# Morphology Matters: Probing the Cross-linguistic Morphological Generalization Abilities of Large Language Models through a Wug Test

**Anh Dang**
LEADS group, Max Planck Institute for Psycholinguistics, Nijmegen, NL
CLS, Radboud University, NL
Utrecht University, NL
thithaoanh.dangthithaoanh@ru.nl

**Limor Raviv**
LEADS group, Max Planck Institute for Psycholinguistics, Nijmegen, NL
cSCAN, University of Glasgow, UK
limor.raviv@mpi.nl

**Lukas Galke**
LEADS group, Max Planck Institute for Psycholinguistics, Nijmegen, NL
lukas.galke@mpi.nl

## Abstract

We develop a multilingual version of the Wug Test, an artificial word completion experiment that is typically used to test the morphological knowledge of children, and apply it to the GPT family of large language models (LLMs). LLMs' performance on this test was evaluated by native speakers of six different languages, who judged whether the inflected and derived forms generated by the models conform to the morphological rules of their language. Our results show that LLMs can generalize their morphological knowledge to new, unfamiliar words, but that their success in generating the "correct" generalization (as judged by native human speakers) is predicted by a language's morphological complexity (specifically, integrative complexity). We further find that the amount of training data has surprisingly little on LLMs' morphological generalization abilities within the scope of the analyzed languages. These findings highlight that "morphology matters", and have important implications for improving low-resource language modeling.

## 1 Introduction

Large language models (LLMs) have been very successful in learning and generating grammatically-correct language as humans do (Brown et al., 2020; OpenAI, 2023). This poses the question of whether they actually have linguistic capability that would allow them to generalize beyond the training distribution (Hupkes et al., 2023). In addition, does this capability manifest differently in different languages that LLMs were trained on? Here, we investigate whether LLMs' linguistic knowledge with respect to morphology differs between languages. Specifically, we test the ability of multilingual LLMs to generalize their morphological knowledge to nonce words in six languages.

Testing cross-linguistic differences in the morphosyntactic abilities of LLMs trained on large amounts of human-generated text is particularly interesting given recent findings on the behavioral similarity between humans and language models in a variety of language learning and processing tasks (Galke et al., 2023; Webb et al., 2023; Srikant et al., 2022) and syntactic structure in the models' learned attention patterns (Manning et al., 2020; Chen et al., 2023). One of the key concerns of contemporary efforts in language modeling is to improve the ability to generalize well across the variety of human languages, especially regarding low-resource languages (e.g., Schäfer et al., 2024; Zheng et al., 2022; Hedderich et al., 2021; Lauscher et al., 2020; Conneau et al., 2020).

Given the importance of the training data to LLM's abilities (Kandpal et al., 2023), conventional wisdom would suggest that the amount of exposure to a given language would be the dominant factor in determining the models' ability to learn the language's morphological patterns. Here, we argue that factors beyond the amount of training data play an important role for LLMs' generalization abilities, and in particular suggest that languages' morphological complexity needs to be taken into account. Notably, languages vary in their degree of morphological complexity (Dryer and Haspelmath, 2013; Evans and Levinson, 2009; Hengeveld and Leufkens, 2018; Ackerman and Malouf, 2013), for



*Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 177–188

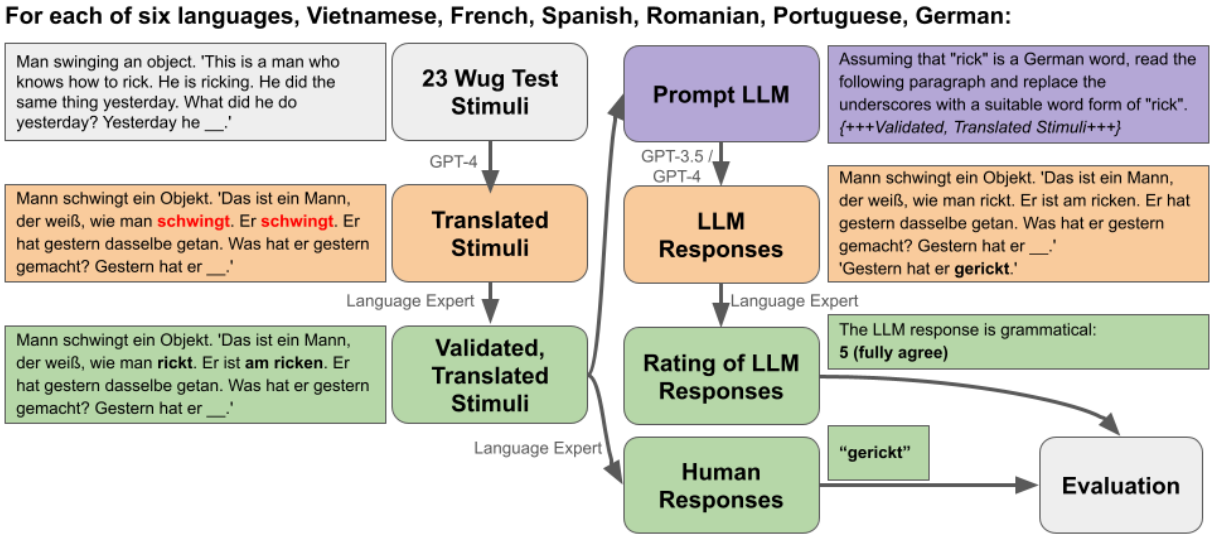


Figure 1: Overview of our experimental procedure with exemplary data and the employed prompt pattern

example in the number of morphological inflection paradigms and their degree of irregularity.

A recent study has shown that LLMs, like humans, are particularly sensitive to the degree of compositional linguistic structure in their input when generating novel forms to new meanings in a matched experiment using a miniature artificial language, with higher degrees of compositionality leading to more systematic generalizations and to a higher agreement with humans (Galke et al., 2023). This finding implies that the morphological learning ability of LLMs across different human languages should similarly be affected by languages' degree of systematic morphological structure, as quantified by measures from typological linguistics (Bentz et al., 2016; Baerman et al., 2015). In the current paper, we test to what extent languages with more systematic structures are indeed learned better by LLMs using an established morphological knowledge test used in the field of child language acquisition: the Wug-test (Berko, 1958).

Even though morphology is heavily studied in the field of computational linguistics (e.g. Batsuren et al., 2022; Wu et al., 2019; Wilson and Li, 2021; Liu and Mao, 2016), and despite its importance to human language learning (Kempe and Brooks, 2008; DeKeyser, 2005; Dressler, 2003, 2010; Slobin, 1985; Raviv et al., 2021), there is little work on the cross-linguistic morphological knowledge of LLMs, especially with respect to the potential effect of languages' morpho-syntactic structure (Weissweiler et al., 2023). Rather, it has been found that LLMs often fail to generate the correct inflected forms of words that were not a part of their training data, regardless of the size of the training set (Liu and Hulden, 2022). Given that only one study to date has probed LLMs' morphological generalization abilities with a multilingual variant of the Wug test (Weissweiler et al., 2023), it is currently unclear to what extent can LLMs generalize their morphological knowledge to new contexts, and to what extent their generalization capabilities are affected by the morphological complexity of language compared to its representation in the training data. Here, we take one step further in this line of work and test the relationship between languages' morphological structure and the generalization ability of multilingual LLMs.

Specifically, as shown in Figure 1, we develop a multilingual version of the Wug Test, an artificial word completion test that is typically used to probe the morphological knowledge of children with respect to inflectional and derivational morphology (Berko, 1958), and apply it to the GPT family of large language models (Brown et al., 2020; Ouyang et al., 2022). We consider six different languages, namely German, Vietnamese, Portuguese, Spanish, French, and Romanian, which vary in their degree of morphological complexity based on several established measures (Lupyan and Dale, 2010; Bentz et al., 2015). For each language, we first employed GPT-4 to translate 23 questions with nonce words from the original Wug Test. The translations were then evaluated and corrected by linguistically-trained native speakers, and the nonce words were adapted to fit each language's phonotactic rules. LLMs were then provided with the translations as prompts (e.g., "This is a Wug.

Now there are two of them. There are two __"), and were prompted to generate the missing inflected form (e.g,. "wugs").

Since the nonce words are new, unfamiliar words, the models need to generalize their morphological knowledge beyond their training data. The model responses were then evaluated by native speakers, who judged whether the inflected and derived forms generated by the LLMs conform to their native language's morphological rules. We then tested LLMs' generalization success across languages against two measures of morphological complexity, namely, the richness of the morphological system and how irregular it is.

In sum, our contributions are

- A multilingual version of the Wug Test for 6 languages
- A human evaluation of GPT-3.5 and GPT-4 responses on this multilingual Wug Test
- A cross-linguistic analysis linking LLM performance to morphological complexity
- An error analysis revealing new patterns of failure modes in morphological generalization

## 2 Related Work

**Morphological capabilities of LLMs** Probing machine learning models for linguistic information is a long-standing endeavour (e.g., Conneau et al., 2018; Jawahar et al., 2019; Manning et al., 2020; Warstadt et al., 2020; Rogers et al., 2021; Zhang et al., 2022; Irwin et al., 2023). In terms of morphological capabilities, Liu and Hulden (2022) conducted a Wug-like test with Transformer models (Vaswani et al., 2017), such as the ones underlying LLMs (but trained from scratch), using the SIGMORPHON 2018 shared task (Cotterell et al., 2018a), and found that models struggled to generalize morphological knowledge to new words.

However, to date, there is only one study that assessed the morphological generalization of LLMs to nonce words: Weissweiler et al. (2023), who also took inspiration from the Wug test (Berko, 1958) and prompted ChatGPT with morphological tasks in 4 different languages. The authors created a new dataset by modifying and re-annotating UniMorph 4.0 (Batsuren et al., 2022), and LLMs were prompted to fill in the blank in example sentences. While instructing LLMs to only emit the inflected form, the first word of the generated response was then compared against human responses and supervised morphology models: the affix rule learner (Liu and Mao, 2016) and the minimal generalization learner (Wilson and Li, 2021). Their results showed that GPT-3.5 is not yet on par with humans regarding its generalization performance on nonce words and also underperforms supervised morphology models.

Our work complements this endeavour in several aspects: First, we opt to manually evaluate every response from the LLMs instead of an automated evaluation strategy (which took on the first word of the model response). Second, we analyze a different set of languages, with only German overlapping across studies. And third, we test the impact of other important factors such as languages' morphological complexity scores.

**The effect of morphological complexity on language modeling** Some studies have explored the relationship between morphological complexity and the learnability of languages by LLMs, but show mixed results. Cotterell et al. (2018b) estimated the predictability of text in a parallel corpora of 21 languages, and found that text in languages with rich inflectional morphology (and thus higher word entropy) was more difficult to predict by n-gram language models and LSTM-based language models. However, when Mielke et al. (2019) use a similar approach with three times more languages and more diverse language families, they did not find a correlation between prediction difficulty and the number of inflectional distinctions that languages have.

Gerz et al. (2018) further showed a positive correlation between multilingual language models' perplexity (how well a language model is able to predict the next work) and type/token ratios (i.e., the ratio between the number of word types and the total number of tokens in the text). More recently, Park et al. (2021) used an even larger parallel corpus of 92 languages, and incorporated more measures of morphological complexity – including corpus-based measures and features from the World Atlas of Languages Structure (WALS) (Dryer and Haspelmath, 2013). Using surprisal as an estimate for difficulty, they found that models' performance was correlated with several complexity measures, and that this correlation was stronger for language models whose tokenizer relied on byte-pair-encoding (Sennrich et al., 2015).

Together, these studies imply that language learnability by LLMs is potentially affected by at least some of the specific morphological features of languages, though which features (and which metrics can capture them) is largely unknown – a question on which we aim to shed new light here.

## 3 Background on measuring morphological complexity

Languages vary in the degree of morphological complexity, which can be measured using a variety of tools (Dryer and Haspelmath, 2013) and dimensions (Ackerman and Malouf, 2013). Morphological complexity measures can be categorized into integrative complexity (I-complexity) and enumerative complexity (E-complexity) (Ackerman and Malouf, 2013). E-complexity refers to the number of cases and inflectional paradigms that exist in a language's grammar. The more inflected forms a language can have (e.g., for gender, number, tense, case, mood etc.), the higher its E-complexity score is. I-complexity refers to the predictability of inflected form from its context. The more irregular a morphological paradigm is (e.g., many verbs in English show an irregular past tense inflection), i. e. how often irregular forms are used, the higher the I-complexity score.

A well-known example of an E-complexity measure is Lupyan and Dale (2010)'s measure for morphological complexity, which was based on 28 morphological features extracted from the World Atlas of Language Structure (WALS, Dryer and Haspelmath, 2013), such as the number of inflectional distinctions. For I-complexity, Wu et al. (2019) introduced an information-theoretic measure to quantify the frequency of irregular forms, and Bentz et al. (2015) proposed three measures to capture the the variety of word types used to encode identical information ("lexical diversity"). These measures include type-token ratio and Shannon entropy ($H$), which measures the degree of uncertainty of words. The last measure is the Zipf-Mandelbrot parameter ($\alpha$), based on the Zipf's law of word distribution. Languages with higher TTR and Shannon entropy are more lexically diverse, and languages with a higher Zipf parameter are less lexically diverse.

For our study, we chose one representative measure for E-complexity and one for I-complexity, relying on previous comparative work that showed that different measures are highly correlated (Cöltekin and Rama, 2023; Bentz et al., 2016). For E-complexity, we use Lupyan and Dale (2010)'s complexity measure based on WALS features (Dryer and Haspelmath, 2013). For I-complexity, we use Bentz et al. (2015)'s entropy-based measure $H$.

## 4 Methodology

### 4.1 Input Languages

Bloomfield (1933) distinguished between four types of languages with respect to morphological structure. In our study, we consider two of them: inflected languages and isolating languages. While Spanish, German, French, Romanian, and Portuguese are highly inflected languages, Vietnamese is an isolating language which does not have explicit grammatical markers within word boundaries. We briefly describe the considered languages below.

**Vietnamese** is an *isolating language* and thus there are no bound morphemes in the form of suffixes and affixes. As such, there are no inflectional or derivational processes. Instead, semantic and grammatical information is expressed using free morphemes (i.e., standalone words). For instance, Vietnamese does not have plural word forms, but instead expresses plurality by adding a number word before the noun.

**French** is an *inflected* language in the Romance branch. Verbal inflection is used to indicate tense, person, number, mood, and aspect. Verbs are inflected such that they agree with the subject in terms of person and number. For example, the past tense formation process in French includes combining the correct conjugated form of the auxiliary verbs and the participle form of the main verb, which is formed by adding the correct ending morpheme to it. Nouns carry number and grammatical gender, with number being governed by the endings of the nous.

**Spanish** is also *inflected* language belonging to the Romance language family, which also includes French, Portuguese and Romanian. The choice of morphemes is governed by grammatical gender when inflecting nouns, pronouns, and adjectives. Verbs are conjugated differently depending on whether the endings of the infinitive forms are *-ar*, *-er*, or *-ir*. They also include inflectional agreement with the person and number of the subject. Another characteristic of Spanish and other Romance languages is that it has fusional morphology, such that a single word form can expresses various grammatical features.

**Romanian**, as another member of the Romance family, is also highly inflected language with both nominal and verbal inflection, indicating a wide range of grammatical features. The inflected forms of nouns and adjectives are determined by the grammatical gender of the nouns as well as their endings. For verbss, there are 4 conjugation classes, depending on the endings of the infinitive forms.

**Portuguese** is an Indo-European language in the Romance branch, Portuguese is also an inflectional language that bears similarity to Spanish, although the exact number of possible distinctions/inflections and the degree of irregularity is different. For certain word endings (e.g., *-s* or *-z*), plural and singular Portuguese forms are the same.

**German** is an *inflected and fusional language* where affixes are added to the stem to convey grammatical information. such as number, case, aspect, and gender. There can be several affixes that encode the same grammatical information. The choice of affixes usually depends on the gender of the noun. If it is masculine, plurality is often expressed by adding *-e*. Feminine nouns often end with *-en*. However, there are many additional rules in German, often involving changing the vowel to an umlaut (e.g., plural of "Zug" is "Züge").

### 4.2 The Wug Test in Different Languages

The Wug test (Berko, 1958) was originally designed to test the morphological knowledge of children. It tests knowledge of both inflectional morphology and derivational morphology in English. In 23 out of 28 questions, children hear a nonce word embedded in the context of an utterance, and need to complete the utterance with the nonce word's inflected form. The questions test knowledge of a wide range of morphological features, including numbers, tenses, diminutive, possessive, and derivation inflections.

We first used GPT-4 to translate the original Wug-test questions from English into the considered 6 different languages (see Figure 1). Then, to ensure the translation is correct, we had language experts (linguistically-trained native speakers) evaluate the machine-translated questions and correct the translation if necessary. In addition, we adjusted the nonce words to fit the phonotactic rules of the language, according to feedback from the language experts. That is, in many languages word have certain rules regarding the combination of different sounds. For example, French verbs must end with *-er* or *-ir*, while a consonant cluster like *zmrzl* would be unacceptable in English but fine in Czech. Thus, we also asked native speakers to modify the original nonce words so that they become phonotactically valid in the corresponding language. If there were any words that already existed in the language, we also removed those from the test.

After checking all translations, we prompted the LLMs to complete the Wug test in each of the 6 languages. Specifically, we consider GPT-4 (OpenAI, 2023) (version: `gpt4-0613` and GPT-3.5 (Ouyang et al., 2022) (version: `gpt-3.5-turbo-0613`). Since these models are fine-tuned on instruction-following, they can deal with prompts that are phrased as an instruction (Ouyang et al., 2022), we opted to prompt the language models in a zero-shot way, i. e., without supplying similar examples. We did not provide an explicit instruction about which word form that should be generated (e.g., past or plural) such that the LLMs have to infer that information from the context, yet instruct the model to assume that the nonce word is a word of the respective language.

Specifically, we employed the following English-language prompt prefix "Assuming that "{word}" is a {language} word, read the following paragraph and replace the underscores with a suitable word form of "{word}"" to each question (see Figure 1). We repeat this procedure to have the two LLMs complete the Wug Test across the six languages. Below we show an example for one question of the Vietnamese Wug test.

**Vietnamese Wug Test:** Assuming that "bing" is a Vietnamese word, read the paragraph and replace the underscores with a suitable form of "bing". *Người đàn ông đứng trên trần nhà. "Đây là một người biết cách bing. Anh ta đang **bing**. Anh ta đã làm điều tương tự ngày hôm qua. Anh ta đã làm gì hôm qua? Hôm qua anh ta __. ( bing/đã bing)"*

In this example, a word should be filled in to indicate past tense of the nonce word "bing". Past tense in Vietnamese does not require changing the word form. The correct form should be the same nonce word. The word "đã" can be optionally added to further clarify that the action is in the past.

Notably, the original Wug test had a pre-defined ground truth response for each question, which were not available for our newly translated languages. Therefore, we asked the language experts to judge whether the model's responses conform to the morphological rules of their language, eval-

uating the correctness of each answer on a scale of 1 (fully disagree) to 5 (fully agree). Finally, we asked these speakers to provide their own preferred completion of the task.

## 5 Results

### 5.1 Accuracy

To calculate accuracy, we binarized the ratings from native speakers into correct and wrong. We consider responses with human ratings of score 4 and 5 to be correct, and those responses rated as 1 and 2 to be wrong. For responses rated with 3, we assign "correct" if the human response matches exactly with the model's response, and "wrong" otherwise.

| Language | T | E | I | Model | Acc. |
|---|---|---|---|---|---|
| Vietnamese | 0.03% | -16 | -1.2099 | GPT-3.5<br>GPT-4 | 87%<br>91% |
| French | 1.78% | -11 | 0.0469 | GPT-3.5<br>GPT-4 | 52%<br>87% |
| Spanish | 0.79% | -11 | 0.0470 | GPT-3.5<br>GPT-4 | 78%<br>70% |
| Romanian | 0.17% | -8 | 0.1106 | GPT-3.5<br>GPT-4 | 56%<br>65% |
| Portuguese | 0.54% | -6 | 0.2948 | GPT-3.5<br>GPT-4 | 56%<br>74% |
| German | 1.68% | -12 | 0.4648 | GPT-3.5<br>GPT-4 | 66%<br>62% |

Table 1: Results from the Human Evaluation of LLM's completions on the Multilingual Wug Test. Column *T* lists the representation of each language GPT-3's training data. E-complexity (column *E*) is the Lupyan and Dale (2010)'s morphological complexity score. I-complexity (*I*) is Bentz et al. (2015)'s entropy-based measure for lexical diversity.

Table 1 shows the results for the six tested languages, as judged by native speakers. Descriptive statistics reveals that both GPT-4 and GPT-3.5 are generally able to generate correct morphemes for the nonce words. The mean accuracy is 0.69 (SD = .46). GPT-4 scores slightly higher than GPT-3.5 (M = .74, SD = .44 versus M = .64, SD = .48), but this difference is not significant under a *t*-test, $t(2{,}274) = -1.69$, $p = .09$. We cannot conclude that GPT-4 is more capable than GPT-3.5.

### 5.2 Effect of Morphological Complexity

To connect our results per with the languages' morphological complexity, we quantify to what extent accuracy is affected by a language's E-complexity using a measure from Lupyan and Dale (2010) and I-complexity using a measure from Bentz et al. (2015), as well as the percentage of GPT-3's training data per language, for which we take the dataset statistics from GPT-3[1] as estimates.

To test whether morphological complexity scores predict LLMs' performance on the Wug test, we fitted mixed-effect logistic regression models to predict accuracy from morphological complexity values, potentially modulated by the amount of training data. The analysis is conducted in R using the `lme4` package. All variables were centered and scaled before the analysis. We consider the question number as a random effect because it is expected that the difficulty varies per question. We have experience with adding more random effects (e.g., type of GPT model, evaluator, language), yet those did not yield a better fit as tested via ANOVA. Due to high co-linearity (VIF>10 for I-complexity and VIF>5 for E-complexity), we split the model into Model 1 with I-complexity and Model 2 with E-complexity – with the language's representation in the training data being present in both.

The results of Model 1 (see Table 2) show that I-complexity scores have a significant weak negative effect on accuracy scores ($\beta = -.67$, $p = .0187$). The results of Model 2 (see Table 3) show that E-complexity scores do not predict LLMs' performance on the Wug test ($\beta = .10$, $p = .7463$). The amount of training data was found not predictive of Wug test performance in both models ($\beta = .10$, $p = .5853$ and $\beta = -.02$, $p = .9137$, respectively). Further, there is no interaction effect between I and the amount of training data. The interaction effect between E-complexity and training data is, however, nearly significant. These results suggest that it is the irregularity of the morphological system rather than the number of inflectional categories that predicts the morphological capabilities of the investigated LLMs. Notably, the amount of training data does not seem to affect the morphological knowledge learned by LLMs. Figure 2 visualizes the relationship between binary accuracy and each of the predictors (E-complexity, I-complexity, and training data percentage).

### 5.3 Error Analysis

We also analyzed the models' incorrect responses (rated 2 or lower, or 3 with mismatching responses)

[1] https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_character_count.csv

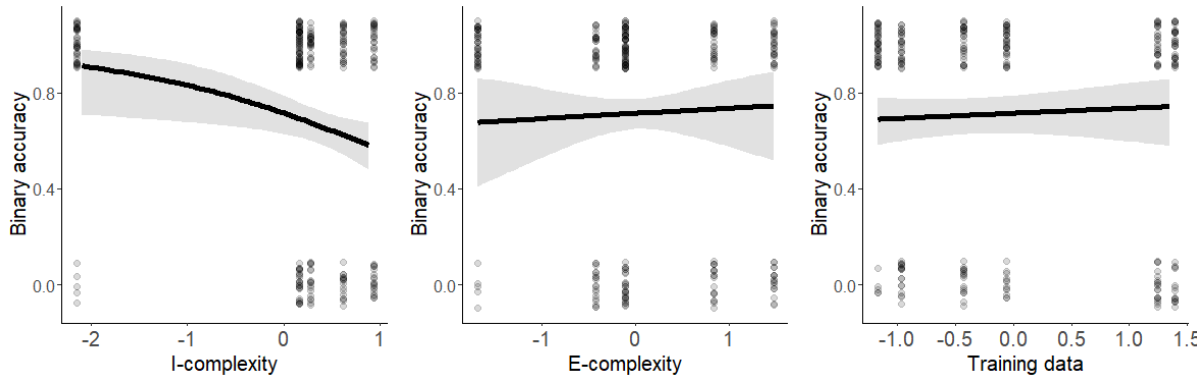


Figure 2: Binary accuracy based on human ratings of LLM responses (y-axis, added jitter) with respect to I-complexity *(Left)* and E-complexity *(Center)*, with higher being more complex, as well as training data percentage *(Right)*. Regression lines are logistic regression with the factor of the (scaled) x-axis as sole predictor and question number as random effect. Results show a trend that LLM responses to the Wug Test in languages that are more complex under these measures receive lower ratings from native speakers.

| Var | Estimate | SE | z value | *p*-value |
|---|---|---|---|---|
| T | 0.1036 | 0.1898 | 0.546 | 0.5853 |
| I | -0.6744 | 0.2869 | -2.351 | ***0.0187** |
| T:I | -0.0461 | 0.2389 | -0.193 | 0.8468 |

Table 2: Results of mixed effect logistic regression with binary accuracy as dependent variable and question number as random effect. Fixed effects are training data (T) and I-complexity (I) and their interaction (T:I).

| Var | Estimate | SE | z value | *p*-value |
|---|---|---|---|---|
| T | -0.0207 | 0.1914 | -0.108 | 0.9137 |
| E | 0.1049 | 0.3243 | 0.324 | 0.7463 |
| T:E | 0.7331 | 0.3857 | 1.901 | 0.0574 |

Table 3: Results of mixed effect logistic regression with binary accuracy as dependent variable and question number as random effect. Fixed effects are training data (T) and E-complexity (E) and their interaction (T:E).

with the goal of detecting any systematic patterns in LLMs' morphological knowledge (or lack thereof). When zooming in on the incorrect responses only, we detected four types of errors:

One type of error is that the models do not inflect the nonce word at all, when it should be inflected, e.g., using an affix (*inflection ignorance*). For example, the correct plural form for the nonce word "tass" in Spanish would be "tasses". However, GPT-3.5 did not add the suffix *-es*, and simply produced the uninflected singular form.

A second type of error was that models occasionally failed to choose the correct affixes (*inflection mismatch*). For example, in German the model generated accusative plural "Lunen", instead of nominative plural "Lune" for the word "Lun".

A third type of error was that the models sometimes applies English morphological rules to nonce words in other languages (*English fall-back*). For example, in Vietnamese, "dã" should be added before the verb to create the past form. In the case of the nonce verb "bing", the models should have responded with "dã bing". However, the model's response was "binged" – which wrongfully follows the grammatical rule of English. We attribute this kind of error to the dominance of English and possibly due to the English Wug test being present in the models' training data. Although "bing" is a phonotactically valid in Vietnamese, the models mistakenly considered it as an English nonce word, as in the original Wug test, and thus completed the sentence with the English past form.

As a fourth type of error, we also observe the *real-word bias*, as reported by Weissweiler et al. (2023), whereby the models sometimes treated the nonce word as if it was a similar existing word in the language, and provide an inflected form for that word. For example, the nonce word "tass" was wrongly pluralized to "Tassen", which is the plural of the very similar existing German word "Tasse".

## 6 Discussion

Our goal was to investigate how well multilingual LLMs learn the underlying morphosyntactic structure of different languages and how this is influenced by languages' degree of morphosyntactic complexity. We did this by applying a Wug test in 6 different languages, and evaluating the models' responses as a function of two measures of

complexity, as well as the representation of the language in the training data.

**Morphology matters** We found that integrative morphological complexity (I-complexity) is more predictive of LLMs' out-of-distribution performance than the language's representation in the pre-training data – a surprising finding given that the amount of training data is usually considered the main driving factor for language modeling performance. For example, despite having the least amount of training data (0.03%), the models' performance was much better on Vietnamese compared to other languages (average accuracy of 85%), which has the lowest E- and I-complexity scores. Notably, all of the observed failures on the Vietnamese Wug test belong to the first error category: the misuse of English morphological rules.

Our results also show that different dimensions of morphological complexity affect LLMs' performance to different degrees. Specifically, we found that only I-complexity (which corresponds to predictability of word forms from context) predicts Wug test accuracy, but not E-complexity. Thus, while languages with a lot of word forms are more challenging for LLMs to learn, the predictability of these word forms given appears to have a greater impact on LLM performance.

Lastly, our results show that the amount of training data seems to be less important than morphological complexity. Specifically, we did not find that the language's representation in the model's training data is not predictive of its morphological capabilities. With this finding, we further support Liu and Hulden (2022), who found that Transformer-based models fail to inflect unknown words despite them being trained on a large amount of data.

**Error types** Our error analysis shows that LLM's occasionally make mistakes in inflecting the nonce words. Besides the real word bias revealed by Weissweiler et al. (2023), our error analysis revealed three more types of errors beyond real-world bias: inflection ignorance, inflection mismatch, and English fall-back. We attribute the English fall-back to the high prominence of English in the model's training data (90%+).

**Comparison with previous studies** Previous studies found the effect of E-complexity on LLMs' performance (Cotterell et al., 2018b; Park et al., 2021; Gerz et al., 2018). However, we do not find any effect of E-complexity on LLMs' Wug test accuracy. Rather, we found that I-complexity predicts morphological capability of LLMs. It should be noted that these studies measure the relationship between morphological complexity and different metrics of LLMs. While we attempt to use behavioral probing to measure morphological knowledge of LLMs, other work uses modeling difficulty (Cotterell et al., 2018b), perplexity (Gerz et al., 2018), and surprisal (Park et al., 2021; Mielke et al., 2019). Furthermore, the high correlation between the analyzed morphological complexity measures confirms the findings of Cöltekin and Rama (2023).

Comparing our results with Weissweiler et al. (2023), we can confirm that LLM's accuracy on the morphological completion of nonce words is not perfect. A particularly interesting case is German: Among the languages studied here, German has a relatively low E-complexity score, but the highest I-complexity score. Weissweiler et al. (2023) found German to be the best-performing language, with 86.49% accuracy, taking into account the five most probable completions for each stimulus $k = 5$. However, comparing German on a $k = 1$ setup with long prompts (most similar to ours), the other study reports 62.18% accuracy, which is indeed comparable with our results for German: 62% (GPT-4) and 66% (GPT-3.5). Therefore, we assume that this drastic drop in accuracy (86% to 62%) can be attributed to the number of possible generation attempts that are taken into account ($k = 5$ vs. $k = 1$). For future studies, it is therefore important to take into account the number of generation attempts.

In the context of comparing large language models to humans, our results suggest that what is more complex for us is also more complex for LLMs. Specifically, work on first and second language acquisition suggests that languages with more complex morphosyntactic structures are harder to learn (Kempe and Brooks, 2008; DeKeyser, 2005; Dressler, 2003, 2010; Slobin, 1985). Our study is in line with this conclusion, and extend it to LLMs. It also confirms recent insights from artificial language learning experiments, which found that artificial miniature languages with more systematic structures are easier to learn and generalize across adult humans, small recurrent neural networks trained from scratch, and large language models (Galke et al., 2023; Raviv et al., 2021).

**Implications** Our findings have important implications for low-resource language processing.

Specifically, it is worth paying attention to languages' morphological complexity. When aiming for equal capabilities across languages in multilingual LLMs, the classic approach would be to counterbalance the representation of low-resource languages in the training data. However, our results suggest that this is not sufficient: we found no significant effect of training data representation on Wug test accuracy (within the frame of the analyzed data). Potentially, other tokenization strategies, such as single-byte tokenization (Xue et al., 2022) or morphology-guided tokenization (Creutz and Lagus, 2007) could help improve LLM's performance on low-resource language processing.

## 7 Conclusion

We tested whether languages' morphological complexity affected the performance of multilingual large language models on a classic language task. We ran the Wug test in 6 languages and analyzed how task performance was affected by the degree of morphological complexity in each language. Our results show that languages' morphological complexity (specifically, integrative complexity), is more important than its relative representation in the training data of large language models – a finding that challenges conventional wisdom and comes with important implications for low-resource language modeling. We have further identified additional error types beyond real-world bias such as English fall-back and inflection ignorance, whose cause we will explore in future work by investigating the role of tokenization.

## Data Availability

The translations of the Wug test into the six considered languages, the script for querying the language models, and the script for our statistical analysis is available under `https://github.com/dangthithaoanh/multilingual-wug-test-on-LLMs`.

## Limitations

We have limited ourselves to comparing only two large language models because we prioritized having an expert judgement for each individual model response. The share of each language in the LLM's pre-training data is taken from the original GPT-3 repository as estimates for GPT-3.5 and GPT-4. Another limitation is that the nonce words could appear more irregular in some languages than in others. Moreover, for most languages, we only had one language expert providing the ratings of grammatical correctness. However, we have qualitatively checked the interrater agreement on Vietnamese and found high agreement. Lastly, we have only considered one language (Vietnamese) for the category of isolating morphology.

## Ethical Considerations

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


## Acknowledgements

We are extremely grateful for the effort that the language experts have put into evaluating the responses of the large language models: Mathilde Josserand, Oxana Grosseck, Sergio Miguel Pereira Soares, Lucia de Hoyos, Tin Le, and Lois Dona. We further appreciate the helpful comments and feedback that we received from the audience of the Evolang conference in Madison, WI, USA, 2024.


## References


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

Matthew Baerman, Dunstan Brown, and Greville G Corbett. 2015. *Understanding and measuring morphological complexity*. Oxford University Press, USA.

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóǧa, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North

Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

Christian Bentz, Tatyana Ruzsics, Alexander Koplenig, and Tanja Samardžić. 2016. A comparison between morphological complexity measures: Typological data vs. language corpora. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 142–153, Osaka, Japan. The COLING 2016 Organizing Committee.

Christian Bentz, Annemarie Verkerk, Douwe Kiela, Felix Hill, and Paula Buttery. 2015. Adaptive Communication: Languages with More Non-Native Speakers Tend to Have Fewer Word Forms. *PLOS ONE*, 10(6):e0128254.

Jean Berko. 1958. The child's learning of english morphology. *Word*, 14(2-3):150–177.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Angelica Chen, Ravid Schwartz-Ziv, Kyunghyun Cho, Matthew L. Leavitt, and Naomi Saphra. 2023. Sudden Drops in the Loss: Syntax Acquisition, Phase Transitions, and Simplicity Bias in MLMs.

Cagri Cöltekin and Taraka Rama. 2023. What do complexity measures measure? Correlating and validating corpus-based measures of morphological complexity. *Linguistics Vanguard*, 9(s1):27–43.

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

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018a. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels. Association for Computational Linguistics.

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

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(1):1–34.

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

Wolfgang U Dressler. 2003. Degrees of grammatical productivity in inflectional morphology. *Italian Journal of Linguistics*, 15:31–62.

Wolfgang U Dressler. 2010. A typological approach to first language acquisition. *Language acquisition across linguistic and cognitive systems*, 52:109–124.

Matthew S Dryer and Martin Haspelmath. 2013. Wals online (v2020. 3). *Zenodo https://doi.org/10.5281/zenodo*, 7385533.

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

Lukas Galke, Yoav Ram, and Limor Raviv. 2023. What makes a language easy to deep-learn?

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 316–327, Brussels, Belgium. Association for Computational Linguistics.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Kees Hengeveld and Sterre Leufkens. 2018. Transparent and non-transparent languages. *Folia Linguistica*, 52(1):139–175.

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

Tovah Irwin, Kyra Wilson, and Alec Marantz. 2023. BERT Shows Garden Path Effects. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3220–3232, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR.

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

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Ling Liu and Mans Hulden. 2022. Can a Transformer Pass the Wug Test? Tuning Copying Bias in Neural Morphological Inflection Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 739–749, Dublin, Ireland. Association for Computational Linguistics.

Ling Liu and Lingshuang Jack Mao. 2016. Morphological reinflection with conditional random fields and unsupervised features. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 36–40, Berlin, Germany. Association for Computational Linguistics.

Gary Lupyan and Rick Dale. 2010. Language structure is partly determined by social structure. *PLoS ONE*, 5(1):e8559.

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

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

OpenAI. 2023. GPT-4 Technical Report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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

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

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

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Anton Schäfer, Shauli Ravfogel, Thomas Hofmann, Tiago Pimentel, and Imanol Schlag. 2024. Language imbalance can boost cross-lingual generalisation. *arXiv preprint arXiv:2404.07982*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Dan I Slobin. 1985. The child as a linguistic icon-maker. *Iconicity in syntax*, pages 221–248.

Shashank Srikant, Ben Lipkin, Anna A. Ivanova, Evelina Fedorenko, and Una-May O'Reilly. 2022. Convergent Representations of Computer Programs in Human and Artificial Neural Networks. In *Advances in Neural Information Processing Systems*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, pages 1–16.

Leonie Weissweiler, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Haofei Yu, et al. 2023. Counting the bugs in chatgpt's wugs: A multilingual investigation into the morphological capabilities of a large language model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6508–6524.

Colin Wilson and Jane S.Y. Li. 2021. Were we there already? Applying minimal generalization to the SIGMORPHON-UniMorph shared task on cognitively plausible morphological inflection. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 283–291, Online. Association for Computational Linguistics.

Shijie Wu, Ryan Cotterell, and Timothy O'Donnell. 2019. Morphological Irregularity Correlates with Frequency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5117–5126, Florence, Italy. Association for Computational Linguistics.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Lining Zhang, Mengchen Wang, Liben Chen, and Wenxin Zhang. 2022. Probing GPT-3's linguistic knowledge on semantic tasks. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 297–304, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yanan Zheng, Jing Zhou, Yujie Qian, Ming Ding, Chonghua Liao, Li Jian, Ruslan Salakhutdinov, Jie Tang, Sebastian Ruder, and Zhilin Yang. 2022. FewNLU: Benchmarking state-of-the-art methods for few-shot natural language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 501–516, Dublin, Ireland. Association for Computational Linguistics.