 for discussion). In this research line, studies have compared the performance of these two types of probes, namely linear probes and MLPs. Acs et al. (2023) found comparable performance while Belinkov, Durrani, et al. (2020) and Qian et al. (2016) found that they produce similar trends, yet linear classifiers achieved lower accuracy scores. Here, we also compare the accuracy scores of MLP probes and linear probes for mT5 and ByT5, as shown in Table 4.3.

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

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

Table 4.3 Probing accuracy of mT5 and ByT5 when using MLP and linear classifiers, averaged over languages and tasks

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.

### 4.3.2 Subword Pooling Methods

The approach to approximating word embeddings from token or character embeddings may have an impact on probing accuracy. Recall that we used two different methods to obtain word-level representations. We computed the mean embeddings of all subwords or characters that belong to the target word. Additionally, we considered the embedding of the last subword or character. In this section, we compare the probing accuracy when using these two pooling methods. Table 4.4 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 4.4 Probing accuracy of mT5 and ByT5 when using different subword pooling methods, namely taking the mean embeddings of all subwords or characters and taking the embedding of the last subword or character. 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 (Ács et al., 2023; Ács et al., 2021; Belinkov, Durrani, et al., 2020).

## 4.4 Effect Of Linguistic Factors

### 4.4.1 Task

Morphological knowledge includes knowledge of different features, each has distinct governing rules and syntactic relationship with other words. To investigate whether morphological features are learned differently by mT5 and ByT5, we averaged over tasks in all languages, resulting in an overall accuracy score for each task, as shown in Table 4.5. The results strongly suggest that each morphological feature is encoded differently. Interestingly, mT5 and ByT5 show somewhat different patterns. Both models perform equally well at number and worst at case. However, tense is learned better than gender by mT5 while the reverse is true for ByT5. Comparing both LLMs the baseline, it appears that they surpass the baseline in all tasks except for tense, where ByT5 performs considerably worse than fastText.

<b>Task</b>	<b>mT5</b>	<b>ByT5</b>	<b>fastText</b>
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 4.5 Probing accuracy of mT5, ByT5, and fastText, averaged over languages and tasks

Case seems to be the hardest task for both models. In this task, probing classifiers perform just slightly above random guessing. Despite the seemingly obvious reason that case often have more value than other morphological features (up to 8 in Turkish) and thus make it a more complicated task, case is 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 how individual languages, 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 in all other languages.

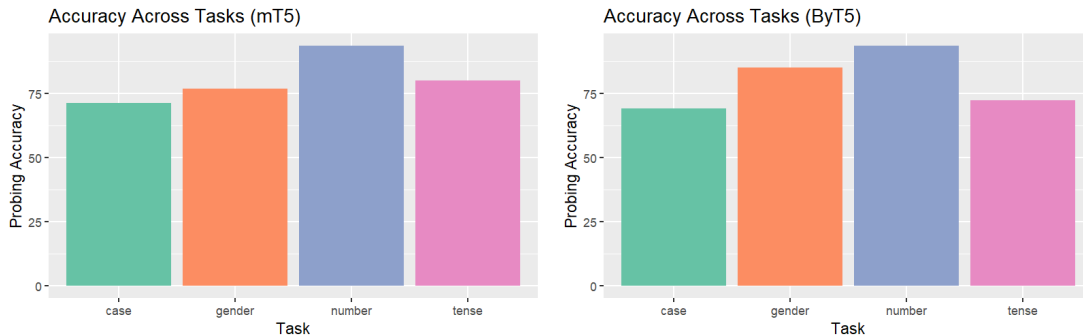


Figure 4.3 Probing accuracy of mT5 (left) and ByT5 (right) by tasks

This difference in learnability of morphological features is also reported by Acs et al. (2023), Bisazza and Tump (2018), and Edmiston (2020). They trained NMT systems using word-level tokenization and found that number is the best captured feature, followed by tense and gender.

#### 4.4.2 Morphological Complexity and Training Data

While languages are morphologically diverse, in the context of multilingual language modeling, the variety in available resources for each language is another factor that makes it different from all the other and may influence morphological learning. Warstadt et al. (2020) trained a series of RoBERTa models (Y. Liu et al., 2019) on 1M, 10M, 100M, and 1B English tokens and examined how linguistic knowledge is represented and used in those models. They found that even models with little training data (1M) and those trained on billions of tokens (30B) performed comparably well on morphological probing tasks, suggesting that morphological knowledge can be encoded with a small amount of training resources. However, by measuring the inductive bias toward linguistic knowledge of these models, they found that only those trained on over 1 billion tokens can successfully use such knowledge to develop linguistic bias. Following up, Y. Zhang et al. (2021) used multiple probing and fine-tuning methods to investigate how linguistic representation learned from various sizes of training corpora is used for downstream tasks. Comparing miniRoBERTas and RoBERTa-base, they observed that the linguistic knowledge encoded in 100M models is nearly equal to that in RoBERTa-base, which was trained on 30B words. However, in terms of downstream tasks, much more training data is needed to

improve miniRoBERTas’ performance on such tasks. While these studies investigated the effect of training resources on English LLMs. Little is known about the role of training data in multilingual settings. Wu and Dredze (2020) experimented on mBERT and found that languages with large amounts of training data outperformed low-resource languages on NLP downstream tasks, namely named entity recognition, part-of-speech tagging, and dependency parsing. Their results strongly suggest the role of training data on LLMs’ performance on downstream linguistic tasks.

Given the difference in the way morphological marking is established in different languages, it is worthwhile to analyze whether they affect how multilingual LLMs learn morphology. In this section, we explore the effects of two types of morphological complexity, namely TTR and the degree of irregularity on probing accuracy. To further examine how morphological complexity affects probing performance, we fitted a generalized mixed effect logistic regression model on the output of probing tasks, the two morphological complexity measures, and the amount of training data. For the dependent variable, we considered the output of the probing classifier on each test sample. As such, there are 200 data points for each task in each language. If the classifiers assigned the correct label to the target word in the test sample, it would be labeled 1, otherwise 0. The analysis was conducted using the `lme4` package in R. We scaled and centered all independent variables before running the statistical test. The fixed effects are *Irregularity*, *TTR*, and the percentage of *training data* of each language. We entered *languages* and *tasks* as random effects. The model is formulated as follows.

$$\text{classifier outputs} \sim \text{training data} * \text{irregularity} + \text{training data} * \text{TTR} + (1|\text{task}) + (1|\text{language})$$

Table 4.6 show the statistics of the regression models. It can be seen that both the effect of *training data* and *irregularity* on probing accuracy is significant (training data:  $\beta = 2.11$ ,  $SE = .55$ ,  $p < .001$ , irregularity:  $\beta = 2.83$ ,  $SE = .42$ ,  $p < .001$ ). In addition, there is an interaction effect between the training data ( $\beta = 2.03$ ,  $SE = .31$ ,  $p < .001$ ). These results show that there is no causal relationship between *TTR* and morphological representation of mT5 and ByT5 ( $\beta = -.62$ ,  $SE = .57$ ,  $p = .272$ ). The interaction between *TTR* and *training data* is also not significant. Our results suggest positive effects of training data size and I-complexity on the morphological representation of LLMs, as well as their interaction. The effect of training data size on probing accuracy is stronger when the degree of irregularity is higher or the morphological paradigm is more irregular. This means that the presence of high irregularity in the morphological systems amplifies the

<b>Fixed Effects</b>				
Variable	Estimate	SE	<i>t</i> -value	<i>p</i> -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
<b>Random Effect</b>				
Group	Name	Variance	Std.Dev.	
language	(Intercept)	3.0920	1.7584	
task	(Intercept)	0.3571	0.5976	

*Table 4.6* 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.

impact of training data on the morphological abilities of LLMs .

## General Discussion and Conclusions

In this study, we investigated the morphological knowledge of 17 different languages using the mT5 and ByT5 models. We employed probing classifiers on contextual word representations to quantify their morphological content. Our analysis compared morphological representations across several factors: languages, morphological features, morphological irregularity, tokenization methods, and the amount of training resources available for each language. We also compared the results of mT5 and ByT5 against fastText, a non-contextualized word representation algorithm. In this final chapter, we summarize the findings and frame them into the big picture of how morphological systems of different languages are captured by multilingual LLMs and provide some methodological implications for linguistic probing research.

**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. English, Dutch, Spanish, Portuguese, Hebrew, Romanian, and Basque achieved nearly perfect accuracy in probing tasks (higher than 90%). On the other hand, both mT5 and ByT5 performed worse at German, French, Russian, and Arabic tasks. These results somewhat contradict Edmiston (2020), who found comparable performance in all languages. Acs et al. (2023) also did not observe performance differences across languages. Rather, they found differences across parts-of-speech and morphological features. We also observed that some tasks are more difficult for the LLMs. Number is the easiest task while case is the hardest one. This difference can be partly attributed to the high number of possible features but also to the contextual-dependent nature of case. This is supported by the baseline results for case, which is much lower than both mT5 and ByT5. These results are also consistent with Bisazza and Tump (2018) and Edmiston (2020).

**Character-level models are on par with sub-word level models in representing**

**morphology** Comparing mT5 and ByT5, it can be seen that both models exhibit highly similar performances. Remarkably, despite being tokenized at the byte level, byT5 captures morphological knowledge in the early layers of the network, on par with mT5, which operates on subwords. Interestingly, we found that morphology is learned later in ByT5 than mT5. The performance gap between the embedding layer and the first layer is considerably higher in ByT5 than mT5. This indicates that although mT5 and ByT5 capture the same amount of morphological knowledge at the end, ByT5 needs more processing time to be able to be comparable with mT5. Belinkov, Durrani, et al. (2020) and Vylomova et al. (2017) tested NMT systems and found character-level tokenizers to surpass BPE in learning morphology. However, through investigating large-scale LLMs, we found no advantage of character-level tokenizer in encoding morphology. In fact, we even found that subword tokenizer even works much better for Turkish and character-level tokenizer benefits Hindi morphology. This indicates that the advantage of character-level over subword-level tokenizers may not be uniform across languages.

**Morphology is best learned in the middle to late layers** We probed all layers of mT5 and ByT5 for their morphological content. Our findings show that morphological knowledge generally improves over layers in both LLMs. There are languages in which performances are comparable across layers, namely English, Spanish, Dutch, Portuguese, Romanian, and Hebrew. 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). Acs et al. (2023) found the embedding layers to be 10 percent lower in accuracy compared to the middle layers. Edmiston (2020) found peak accuracy in the middle layer for German and Russian and similar scores across layers for English, French, and Spanish. Edmiston (2020) attributed the stable performance of these languages to their simple morphological systems. This is also true for our results. English, Spanish, Dutch, and Romanian are among the most regular languages according to the I-complexity scores by Wu et al. (2019). Our results are different from Belinkov, Màrquez, et al. (2017), Peters et al. (2018), and Tenney, Das, and Pavlick (2019), 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 mediates the effect of training data size** One of the topics that we would like to discover is the relationship between the morphological knowledge of LLMs and morphological complexity. We quantified this complexity using

two measures, namely I-complexity (irregularity) and TTR. Our logistic regression analysis reveals the significant effects of morphological irregularity and training data sizes on the performance of probing classifiers, such that the effect of training data is mediated by irregularity. When the language is highly irregular, a larger amount of training data is necessary to fully capture its morphological system.

Our study is the first to systematically examine the effects of morphological complexity on the difficulty of capturing morphological information of LLMs. Previous studies have tested the effects of morphological complexity on modeling difficulty, yet modeling difficulty is measured differently. Cotterell et al. (2018) and Mielke et al. (2019) used sentence surprisal, Gerz et al. (2018) used perplexity. Here, we present the first evidence that the effect of morphological irregularity is present at the representation level. However, it can be tricky to compare our results with theirs because we examined the amount of morphological knowledge embedded in the hidden representation while their measure is on a higher level (i.e., the difficulty when predicting a sentence) and are computed after the training process. Another reason which leads to the difference between their results and ours is the measure of morphological complexity, which varies across studies. Mielke et al. (2019) and Gerz et al. (2018) correlated modeling difficulty with Morphological Counting Complexity (MCC) (Sagot, 2013), vocabulary sizes of languages, and dependency length. They found vocabulary sizes to be the best predictor. Studies on the effect of training sizes show that its effect is not present at the representation level yet at the downstream level (Warstadt et al., 2020; K. Zhang & Bowman, 2018). These studies show that factors like training data and morphological complexity may affect LLMs differently at different training steps. As such, our results do not eliminate the possibility that the effect of morphological complexity is not present after the end-to-end training process. Thus, an interesting topic for future research is to investigate how the effect of morphological complexity is manifested in other levels of representations, such as in downstream performance. For example, Dang et al. (2024) also found the effect of irregularity on the abilities of LLMs to generalize their morphological knowledge to nonce words.

**MLPs perform as well as linear probes in morphological probing tasks** Our results also provide methodological implications for linguistic probing practice. First, the use of probing classifiers to investigate the linguistic content of language models has received critical concern Belinkov (2022) and Hewitt and Liang (2019). Hewitt and Liang (2019) showed that MLPs remember much more about the training data than linear probes. As such, they may tell more about the probes than the representations. We implemented

and compared different variations of probing classifiers. Our results, however, align more closely with Acs et al. (2023) and Qian et al. (2016) that both MLPs and linear probes achieve consistent performance. This shows that probing results are not affected by the type of probing classifiers. Second, there has been little consensus over subword pooling methods. Acs et al. (2023) and Belinkov, Durrani, et al. (2020) found that using the last token (or character) embedding is better than the embedding of the first token and the mean embedding in encoding morphology. Our results further support theirs. We found that the last token embedding is the best representative for both mT5 and ByT5. They may accumulate all information processed with all previous tokens or characters of the word.

**Limitations** Our study provides evidence for disparity in morphological knowledge of LLMs and the mediated effect of morphological irregularity on such knowledge. However, it has several limitations that should be taken into account. Firstly, we ran the experiment only once due to the high volume of experiments. In linguistic probing, it is standard practice to run experiments multiple times to control for randomness. For instance, Acs et al. (2023) averaged accuracy scores over 10 runs, while Edmiston (2020) reported the best score out of 10 runs. Although we followed established procedures, our single-run approach may introduce some degree of variability in the results. Secondly, while we aim to cover as many typologically different languages as possible, the dataset that we use supports mostly Indo-European languages because there is not enough sampling data for other languages. Therefore, we limited ourselves to testing the morphological knowledge of European languages only. 12 out of 17 languages that we investigated are Indo-European languages. For example, we did not investigate the Celtic, Japonic, and Sino-Tibetan families. Moreover, the current study only focuses on inflectional morphology. This decision was made also because of the lack of datasets for probing knowledge of derivational morphology. Another limitation lies in our statistical test. Our sample is relatively large (17200 in total). Large samples may result in small p-value even when there is no effect (Søgaard et al., 2014). As such, significant results in our model should be interpreted with caution. Further research is needed to generalize beyond our suggestive evidence.

# List of Figures

- 3.1 The difference in tokenization method of ByT5 and mT5 from Xue et al. (2022) . . . . . 21
  
- 4.1 Probing accuracy of mT5 (left) and ByT5 (right) across languages . . . . . 31
- 4.2 Probing accuracy of mT5 (right) and ByT5 (left) 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. . . . . 33
- 4.3 Probing accuracy of mT5 (left) and ByT5 (right) by tasks . . . . . 37

# List of Tables

- 2.1 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. . . . . 15
- 2.2 An Example, taken from Belinkov, Durrani, et al. (2020) of how Morfessor (Smit et al., 2014), BPE (Sennrich et al., 2016), SentencePiece (Kudo & Richardson, 2018), and byte-level tokenizer (Xue et al., 2022) tokenize the sentence *"Professor admits to shooting his girlfriend"*. . . . . 17
- 3.1 Configuration of mT5 and ByT5, from Xue et al. (2022) (Param = Number of parameters, Enc. = Number of encoder layers; Dec. = Number of decoder layers; Embed. Dim. = Embedding dimensions) . . . . . 20
- 3.2 Morphological properties of 12 investigated languages (N = noun; V = verb) 29
- 4.1 Probing accuracy of mT5, ByT5 and fastText, averaged over languages and tasks . . . . . 31
- 4.2 Probing accuracy of mT5, ByT5, and fastText by languages, language families, averaged over tasks. Scores in bold indicate considerable differences between models. . . . . 32
- 4.3 Probing accuracy of mT5 and ByT5 when using MLP and linear classifiers, averaged over languages and tasks . . . . . 35
- 4.4 Probing accuracy of mT5 and ByT5 when using different subword pooling methods, namely taking the mean embeddings of all subwords or characters and taking the embedding of the last subword or character. The results were averaged over languages and tasks and the reported probes are MLPs 35
- 4.5 Probing accuracy of mT5, ByT5, and fastText, averaged over languages and tasks . . . . . 36

4.6	Results of linear mixed effect regression with <i>probing accuracy</i> of mT5 as outcome variable and <i>language</i> and <i>task</i> as random effects. Fixed effects are irregularity (I) and TTR (T), training data (TD) and their two-way interactions. . . . .	39
5.1	Accuracy scores of each task in each language from mT5, ByT5, and fastText	59

# Bibliography

- Ackerman, F., & Malouf, R. (2013). Morphological organization: The low conditional entropy conjecture. *Language*, 89(3), 429–464. <https://doi.org/10.1353/lan.2013.0054>
- Acs, J., Hamerlik, E., Schwartz, R., Smith, N. A., & Kornai, A. (2023). Morphosyntactic probing of multilingual bert models. *Natural Language Engineering*, 1–40.
- Ács, J., Kádár, Á., & Kornai, A. (2021). Subword pooling makes a difference. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2284–2295.
- Adams, O., Makarucha, A., Neubig, G., Bird, S., & Cohn, T. (2017). Cross-lingual word embeddings for low-resource language modeling. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 937–947.
- Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., & Goldberg, Y. (2016). Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*.
- Ahmed, T., & Devanbu, P. (2022). Few-shot training llms for project-specific code-summarization. *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, 1–5.
- Ali, M., Fromm, M., Thellmann, K., Rutmann, R., Lübbering, M., Leveling, J., Klug, K., Ebert, J., Doll, N., Buschhoff, J. S., et al. (2023). Tokenizer choice for llm training: Negligible or crucial? *arXiv preprint arXiv:2310.08754*.
- Belinkov, Y. (2022). Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1), 207–219.
- Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., & Glass, J. (2017). What do neural machine translation models learn about morphology? *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 861–872.

- Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., & Glass, J. (2020). On the linguistic representational power of neural machine translation models. *Computational Linguistics*, 46(1), 1–52.
- Belinkov, Y., Gehrmann, S., & Pavlick, E. (2020, July). Interpretability and analysis in neural NLP. In A. Savary & Y. Zhang (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics: Tutorial abstracts* (pp. 1–5). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-tutorials.1>
- Belinkov, Y., Màrquez, L., Sajjad, H., Durrani, N., Dalvi, F., & Glass, J. (2017). Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1–10.
- Bentz, C., Verkerk, A., Kiela, D., Hill, F., & Buttery, P. (2015). Adaptive communication: Languages with more non-native speakers tend to have fewer word forms. *PloS one*, 10(6), e0128254.
- Berko, J. (1958). The child’s learning of english morphology. *Word*, 14(2-3), 150–177.
- Bisazza, A., & Tump, C. (2018). The lazy encoder: A fine-grained analysis of the role of morphology in neural machine translation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2871–2876.
- Bloomfield, L. (1933). *Language*. H. Holt.
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5, 135–146.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877–1901.
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.
- Chen, A., Schwartz-Ziv, R., Cho, K., Leavitt, M. L., & Saphra, N. (2023). Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in mlms. *arXiv preprint arXiv:2309.07311*.
- Chung, J., Cho, K., & Bengio, Y. (2016). A character-level decoder without explicit segmentation for neural machine translation. *Proceedings of the 54th Annual Meeting*

- of the *Association for Computational Linguistics (Volume 1: Long Papers)*, 1693–1703.
- Clark, J. H., Garrette, D., Turc, I., & Wieting, J. (2022). Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10, 73–91.
- Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What does bert look at? an analysis of bert’s attention. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 276–286.
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451.
- Conneau, A., Kruszewski, G., Lample, G., Barrault, L., & Baroni, M. (2018). What you can cram into a single &!#\* vector: Probing sentence embeddings for linguistic properties. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2126–2136.
- Cotterell, R., Kirov, C., Hulden, M., & Eisner, J. (2019). On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics*, 7, 327–342.
- Cotterell, R., Mielke, S. J., Eisner, J., & Roark, B. (2018). Are all languages equally hard to language-model? *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 536–541.
- Dang, A., Galke, L., & Raviv, L. (2024). Morphology matters: Probing the cross-linguistic morphological generalization abilities of large language models through a wug test. *To appear in proceedings of the 13th edition of the Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2024)*.
- DeKeyser, R. M. (2005). What makes learning second-language grammar difficult? a review of issues. *Language learning*, 55.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Dryer, M. S., & Haspelmath, M. (2013). Wals online (v2020. 3). *Zenodo* <https://doi.org/10.5281/zenodo.7385533>.

- Durán, P., Malvern, D., Richards, B., & Chipere, N. (2004). Developmental trends in lexical diversity. *Applied Linguistics*, 25(2), 220–242.
- Edman, L., Sarti, G., Toral, A., Noord, G. v., & Bisazza, A. (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.
- Edmiston, D. (2020). A systematic analysis of morphological content in bert models for multiple languages. *arXiv preprint arXiv:2004.03032*.
- Ettinger, A. (2020). What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8, 34–48.
- Ettinger, A., Elgohary, A., Phillips, C., & Resnik, P. (2018). Assessing composition in sentence vector representations. *Proceedings of the 27th International Conference on Computational Linguistics*, 1790–1801.
- Fleshman, W., & Van Durme, B. (2023). Toucan: Token-aware character level language modeling. *arXiv preprint arXiv:2311.08620*.
- Galke, L., Ram, Y., & Raviv, L. (2023). What makes a language easy to deep-learn? *arXiv preprint arXiv:2302.12239*.
- Gandhe, A., Metze, F., & Lane, I. (2014). Neural network language models for low resource languages. *Fifteenth Annual Conference of the International Speech Communication Association*.
- Gao, Y., Nikolov, N. I., Hu, Y., & Hahnloser, R. H. (2020, July). Character-level translation with self-attention. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 1591–1604). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.145>
- Gerz, D., Vulić, I., Ponti, E., Naradowsky, J., Reichart, R., & Korhonen, A. (2018). Language modeling for morphologically rich languages: Character-aware modeling for word-level prediction. *Transactions of the Association for Computational Linguistics*, 6, 451–465.
- Goldman, O., & Tsarfaty, R. (2022). Morphology without borders: Clause-level morphology. *Transactions of the Association for Computational Linguistics*, 10, 1455–1472.
- Hewitt, J., Ethayarajh, K., Liang, P., & Manning, C. D. (2021). Conditional probing: Measuring usable information beyond a baseline. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1626–1639.

- Hewitt, J., & Liang, P. (2019). Designing and interpreting probes with control tasks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2733–2743.
- Hofmann, V., Pierrehumbert, J., & Schütze, H. (2021). Superbizarre is not superb: Derivational morphology improves bert’s interpretation of complex words. *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)*, 3594–3608.
- Hupkes, D., Veldhoen, S., & Zuidema, W. (2018). Visualisation and ‘diagnostic classifiers’ reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61, 907–926.
- Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al. (2017). Google’s multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5, 339–351.
- Kempe, V., & Brooks, P. J. (2008). Second language learning of complex inflectional systems. *Language Learning*, 58(4), 703–746.
- Kim, Y., Jernite, Y., Sontag, D., & Rush, A. (2016). Character-aware neural language models. *Proceedings of the AAAI conference on artificial intelligence*, 30(1).
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Köhn, A. (2015). What’s in an embedding? analyzing word embeddings through multilingual evaluation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2067–2073.
- Kudo, T., & Richardson, J. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 66–71.
- Lample, G., & Conneau, A. (2019). Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.
- Lee, J., Cho, K., & Hofmann, T. (2017). Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5, 365–378.

- Lin, Y., Tan, Y. C., & Frank, R. (2019). Open sesame: Getting inside bert’s linguistic knowledge. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 241–253.
- Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4, 521–535.
- Liu, L., & Hulden, M. (2022). Can a transformer pass the wug test? tuning copying bias in neural morphological inflection models. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 739–749.
- Liu, N. F., Gardner, M., Belinkov, Y., Peters, M. E., & Smith, N. A. (2019). Linguistic knowledge and transferability of contextual representations. *Proceedings of NAACL-HLT*, 1073–1094.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Luong, M.-T., Sutskever, I., Le, Q. V., Vinyals, O., & Zaremba, W. (2014). Addressing the rare word problem in neural machine translation. *arXiv preprint arXiv:1410.8206*.
- Madsen, A., Reddy, S., & Chandar, S. (2022). Post-hoc interpretability for neural nlp: A survey. *ACM Computing Surveys*, 55(8), 1–42.
- Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48), 30046–30054.
- McCoy, T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3428–3448.
- Mielke, S. J., Cotterell, R., Gorman, K., Roark, B., & Eisner, J. (2019). What kind of language is hard to language-model? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4975–4989.
- Mikhailov, V., Serikov, O., & Artemova, E. (2021). Morph call: Probing morphosyntactic content of multilingual transformers. *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, 97–121.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

- Park, H. H., Zhang, K. J., Haley, C., Steimel, K., Liu, H., & Schwartz, L. (2021). Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, 9, 261–276.
- Peters, M. E., Neumann, M., Zettlemoyer, L., & Yih, W.-t. (2018). Dissecting contextual word embeddings: Architecture and representation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1499–1509.
- Qian, P., Qiu, X., & Huang, X.-J. (2016). Investigating language universal and specific properties in word embeddings. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1478–1488.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 9.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140), 1–67.
- Raviv, L., de Heer Kloots, M., & Meyer, A. (2021). What makes a language easy to learn? a preregistered study on how systematic structure and community size affect language learnability. *Cognition*, 210, 104620.
- Rogers, A., Kovaleva, O., & Rumshisky, A. (2021). A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8, 842–866.
- Roumeliotis, K. I., Tselikas, N. D., & Nasiopoulos, D. K. (2024). Llms in e-commerce: A comparative analysis of gpt and llama models in product review evaluation. *Natural Language Processing Journal*, 6, 100056.
- Ruder, S., Vulić, I., & Søgaard, A. (2019). A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65, 569–631.
- Sagot, B. (2013). Comparing complexity measures. *Computational approaches to morphological complexity*.
- Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1715–1725.
- Shapiro, N., Paullada, A., & Steinert-Threlkeld, S. (2021). A multilabel approach to morphosyntactic probing. *Findings of the Association for Computational Linguistics: EMNLP 2021*, 4486–4524.

- Smit, P., Virpioja, S., Grönroos, S.-A., & Kurimo, M. (2014). Morfessor 2.0: Toolkit for statistical morphological segmentation. *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 21–24.
- Søgaard, A., Johannsen, A., Plank, B., Hovy, D., & Alonso, H. M. (2014). What’s in a p-value in nlp? *Proceedings of the eighteenth conference on computational natural language learning*, 1–10.
- Stanczak, K., Ponti, E., Hennigen, L. T., Cotterell, R., & Augenstein, I. (2022). Same neurons, different languages: Probing morphosyntax in multilingual pre-trained models. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1589–1598.
- Tenney, I., Das, D., & Pavlick, E. (2019). Bert rediscovers the classical nlp pipeline. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4593–4601.
- Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Van Durme, B., Bowman, S. R., Das, D., et al. (2019). What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.
- Toporkov, O., & Agerri, R. (2024). On the role of morphological information for contextual lemmatization. *Computational Linguistics*, 1–35.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, 32, 323–352.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- Voita, E., Talbot, D., Moiseev, F., Sennrich, R., & Titov, I. (2019). Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5797–5808.

- Vylomova, E., Cohn, T., He, X., & Haffari, G. (2017). Word representation models for morphologically rich languages in neural machine translation. *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, 103–108.
- Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 353–355.
- Wang, L., Lyu, C., Ji, T., Zhang, Z., Yu, D., Shi, S., & Tu, Z. (2023). Document-level machine translation with large language models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 16646–16661.
- Warstadt, A., Zhang, Y., Li, X., Liu, H., & Bowman, S. (2020). Learning which features matter: Roberta acquires a preference for linguistic generalizations (eventually). *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 217–235.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35, 24824–24837.
- Weissweiler, L., Hofmann, V., Kantharuban, A., Cai, A., Dutt, R., Hengle, A., Kabra, A., Kulkarni, A., Vijayakumar, A., Yu, H., et al. (2023). Counting the bugs in chatgpt’s wugs: A multilingual investigation into the morphological capabilities of a large language model. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6508–6524.
- Wu, S., Cotterell, R., & O’Donnell, T. (2019). Morphological irregularity correlates with frequency. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5117–5126.
- Wu, S., & Dredze, M. (2020). Are all languages created equal in multilingual bert? *Proceedings of the 5th Workshop on Representation Learning for NLP*, 120–130.
- Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M., Roberts, A., & Raffel, C. (2022). Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10, 291–306.
- Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., & Raffel, C. (2021, June). MT5: A massively multilingual pre-trained text-to-text transformer. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, & Y. Zhou (Eds.), *Proceed-*

- ings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 483–498). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.naacl-main.41>
- Zhang, K., & Bowman, S. (2018). Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 359–361.
- Zhang, Y., Warstadt, A., Li, H.-S., & Bowman, S. R. (2021). When do you need billions of words of pretraining data? *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.

# Appendix

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

No.	Language	Task	mT5	ByT5	fastText
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 5.1* Accuracy scores of each task in each language from mT5, ByT5, and fastText