A Morphology-Aware Evaluation of Turkish Syntax in Large Language Models

Ezgi Başar **Arianna Bisazza**
Center for Language and Cognition (CLCG), University of Groningen
{e.basar, a.bisazza}@rug.nl

Abstract

Minimal pair benchmarks have become a common approach for evaluating the syntactic knowledge of language models (LMs). However, the creation of such benchmarks often overlooks language-specific confounders that may affect model performance, particularly in the case of morphologically rich languages. In this paper, we investigate how surface-level factors such as morpheme count, subword count, and sentence length influence the performance of LMs on a Turkish benchmark of linguistic minimal pairs. We further analyze whether a tokenizer's degree of alignment with morphological boundaries can serve as a proxy for model performance. Finally, we test whether the distribution of morphemes in a minimal pair benchmark can skew model performance. Our results show that while surface factors have limited predictive power, they might still serve as a systematic source of bias. Moreover, we find that morphological alignment can roughly correspond to model performance, and morpheme-level imbalances in the benchmark may have a significant influence on results.

1 Introduction

The evaluation of language models' linguistic capabilities has increasingly relied on minimal pair benchmarks, where models must distinguish between a set of syntactically acceptable and unacceptable sentence pairs. Following the seminal work by Warstadt et al. (2020), BLiMP-style benchmarks have become the standard approach for the linguistic evaluation of LMs. While these benchmarks now exist for numerous languages, their design does not always account for language-specific confounding factors that could skew performance metrics.

Turkish, an agglutinative language with rich morphology, presents particular challenges in this respect. Words in Turkish typically consist of multiple morphemes concatenated to a root, creating substantial morphological complexity. The highly productive nature of agglutinative morphology means that speakers can generate an overwhelming number of legitimate word forms through regular combinations of roots and suffixes. This leads to an explosion of possible word forms beyond what can be included in a fixed tokenizer vocabulary. Consequently, the subword tokenizers on which current language models are based segment most morphologically complex words into subword units, which may or may not correspond with morphological boundaries. The same morpheme may be represented by different subword tokens depending on the surrounding context. This poses another complication for learning morphosyntactic patterns.

This paper presents a morphology-aware analysis of LM performance on a recently introduced benchmark of Turkish syntactic minimal pairs (Başar et al., 2025), focusing on three main aspects. We first investigate how well surface-level differences in character, word, subword, and morpheme counts predict model performance across minimal pairs. Subsequently, we test whether the morphological alignment of tokenizers corresponds to how models make grammatical judgments. Finally, we explore whether balancing morphological distributions within a benchmark can suggest important considerations for benchmark creation.

Based on our analyses, we find that although differences in character, word, and morpheme counts are overall weak predictors of model performance, they can influence model behavior in predictable ways. We further observe that the degree to which a tokenizer respects morphological boundaries can inform us about model performance. Critically, we demonstrate that imbalanced morpheme distributions in a minimal pair benchmark can have a significant effect on the results. Our work underlines the importance of controlling for morphological confounders when evaluating language models on agglutinative languages.

Proceedings of the Second Workshop Natural Language Processing for Turkic Languages (SIGTURK 2026), pages 95–102
March 29, 2026

2 Background

The BLiMP benchmark provides a framework for testing the linguistic knowledge (Warstadt et al., 2020) of language models for English, and it has since been adapted for many languages including Turkish (Başar et al., 2025), Chinese (Xiang et al., 2021), Russian (Taktasheva et al., 2024), and Dutch (Suijkerbuijk et al., 2025), among others. These benchmarks are designed to isolate specific grammatical contrasts by creating sentence pairs that differ minimally, often by only one or two words.

However, even carefully constructed benchmarks may contain unintended but systematic differences between acceptable and unacceptable sentences. These differences could potentially act as confounders in minimal pair evaluations. For morphologically rich languages like Turkish, this concern is particularly relevant. In Turkish, grammatical acceptability distinctions often hinge on morphological choices, leading to systematic differences in word length and morphological complexity between acceptable and unacceptable sentences.

The Turkish Benchmark of Linguistic Minimal Pairs, also referred to as TurBLiMP (Başar et al., 2025), covers sixteen grammatical phenomena comprising sixteen thousand minimal pairs. Evaluations on TurBLiMP reveal that model architecture significantly influences performance, with masked language models showing different patterns than causal models. Training data characteristics also play a crucial role, with monolingual Turkish models often outperforming larger multilingual models.

Turkish belongs to the agglutinative language family, characterized by rich morphology and flexible word order (Göksel and Kerslake, 2005). Words in Turkish can be formed productively by concatenating a root and one or more suffixes. This structure allows for potentially infinite word forms from finite morpheme inventories. Speakers routinely generate and parse various legitimate yet low-frequency word forms through regular morphological processes, imposing a significantly larger vocabulary burden on LMs compared to analytic or fusional languages. The extensive use of morphology for syntactic functions also implies that morphosyntactic cues play a crucial role in linguistic acceptability tasks.

Previous research suggests that agglutinative languages, as opposed to those that are fusional or analytic, present idiosyncratic challenges for language models (Cotterell et al., 2016; Park et al., 2021; Gerz et al., 2018; Arnett and Bergen, 2025). It has been shown that increased morphological complexity can considerably degrade neural machine translation performance (Ataman et al., 2017). Additionally, research has demonstrated that language models struggle with morphological generalization in Turkish, particularly when encountering novel combinations of morphemes (Ismayilzada et al., 2025). Recent work by Poelman et al. (2025) observes no conclusive differences in the morphological alignment of tokenizers for agglutinative and fusional languages, and proposes a subword-based bigram metric to explain performance disparities.

Our study builds on this foundation by conducting a detailed analysis of morphological or subword-based factors that may confound evaluation results. We extend previous work by systematically investigating surface-level confounds, analyzing the relationship between tokenizer alignment and model performance, and testing the impact of morphological balancing on benchmark performance. Our approach provides insights not only into model behavior as it relates to syntactic evaluations, but also into minimal pair benchmark design considerations for agglutinative languages.

3 Experimental Setup

We selected five language models representing different architectures, training objectives, and performance levels on the TurBLiMP[1] benchmark (Başar et al., 2025). The selection includes both top-performing models and models with more modest accuracy to enable comparative analysis.

The three best-performing models on TurBLiMP were EuroLLM with 9B parameters (Martins et al., 2024), BERTurk with a vocabulary size of 128k (Schweter, 2020), and the large variant of cosmosGPT (Kesgin et al., 2024). EuroLLM is a multilingual causal language model with strong performance across European languages. BERTurk is a Turkish-only masked language model with a WordPiece tokenizer. It was the only model reported to correlate with human judgments on the TurBLiMP benchmark. cosmosGPT is a Turkish-only causal language model trained on extensive Turkish web and book corpora.

The two lowest-performing models were the Goldfish models and Qwen 2.5 with 7B parame-

[1] `https://github.com/ezgibasar/turblimp`

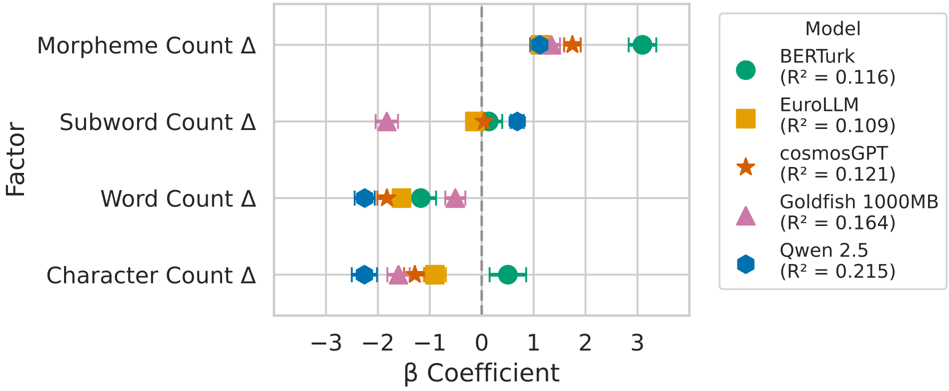

Figure 1: β coefficients fitted for the BERTurk, EuroLLM, cosmosGPT, Goldfish, and Qwen 2.5 models.

ters. Goldfish is a series of causal language models trained on various training data sizes. In the subsequent experiments, we will be using the variant that was trained on one gigabyte of Turkish text. Qwen 2.5 is a multilingual model that, despite its larger size, showed lower accuracy on the benchmark.

For the purposes of our paper, accuracy refers to the percentage of minimal pairs where the model assigns a higher sequence log-probability to the acceptable sentence than to its unacceptable counterpart. This performance metric provides a measure of how well each model captures the grammatical distinctions tested in the benchmark.

4 Analysis of Surface-Level Factors

In this section, we investigate the extent to which our tested language models rely on surface differences between the acceptable and unacceptable sentences. We analyze four predictors of model performance, namely the differences in character counts, word counts, subword counts, and morpheme counts.

4.1 Factors

Character count refers to sentence length based on the total number of characters in each sentence. For this factor, we take the difference in character counts between the acceptable sentence and the unacceptable one. **Word count** corresponds to the number of whitespace-separated tokens. **Subword count** measures the number of smaller units a tokenizer splits a sequence into. Subwords might loosely follow word or morpheme boundaries. However, particularly for agglutinative languages, they often do not display a one-to-one correspondence with either. For each minimal pair, we tokenize the cue words using each model's specific tokenizer and record the number of subwords. We compute subword count differences by taking the subword count of the acceptable choice and subtracting that of the unacceptable one. **Morpheme count** measures the number of morphemes as identified by the Zemberek morphological pipeline by Akın and Akın (2007). We include morpheme count as an analysis dimension even though our tested language models have no access to morphological parses. Including both subword count and morpheme count differences as predictors allows us to examine whether models behave differently when the acceptable choice is a more morphologically complex word compared to when they simply encounter more subword tokens. If models are sensitive to morphological complexity rather than surface-level patterns, we would expect different patterns to emerge.

Using these four fixed effects, we fit separate linear regression models for each language model. Each regression model predicts the log-probability difference between acceptable and unacceptable sentences based on four factors. This approach quantifies whether surface-level factors are good predictors of model performance and allows us to see if these influences differ between high-performing and low-performing models.

4.2 Results

We visualize the effects of our predictors by plotting the standardized β coefficients for four factors across five models. Figure 1 further includes each model's corresponding R^2 value as an indication of how well these predictors explain the variance in log probability differences.

The results reveal several important patterns. Character count differences show a clear architectural divide. The autoregressive models, including EuroLLM, cosmosGPT, Qwen 2.5, and Goldfish, all exhibit significant negative coefficients. This reflects the tendency of autoregressive models to prefer shorter sequences. The masked language model BERTurk, however, shows a positive coefficient. In other words, BERTurk is more likely to assign a higher probability to the acceptable sentence when the acceptable sentence is longer. This difference indicates that masked language models may benefit from longer contexts, providing more information to make the correct prediction.

Word count differences show negative coefficients across models. This suggests that models tend to prefer more economical expressions when making grammaticality judgments. The strength of this preference varies, with Qwen 2.5 showing the strongest negative coefficient and Goldfish showing the weakest.

Subword count differences exhibit the most variable patterns across models. Goldfish shows a strong negative coefficient. Qwen 2.5 shows a strong positive coefficient. BERTurk, EuroLLM, and cosmosGPT show coefficients near zero with varying levels of statistical significance. The varying patterns motivate our investigation of tokenizer alignment in the next section. Informed by prior work (Goldman et al., 2024; Jumelet et al., 2025), we would have expected to see degraded performance when the acceptable variant contains a greater number of subwords. However, only the Goldfish model behaves in line with this expectation.

Finally, the effect of morpheme count difference is consistently positive and statistically significant across all five models. This implies that models are more likely to correctly identify the acceptable sentence when the acceptable sentence offers a word choice with greater morphological complexity. The magnitude of this effect varies across models, with BERTurk showing the strongest coefficient and Qwen 2.5 showing the weakest.

It is important to note that these surface factors collectively explain only a very small proportion of variance in model judgments. The remarkably low R^2 values indicate that while surface factors may introduce systematic biases, they are not the primary drivers of model performance on TurBLiMP. This finding is desirable for a benchmark intended to reflect syntactic abilities rather than relying on surface-level features.

5 Role of Tokenization Quality

The inconsistent patterns we observed for subword count differences across models point to the importance of understanding how tokenizers handle Turkish morphology. Accordingly, this section investigates the interplay between a tokenizer's morphological alignment and model performance. Morphological alignment refers to how well a subword tokenizer segments words according to morpheme boundaries. This investigation allows us to identify whether models with better-aligned tokenizers also perform better on the benchmark.

Recent work by Arnett and Bergen (2025) examined the morphological alignment of tokenizers in various languages, including Turkish. Their goal was to determine if poor morphological alignment contributes to performance disparities in typologically diverse languages. Their findings ultimately dismissed this factor as significant. It is worth noting, however, that their method for quantifying morphological alignment is different from ours. Rather than evaluating whether a tokenizer identifies all morpheme boundaries (e.g., splitting *elmalarıma* into *elma* + *lar* + *ım* + *a*), they measured whether the tokenizer separates the stem from all other suffixes combined (e.g., *elma* + *larıma*). Poelman et al. (2025) later noted this measure of morphological alignment may be problematic for agglutinative languages, and adopted a more rigorous definition of morphological alignment, taking all morpheme boundaries into account. We adopt the same definition to operationalize morphological alignment. Our study narrows the focus of this inquiry by evaluating the relationship between morphological alignment and a model's performance across a suite of linguistic abilities in one language.

To evaluate tokenizer morphological alignment, we construct a dataset of four hundred morphologically segmented Turkish words using the morphological inflection pipeline by Akın and Akın (2007). This dataset covers four common suffix combinations in Turkish.

Verb + Nominalizer + Possessive + Case Contains 100 verb forms with the nominalizer *-DIK* followed by possessive markers and case endings (e.g., *unuttuğunu* → [unut, tuğ, un, u]). This category targets the segmentation of nominalized verbs.

Verb + TAM + Person Includes 100 finite verbs segmented into stems, tense/aspect/modality (TAM) markers, and person markers (e.g., *olacak-*

tılar → [ol, acak, tı, lar]), targeting inflectional morphology.

Noun + (Plural) + Possessive + Case Comprises 100 possessed nouns decomposed into stems, optional plural markers (*-lAr*), possessive markers, and case endings (e.g., *elmalarıma* → [elma, lar, ım, a]).

Noun + (Plural) + Case Features 100 simpler noun forms with optional plural and case markers (e.g., *elmalara* → [elma, lar, a]), providing a baseline for bare nominal inflection.

We process all four hundred words using each model's tokenizer and compare the resulting segmentations against the gold-standard morpheme boundaries. We use four different metrics to quantify morphological alignment. The first metric is the average Damerau-Levenshtein **distance** between the tokenizer's output and the gold segmentation. Lower distance indicates better alignment. The second metric is the proportion of **undersegmented** words, where the tokenizer produces fewer subwords than the actual number of morphemes. The third metric is the proportion of **oversegmented** words, where the tokenizer produces more segments than the actual morphemes. The fourth metric is the proportion of **exact** matches, where the tokenizer's segmentation fully matches the gold standard.

Tokenizer	Dist.	Underseg.	Overseg.	Exact
BERTurk	2.48	82.2	0.8	5.8
EuroLLM	2.36	24.8	27.5	2.5
cosmosGPT	4.49	51.2	14.0	0.8
Goldfish	1.62	83.0	3.5	9.8
Qwen 2.5	5.50	7.0	54.2	0.5

Table 1: Tokenizer morphological alignment evaluation.

Table 1 presents the results of our tokenizer alignment evaluation. The Goldfish tokenizer achieves the lowest average Levenshtein distance, indicating the best overall alignment with morphological boundaries. However, this alignment seems to be accompanied by a tendency towards undersegmentation, effectively treating many morphologically complex words as single units or minimally segmented forms. This observation could be interpreted in conjunction with the Goldfish model's strong negative subword count coefficient in our earlier experiment. However, we should also note that the masked language model BERTurk does not display the same behavior despite having comparable results in our morphological alignment evaluation.

The Qwen 2.5 tokenizer shows the poorest alignment with the highest Levenshtein distance. This poor alignment results from a reliance on oversegmentation. This tokenizer tends to break words into many small pieces that exceed the number of morphemes. We can also note that Qwen 2.5 had a prominent positive subword count coefficient, mirroring the behavior of the Goldfish model in the opposite direction.

The EuroLLM tokenizer achieves the second-best alignment score through a more balanced approach with both undersegmentation and oversegmentation while cosmosGPT has the second-worst alignment. When we examine the relationship between morphological alignment and overall model performance, we find a loose correspondence but no strict correlation. Goldfish has excellent alignment but lower overall accuracy. BERTurk has moderate alignment but top-tier performance. This suggests that while tokenizer quality affects how morphological information is represented, it is not the sole determinant of a model's ability to make linguistic judgments. It may partially provide a clue as to why models behave differently with respect to surface-level factors, but other considerations, such as model architecture, should also be taken into account.

6 Morpheme Count-Balanced Results

Our findings that models are sensitive to morpheme count differences and that this sensitivity varies across models raises an important question for benchmark design. In this section, we investigate whether morphological balancing should be a consideration when creating minimal pair benchmarks for agglutinative languages. In languages like Turkish where syntax is largely realized through morphology, grammatical contrasts often involve choices between different morphemes. If those morphemes appear with imbalanced frequencies in acceptable versus unacceptable sentences, models could achieve high accuracy only due to a potential bias towards words containing certain morphemes, rather than reflecting linguistic abilities.

With this concern in mind, we create balanced versions of six TurBLiMP phenomena where grammaticality judgments depend primarily on morphological choices. These phenomena include: Argument Structure Transitive, Argument Structure

Ditransitive, Nominalization, Anaphor Agreement, Subject Agreement, and Relative Clauses.

For each phenomenon, we adjust the dataset to ensure a one-to-one ratio between critical morpheme alternatives in acceptable and unacceptable sentences. For example, in the Nominalization phenomenon, grammaticality hinges on whether the choice verb should take the -DIK or -mA nominalizing suffix. In the original benchmark, the -DIK nominalizer appears more frequently in acceptable sentences. In our balanced version, we ensure that -DIK appears in acceptable sentences in exactly fifty percent of pairs and in unacceptable sentences in the other fifty percent. In a similar vein, we can note a further example from the Argument Structure phenomena in which acceptability is determined on the basis of different case morphemes. For instance, the acceptable sentence may feature the dative case suffix -A, while the unacceptable sentence involves the accusative case suffix -I. In other cases, this may be the other way around. We ensure that the alternative case morphemes occur an equal number of times for each condition in our balanced set of minimal pairs. This intervention removes the possibility of achieving high accuracy due to a general preference for one morpheme over the other.

	BERTurk	EuroLLM	cosmosGPT	Goldfish	Qwen 2.5
Original Accuracy	93.7	93.8	93.6	89.1	87.0
Balanced Accuracy	93.4	94.0	90.8	89.7	84.2
Difference	-0.3	+0.2	-3.0	+0.7	-3.2
t-value	-0.85	0.62	-2.12	0.97	-1.95
p-value	0.402	0.540	**0.042**	0.339	0.061

Table 2: Comparison of average accuracy between original and morpheme-balanced benchmark subsets. Results are based on independent sample t-tests.

Table 2 compares average model performance on the original versus morpheme-balanced subsets. The results reveal important differences in how models respond to this intervention. Three models show minimal changes in accuracy with no statistically significant differences. BERTurk shows a negligible decrease of 0.3 percent. EuroLLM shows a negligible increase of 0.2 percent. Goldfish shows a small increase of 0.7 percent. The stability of these models might suggest that their judgments reflect syntactic knowledge rather than an exploitation of morphological distributional biases.

In contrast, two models show notable declines in accuracy. cosmosGPT shows a decrease of 3.0 percent, which is statistically significant. Qwen 2.5 shows a decrease of 3.2 percent, albeit not being statistically significant. The degraded performances indicate that a substantial portion of these models' original high performance stemmed from exploiting imbalanced morphological distributions in the benchmark. When this distributional cue is neutralized through balancing, their weaker underlying syntactic capabilities are revealed.

This finding has direct implications for benchmark design. The fact that some models maintain their performance while others are affected suggests that morphological balancing can be an important consideration when designing such benchmarks. For agglutinative languages like Turkish, where grammatical distinctions are often realized through morphological contrasts, ensuring balanced morpheme distributions should be considered a best practice in benchmark creation.

The differential impact across models also ties into some of our earlier observations. Both cosmosGPT and Qwen 2.5 showed stronger dependencies on morpheme count in our regression analysis. Correspondingly, both models show performance degradation when morphological distributions are balanced. Our results suggest that minimal pair benchmarks for languages where morphosyntax plays a crucial role should employ morphological balancing for robust evaluations.

7 Conclusion

Our morphology-aware analysis of Turkish language models reveals how typological features and benchmark design can affect evaluation outcomes. First, while surface-level factors like character, word, subword, and morpheme counts have limited predictive power overall, they can introduce systematic biases. Notably, all models we examined displayed a sensitivity to morpheme count differences, tending to favor the acceptable sentence when it was morphologically more complex.

The morphological alignment of a model's tokenizer corresponds only loosely with overall performance. We found that high-performing models can show varied tokenization behaviors, relying on undersegmentation, oversegmentation, or a more balanced approach. This suggests that while tokenization quality can impact how morphological cues are processed, it does not provide a direct proxy for model performance.

Most critically, we demonstrate that the morphological composition of a benchmark can be a significant confounder. When we balanced the distribution of critical morphemes in a subset of the TurBLiMP phenomena, the performance of some models declined substantially, while others remained stable. This indicates that imbalanced morphological distributions can lead to a misleading representation of the linguistic abilities of some models.

Limitations

One limitation of our line of inquiry is that we only explored a single agglutinative language. To establish whether our findings reflect some general properties of the modeling of agglutinative languages, similar analyses should be applied to other typologically similar languages, such as Finnish, Hungarian, or Korean.

The number of models in our study, while chosen to represent diverse architectures and performance levels, is relatively small. A broader investigation involving a greater number of models is needed to better disentangle the interactions between model architecture, training data composition, morphological alignment, and model performance. This would help determine if the patterns we observed hold for their representative categories. Using a larger set more comparable models would also enable a more fine-grained analysis of phenomena-specific outcomes.

Our operationalization of morphological alignment is based on a fixed set of 400 constructed word forms. Although this set covers common suffix combinations, it certainly does not capture the full extent of the morphological complexity encountered in natural corpora or the TurBLiMP benchmark.

Addressing these limitations would strengthen the generalizability of our conclusions and further refine best practices for creating linguistically informed evaluations for agglutinative languages.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. Arianna Bisazza was supported by the Talent Programme of the Dutch Research Council (grant VI.Vidi.221C.009).

References

Ahmet Afşın Akın and Mehmet Dündar Akın. 2007. Zemberek, an open source nlp framework for turkic languages. *Structure*, 10:1–5.

Catherine Arnett and Benjamin Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6607–6623, Abu Dhabi, UAE. Association for Computational Linguistics.

Duygu Ataman, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Linguistically motivated vocabulary reduction for neural machine translation from turkish to english. *The Prague Bulletin of Mathematical Linguistics*, 108:331 – 342.

Ezgi Başar, Francesca Padovani, Jaap Jumelet, and Arianna Bisazza. 2025. TurBLiMP: A Turkish benchmark of linguistic minimal pairs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16506–16521, Suzhou, China. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared Task—Morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22, Berlin, Germany. Association for Computational Linguistics.

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.

Aslı Göksel and Celia Kerslake. 2005. *Turkish: A Comprehensive Grammar*. Comprehensive grammars. Routledge.

Omer Goldman, Avi Caciularu, Matan Eyal, Kris Cao, Idan Szpektor, and Reut Tsarfaty. 2024. Unpacking tokenization: Evaluating text compression and its correlation with model performance. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2274–2286, Bangkok, Thailand. Association for Computational Linguistics.

Mete Ismayilzada, Defne Circi, Jonne Sälevä, Hale Sirin, Abdullatif Köksal, Bhuwan Dhingra, Antoine Bosselut, Duygu Ataman, and Lonneke Van Der Plas. 2025. Evaluating morphological compositional generalization in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1270–1305, Albuquerque, New Mexico. Association for Computational Linguistics.

Jaap Jumelet, Leonie Weissweiler, and Arianna Bisazza. 2025. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs. *Preprint*, arXiv:2504.02768.

H. Toprak Kesgin, M. Kaan Yuce, Eren Dogan, M. Egemen Uzun, Atahan Uz, H. Emre Seyrek, Ahmed Zeer, and M. Fatih Amasyali. 2024. Introducing cosmosgpt: Monolingual training for turkish language models. *arXiv preprint arXiv:2404.17336*.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. Eurollm: Multilingual language models for europe. *Preprint*, arXiv:2409.16235.

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.

Wessel Poelman, Thomas Bauwens, and Miryam de Lhoneux. 2025. Confounding factors in relating model performance to morphology. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7273–7298, Suzhou, China. Association for Computational Linguistics.

Stefan Schweter. 2020. Berturk - bert models for turkish.

Michelle Suijkerbuijk, Zoë Prins, Marianne de Heer Kloots, Willem Zuidema, and Stefan L. Frank. 2025. Blimp-nl: A corpus of dutch minimal pairs and acceptability judgments for language model evaluation. *Computational Linguistics*, pages 1–39.

Ekaterina Taktasheva, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, Ekaterina Artemova, and Vladislav Mikhailov. 2024. RuBLiMP: Russian benchmark of linguistic minimal pairs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9268–9299, Miami, Florida, USA. Association for Computational Linguistics.

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.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A benchmark for Chinese language model evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790, Online. Association for Computational Linguistics.