Autoregressive Transformer Models Are Systemically Biased Toward English-like Languages

Anonymous ACL submission

Abstract

Language models exhibit systematic performance differences across languages, but it remains unclear whether these disparities stem from intrinsic linguistic properties or artifacts of engineering choices. To disentangle these factors, we train over 200 monolingual language models from scratch on parallel data across 10 typologically diverse languages from 6 language families, systematically varying tokenization algorithms, vocabulary sizes, and pretokenization strategies, and evaluate model performance on four tasks on cross-linguistically comparable benchmarks. We observe systematic performance gaps across languages across conditions, with languages more typologically distant from English being harder to model. Different languages benefit from different tokenization settings. Further analysis reveals a trend different from previous work: we observe a negative correlation between metrics of tokenization quality and ultimate language model performance. We conclude that the systematic language gaps are due to fundamental architectural choices; our results suggest that autoregressive attention-based decoders are unlikely to yield fair crosslinguistic language modeling.[1]

1 Introduction

Cross-lingual performance disparities in language models have been widely documented (Chang et al., 2024; Ramesh et al., 2023; Ahuja et al., 2023, *inter alia*), yet whether these disparities reflect intrinsic differences in languages or are rather artifacts of engineering choices (e.g., training data, tokenization, and evaluation) remains an open question (e.g., Arnett and Bergen, 2025; Park et al., 2021; Shani et al., 2026). Answering this question has twofold significance. From an NLP perspective, it exposes potential biases in current systems and informs the development of more typologically robust and equitable models. From a cognitive science perspective, a well-established finding in language acquisition is that all natural languages are learned with broadly comparable ease by typically developing children (Gleitman and Newport, 1995; Armon-Lotem et al., 2016); if transformer models fail to meet this criterion even in controlled monolingual settings, this suggests that the architecture itself might encode inductive biases fundamentally different from those of human learners. This would call into question the use of transformers for cognitive modeling.

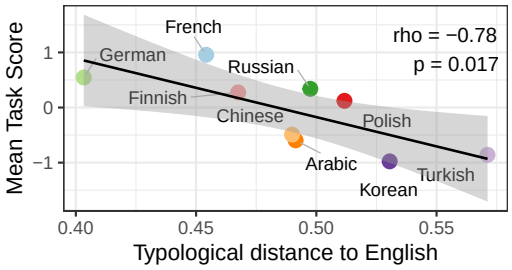


Figure 1: Spearman's rank correlation between typological distance to English and mean performance across our language tasks (English is excluded as the reference point).

Many studies have attempted to address this question and provide useful insights (e.g., Koplenig and Wolfer, 2023; Gerz et al., 2018; Cotterell et al., 2018; Park et al., 2021; Ulčar et al., 2025). However, existing findings remain difficult to interpret, because each study necessarily makes simplifying assumptions, and no single study has simultaneously controlled for all major confounding factors, including training data, tokenization, and evaluation metrics (Shani et al., 2026).

First, previous studies differ in how they handle training data: some control for amount (Arnett and Bergen, 2025) or use parallel corpora (Cotterell et al., 2018; Mielke et al., 2019), while others sacrifice data control for linguistic diversity (Koplenig

[1] This is a pre-registered study [OSF link]; models and code will be released upon acceptance.

and Wolfer, 2023). Second, tokenization has received less systematic attention as a confounding factor, as most studies use fixed vocabulary sizes and tokenization algorithms, despite evidence that different languages benefit from different settings (Gazit et al., 2025; Petrov et al., 2023; Ali et al., 2024; Arnett et al., 2025). Third, most studies focus on a single probability-based metric such as perplexity or surprisal, yet lower perplexity does not always translate to better downstream performance (Liu et al., 2023; Hu et al., 2024). Finally, the effect of scaling has received little attention, leaving it unclear whether existing conclusions, which are largely based on smaller models and data sizes, generalize when scaled up. This is a critical issue given today's very large language models.

In this study, we address these limitations by (a) training monolingual transformer models from scratch exclusively on parallel data across 10 typologically diverse languages from 6 language families; (b) systematically manipulating tokenization configurations, including pre-tokenization strategies, tokenization algorithms, and vocabulary sizes; (c) evaluating models using four tasks on comparable parallel benchmarks; and (d) experimenting with scaled data and model sizes. We ask: after strictly controlling these potential confounding factors, are all languages equally difficult to learn by transformer models? If not, what linguistic typological and engineering choices drive such gaps? And, how can such gaps be mitigated?

We find that systematic performance gaps persist across languages under a wide range of experimental settings. Linguistic factors—in particular, typological distance from English—are the most reliable predictors of relative language modeling difficulty (Fig. 1). Engineering choices are impactful, but their effects vary and are language-specific. Follow-up experiments show that neither controlling for training steps nor tokenization settings closes the gap. Critically, the cross-linguistic gap becomes more pronounced under scaled experimental conditions, suggesting that the English-centric bias of transformer models is not an artifact of scale but a fundamental property of the architecture.

2 Related Work

2.1 Cross-linguistic modeling difficulty

Whether all languages are equally hard to model, and if not, which languages are harder, has been a longstanding question in NLP. Cotterell et al. (2018) and Mielke et al. (2019) use parallel corpora such as the Europarl (Koehn, 2005) and Bible (Mayer and Cysouw, 2014) to evaluate cross-linguistic language modeling difficulty with LSTMs and RNNs. Chang et al. (2024) extend this to over 250 languages using monolingual transformer models and find persistent performance gaps even under largely comparable data conditions. However, Arnett and Bergen (2025) show that once data size is controlled via byte-premium (Arnett et al., 2024a), some of these differences disappear. These conflicting findings raise the question of whether observed cross-linguistic disparities reflect genuine linguistic difficulty or methodological confounds.

2.2 Sources of cross-linguistic difficulty

Prior work has proposed a range of explanations for cross-linguistic performance disparities, spanning both methodological factors, engineering choices, and linguistic properties. Below, we summarize hypotheses from multilingual NLP and cognitive science that relate language structure to learnability. For an extensive recent review of these considerations, see Shani et al. (2026).

Training data Data quantity and quality are difficult to control cross-linguistically, as high-resource languages tend to have richer and higher-quality corpora (Kreutzer et al., 2022; Joshi et al., 2020; Arnett and Bergen, 2025).

Tokenization Tokenization choices can introduce systematic cross-linguistic biases (Gallé, 2019; Schmidt et al., 2024; Zouhar et al., 2023; Bostrom and Durrett, 2020; Goldman et al., 2024; Tao et al., 2024; Shani et al., 2026). Prior work has examined how pre-tokenization strategies (Gazit et al., 2025), tokenization algorithms (Bostrom and Durrett, 2020; Rust et al., 2021; Uzan et al., 2024), and vocabulary sizes (Alrefaie et al., 2024; Mielke et al., 2019) influence language modeling difficulty.[2] Tokenizer quality metrics such as CTC (Schmidt et al., 2024) and Rényi entropy (Zouhar et al., 2023) have been proposed, but their relationship to model performance remains inconsistent (Schmidt et al., 2024; Rust et al., 2021; Cognetta et al., 2024). Further discussion is in App. A.

Evaluation Evaluation design can introduce cross-linguistic biases: intrinsic metrics such as

[2] Pre-tokenization splits raw text into word-level units prior to subword segmentation, while the tokenization algorithm determines how those units are further divided into subwords.

bits-per-byte and perplexity yield different language rankings (Poelman and de Lhoneux, 2026), and different tasks might not correlate with each other (Agarwal et al., 2025; Hu et al., 2020).

Morphology & Orthography Morphologically richer or non-Latin languages are generally harder to model (Park et al., 2021; Cotterell et al., 2018; Shani et al., 2026). This is partly attributable to frequency-based tokenization algorithms such as BPE, which do not respect morpheme boundaries and produce fragmented subword units for languages with complex inflectional systems or different encoding systems. Morphologically aligned or script-aware tokenizers have shown benefits for downstream tasks, though evidence is mixed (Bostrom and Durrett, 2020; Rust et al., 2021; Arnett et al., 2024b; Land and Arnett, 2025).

Speaker population Koplenig and Wolfer (2023) find that across over 1,000 languages, those with more speakers tend to exhibit higher compression-based learning difficulty.

Typological markedness Children tend to acquire typologically less marked features, i.e., those more common cross-linguistically, more easily (e.g., Hughes et al., 2025; Culbertson et al., 2020). However, White and Cotterell (2021) find that this asymmetry does not straightforwardly extend to neural language models.

Distance to English Typological distance from English has received little systematic attention as a predictor of modeling difficulty, despite the fact that mainstream architectures, tokenization schemes, and benchmarks have largely been developed and validated on English-centric pipelines and by English-centric research communities.

3 Experiment Design

3.1 Training Data

Building on OPUS12 (Yang et al., 2025), we construct three multilingual parallel corpora that vary in linguistic diversity and scale. We conduct a data quality audit to ensure that the data for each language is comparable in quality (App. B.2). **parallel-10** covers 10 languages from 6 families.[3] Each language contributes approximately 20M English-aligned words **parallel-3 (100M)** and **parallel-3 (380M)** are two trilingual corpora comprised of Arabic, Chinese, and English, corresponding with approximately 100M and 380M English words, respectively. They allow us to examine whether tokenization and linguistic effects remain stable under larger-scale training regimes. Together, these corpora reflect a trade-off between linguistic diversity and scale; detailed construction procedures are in App. B.

[3] Arabic, Chinese, English, French, German, Finnish, Polish, Korean, Russian, and Turkish.

3.2 Models & Hyperparameters

Following Yang et al. (2026), we use `GPT2-small` (Brown et al., 2020) as our primary architecture, which achieves lower validation loss than smaller models at our data scale. In scaling experiments, we additionally train `GPT2-medium` (See Tab. 4). Each model is trained separately per language following the hyperparameter settings of Chang et al. (2026), with a batch size and context length of 128. Alternative settings were explored in a pilot study but led to overfitting. All final hyperparameters are reported in App. C.1.

3.3 Tokenization

We manipulate three aspects of tokenization: pretokenization strategy, tokenization algorithm, and vocabulary size.

- **Pretokenization.** We train standard BPE (Sennrich et al., 2016) and SuperBPE (Liu et al., 2025), which allows space within a single token, to examine how pretokenization interacts with different typological properties.
- **Tokenization algorithm.** We experiment with the two dominant tokenization algorithms used in language models: BPE and Unigram (Kudo and Richardson, 2018), in addition to SuperBPE.
- **Vocabulary size.** We train tokenizers with five vocabulary sizes commonly used in language model pretraining, including 8k, 33k, 49k, 66k, and 82k. Rather than scaling vocabulary size according to the number of word types in each language (Mielke et al., 2019), we use fixed absolute vocabulary sizes to enable controlled cross-linguistic comparison.

3.4 Linguistic Features

We operationalize the linguistic factors discussed in §2.2 as follows.

Morphological richness We measure morphological richness using two metrics: type-token ratio and word entropy. Higher values indicate greater

morphological richness for both metrics. We calculate the two metrics using UD treebanks v2.17 (de Marneffe et al., 2021).[4] We conduct Spearman's ranking correlation test and find a strong correlation ($\rho = 0.84$) between the two. For consistency, we z-scored the two measures for each language and obtained the ranking of morphological richness.

Speaker Population We use speaker population estimates from Ethnologue,[5] including both first and second language users.

Distance to English We compute typological distance to English using the `featural` distance metric from URIEL+ (Khan et al., 2025).

Markedness We derive a language-level markedness score from WALS, retaining the 20 typological features attested for all experimental languages.[6] Feature values that are less frequent cross-linguistically receive higher scores, and scores are averaged across features per language.

3.5 Evaluation

We evaluate models on four zero-shot tasks designed around parallel or comparable benchmarks, measuring language modeling, generation, grammatical knowledge, and structural representation. We restrict to zero-shot evaluation as no multilingual parallel fine-tuning benchmarks currently exist, and fine-tuning on language-specific splits would introduce confounds inconsistent with our controlled design.

Negative log-likelihood (NLL) NLL corresponds to the sum of token-level negative log probabilities (i.e., token-level surprisal) over a sequence, i.e., $-\sum_{i=1}^{n} \ln(P(t_i \mid \mathbf{t}_{<i}))$. Unlike perplexity or bits-per-character, which normalize by token or character count and are thus sensitive to tokenization or encoding granularity, NLL reflects the total surprisal of the entire sequence. In our test setting, the evaluation data are parallel across languages. If the languages are equally learnable and encode comparable information, we expect sequences conveying the same content to have similar probabilities, and thus similar NLL values. This metric has been used in several cross-linguistic studies, such as Lin et al. (2024), Costa-Jussà et al. (2022), and Chang et al. (2026). NLL is computed on the held-out test split of the respective training corpus.

Kolmogorov-Smirnov (KS) Metric We adopt the KS-based rank-frequency metric (KS_p) of Meister and Cotterell (2021) to quantify how closely model-generated text matches the distributional properties of natural language. In our experiment, for each model, we sample 40k sentences using top-k sampling ($k = 50$, temperature = 1.0). We lowercase all whitespace-separated tokens and compute empirical token frequencies (for Chinese, we use character as a token). These counts are rank-sorted, and the top 10k ranks are retained. Let $p(k)$ and $q(k)$ denote the normalized rank-frequency distributions for the kth most frequent words from the model output and the test set, respectively. The KS distance is then defined as the supremum of the absolute difference between their cumulative distributions:

$$KS_p = \sup_{1 \le k \le K} \left| \sum_{i=1}^{k} p(i) \; - \; \sum_{i=1}^{k} q(i) \right|, \quad (1)$$

where $K = 10{,}000$. The KS distance lies in $[0, 1]$, with lower values indicating a closer match between the model-generated and natural rank–frequency distributions.

Unlabeled Attachment Score (UAS) We follow Aoyama and Wilcox (2025) to compute UAS for autoregressive language models by identifying the attention head that best captures each dependency relation. We restrict evaluation to six core Universal Dependency (UD) relations[7] that are attested across all experimental languages. The attention head with the highest accuracy per relation is selected, and scores are macro-averaged across relations. Details are provided in App. C.3. For each language, we use 1,000 sentences from the PUD treebanks (Zeman et al., 2018).

MultiBLiMP Accuracy (ACC) We use MultiBLiMP (Jumelet et al., 2026) to test models' grammatical knowledge, specifically in subject-verb agreement, a phenomenon shared by most languages in the world. The benchmark contains minimal pairs for 101 languages. We report accuracy on each language's test set.[8]

3.6 Summary

In total, we train 150 monolingual models on **`parallel-10`** (10 languages $\times$ 3 tokenization al-

[4] For equations, see App. C.2.

[5] `https://www.ethnologue.com/`

[6] See App. E.2 for features included.

[7] `nsubj`, `obj`, `ccomp`, `csubj`, `iobj`, and `xcomp`.

[8] As Chinese and Korean lack subject-verb agreement, we substitute the verb-related category from ZhoBLiMP (Liu et al., 2026) for Chinese and exclude Korean from this evaluation.

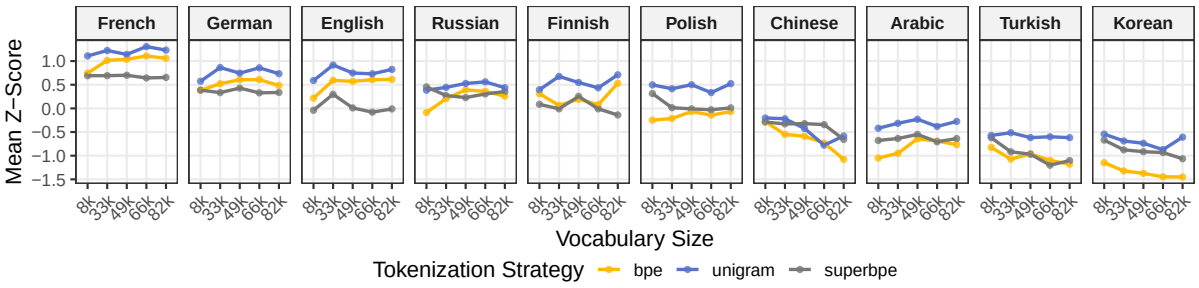


Figure 2: Average z-score across our four language tasks for 10 languages with different tokenization settings

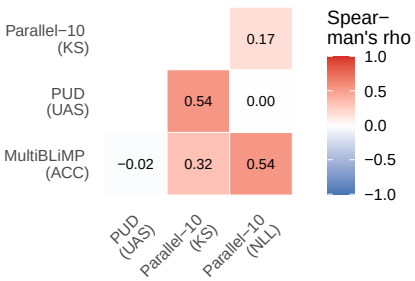


Figure 3: Correlation between metrics

gorithms × 5 vocabulary sizes) and 72 models on **parallel-3 (100M)** and **parallel-3 (380M)** for scaling experiments (2 model sizes × 3 languages × 3 tokenization algorithms × 2 vocabulary sizes). Training data quantity is held constant within each configuration. Models are evaluated on 4 tasks, and analyzed with respect to 3 engineering factors and 4 linguistic features.

4 Main Results

4.1 Correlation Between Metrics

We first computed Spearman's rank correlation coefficients across different metrics. Our analysis reveals moderate-to-weak correlations among the metrics ($\rho < 0.6$), as illustrated in Fig 3. This lack of strong correlation suggests that different evaluation methodologies can yield different model rankings. Additionally, no correlation is observed between NLL and UAS. These findings suggest that relying on a single metric may provide an incomplete picture of model quality.

4.2 Systematic Performance Gaps Persist Under Controlled Conditions

Motivated by this, we analyze the results in two complementary ways. First, we compute an aggregated mean z-score across all four metrics to obtain a holistic measure of overall performance, allowing us to compare languages in terms of their general modeling difficulty. Second, we examine each task separately, since different tasks capture different aspects of language model behavior and may reveal distinct cross-linguistic patterns. Fig. 2 shows the aggregated task scores, and Fig. 9 in the Appendix shows the task-specific ranking.

Across languages, Unigram tokenization leads to higher mean scores, which is consistent with other studies (Bostrom and Durrett, 2020; Vemula et al., 2025). Additionally, different languages have different optimal tokenization settings. Most importantly, there is a systematic gap across languages, with French showing the overall best performance and Korean and Turkish the worst. Note that the highest average scores obtained by a Korean model are lower than the lowest average scores obtained by any French model. Paired t-tests with Holm correction confirm that most language pairs differ significantly; exceptions are largely between typologically similar languages (e.g., German-English, Russian-Polish, see Fig. 16 for more details).

4.3 Linguistic and engineering factors have different effects

Linguistic features predict performance gaps. We fit a linear mixed-effects model to predict z-scored means for each task/condition pair.[9] We did not include `language` as a random effect because it is collinear with the linguistic features. To test the robustness of our findings, we also conducted a leave-one-language-out experiment and fit an item-level mixed-effects model in which each sentence serves as an observation; both yield similar patterns (App. E.1).

First, we find that neither `tokenization` nor `vocab size` is a significant predictor ($p >$

[9] $z \sim$ *tokenization* + *vocab_size_log* + *ttr_z* + *word_entropy_z* + *dist_en_z* + *speakers_z* + *markedness_z* +(*tokenization* + *vocab_size_log* + *dist_en_z* + *speakers_z* + *markedness_z* || *task*); covariates are also z-scored. All linear regression models are fit using the `lme4` package in `R` (Bates et al., 2015)

0.3). We found that morphology metrics (`ttr` and `word_entropy`), number of speakers, and distance from English are significant ($p < 0.05$), all with negative estimates. In leave-one-language-out experiments, given the limited statistical power with only 9 observations, we focus on the consistency of effect directions rather than significance. We find consistent negative estimates for type-token ratio, number of speakers, and distance from English when excluding a language (Fig. 19). We note that the number of speakers is highly correlated with our morphological richness measures (`ttr`: $r = -0.82$; `word_entropy`: $r = -0.73$), which limits our ability to disentangle their independent contributions. We therefore interpret the speaker population effect with caution. To further assess robustness, we refit the model leaving out each linguistic predictor in turn, and find that the negative estimate for distance to English remains consistent across all specifications while other predictors show greater variability (see Figures 20-24). Together, these results suggest that typological distance from English is the most robust predictor: languages more distant from English are systematically harder to model by transformers.

Tokenization effects are language-specific. The null effect of tokenization in the main model likely reflects language-specific optima that cancel out when averaged across languages. To test this, we fit a mixed-effects model, again to predict z-scored task/condition means, this time including the interaction between language and tokenization scheme.[10] BPE-trained English is the baseline factor because of our English-centric bias hypothesis. The statistical results are shown in Fig. 4. Relative to English, Arabic, Korean, and Turkish show significantly lower performance. Tokenization interacts with language systematically: SuperBPE benefits morphologically rich languages (Korean, Polish, Turkish, Russian, Arabic, Finnish) but not English-like languages, while Unigram benefits Korean and Polish. Vocabulary size negatively impacts Chinese, Korean, and Turkish, suggesting that larger vocabularies are not universally beneficial for all languages.

[10] Our `lme4` formula is *z ~ tokenization*language + vocab_size_log*language+ (tokenization + vocab_size_log + language || task)*. As mentioned above, linguistic features are colinear with *language* and therefore excluded.

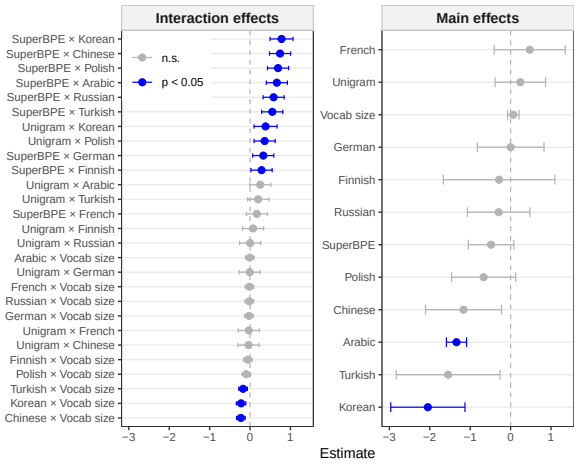


Figure 4: Mixed effect model results (estimate ± 95% CI) on the interaction between language and tokenization settings

5 Training interventions do not close cross-linguistic gaps

Section 4.3 shows that tokenization settings influence language model performance within languages. In this section, we investigate the underlying mechanisms through which tokenization introduces these engineering-driven differences. We focus on two hypotheses motivated by prior work. First, previous studies suggest that tokenizer quality affects downstream performance (e.g., Zouhar et al., 2023; Vemula et al., 2025). Second, because all models are trained for a fixed number of epochs, datasets that yield more tokens (i.e., larger CTC) under a given tokenizer result in more training steps and therefore more parameter updates. We test whether these factors help explain the cross-linguistic performance differences observed above.

5.1 Tokenizer quality measures negatively correlate with model performance

We assess tokenizer quality using three metrics: (1) CTC (Schmidt et al., 2024), which measures compression (higher = worse compression); (2) Rényi Entropy (Zouhar et al., 2023), which measures token distribution uniformity (higher = better utility)[11]; (3) Zipfian α, (Lotz et al., 2025), which measures distribution skewness (higher = less uniform).

We compute all three metrics on the test split of **`parallel-10`** and run Spearman's rank correlations with mean z-score (Fig. 3). We find a negative correlation between tokenizer quality and model per-

[11] We use `tokenization-scorer` from Zouhar et al. (2023) with the recommended setting (α=2.5)

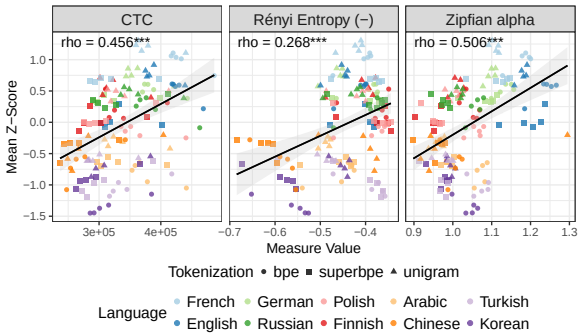


Figure 5: Correlation between tokenizer metrics (CTC, negative Rényi Entropy, Zipfian α) and mean z-score across languages and model configurations.

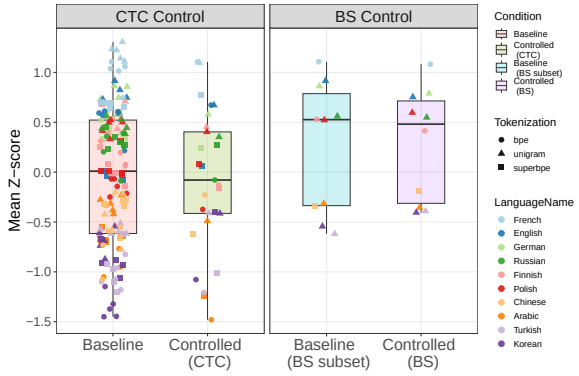


Figure 6: Z-score distribution comparison between the original experiments and experiments with controlled CTC by changing vocab size and with the same training steps by changing the batch size (BS).

formance across all three metrics, contrary to Rust et al. (2021); Zouhar et al. (2023); Lotz et al. (2025), though Schmidt et al. (2024) similarly report inconsistent correlations across settings. Task-level results are reported in App. D.2 (Figures 10–12).

5.2 Controlling for tokenization compression does not close the gap

To test whether compression efficiency alone explains cross-linguistic differences, we conduct a controlled experiment in which CTC in the training dataset is held approximately constant across languages. We use English BPE with a vocabulary size of 100k as a reference point and select vocabulary sizes for each language and tokenizer such that their training-set CTC matches this reference as closely as possible. The vocab size is reported in Tab. 6. Models are then trained and evaluated under identical procedures.

The results are shown in Fig. 6 (left). Despite equalizing CTC, cross-linguistic performance gaps persist. A Levene's test comparing variance across languages between the baseline and controlled conditions finds no significant difference ($F(1, 18) = 0.03$, $p = 0.86$), confirming that compression efficiency alone does not account for the observed disparities. The ranking of languages remains largely the same regardless of whether CTC is controlled, with French, English and German at the top of the distribution and Chinese, Arabic, and Turkish at the bottom.

We attribute this to a fundamental tension in CTC equalization: as shown in Fig 15, matching the English reference requires languages such as Chinese, Korean, Turkish, and Finnish to use substantially smaller vocabularies, resulting in finer-grained and potentially less semantically coherent subword units; equalizing compression introduces a new source of disparity even as it removes another.

5.3 Controlling for training steps does not close the gap

Another possible hypothesis is that the languages with high CTC will receive more training steps if the context length and batch size are the same. To test this, we adjust the batch size for models trained with different languages to make sure that the model will receive an equivalent number of training steps. Results are plotted in Figure 6 (right).

For each language, we select a high-performing tokenization configuration and adjust the batch size based on the CTC of the training dataset so that models are trained for approximately the same number of update steps. The resulting step counts are closely matched across languages ($\sigma = 497$; $CV = 2.73\%$). To assess whether controlling the number of update steps reduces cross-linguistic disparities, we conduct the Pitman-Morgan test for paired comparison. We find cross-linguistic variance decreases under step-controlled training ($\sigma^2_{\text{baseline}} = 0.43$, $\sigma^2_{\text{BS-controlled}} = 0.32$), and this reduction is significant ($t(8) = 2.90$, $p = 0.020$, two-tailed). However, the gap persists, and language rankings remain stable, indicating that engineering choices can partially mitigate but not eliminate the gaps.

6 Scaling Experiments

To examine whether cross-linguistic performance gaps persist at scale, we train `GPT2-medium` on larger datasets (**`parallel-3 (100M)`** & **`parallel-3 (380M)`**). For KS and NLL, we use the FLORES test split rather than in-domain data to ensure

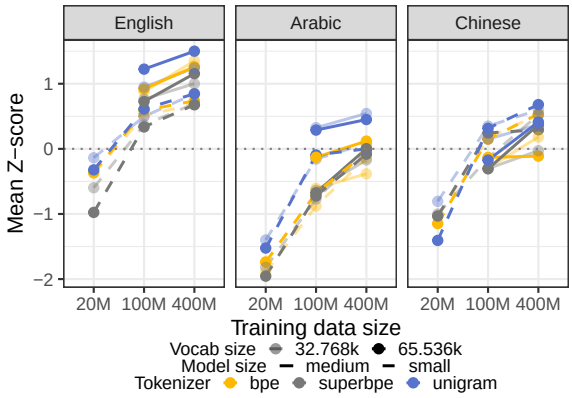


Figure 7: Mean z-score in scaling experiments with different tokenization, model, and training settings

comparability across models trained on different data sizes. The results are visualized in Figure 7. Pairwise paired t-tests after Holm correction show significant differences across all language pairs (Fig. 17). Our mixed-effects model[12] confirms that English ($\beta = 1.16$, $p < 0.001$) and Chinese ($\beta = 0.75$, $p < 0.001$) significantly outperform Arabic. Training data size and model size are both significant positive predictors ($p < 0.05$). However, their effects are language-specific: scaling model size benefits Arabic and English but significantly hurts Chinese ($\beta = -0.55$, $p = 0.006$). This may suggest that larger Chinese models require more training tokens than our parallel corpus provides. We leave this hypothesis for future exploration.

7 Discussion & Conclusion

Implications for NLP. Our scaling experiments reveal a striking contrast: models trained on 100M English words outperform those trained on Arabic data equivalently to 400M English words, suggesting that cross-linguistic performance gaps are not simply a function of data quantity. Our results point to two possible interpretations: either (a) the architecture is sound and better encoding or tokenization schemes could close the gap, or (b) the problem is more systemic; that is, autoregressive attention-based transformers are inherently biased toward certain language types regardless of tokenization/encoding. Our findings more strongly support (b), as within the range of mainstream tokenization approaches examined here, improvements were modest and did not fundamentally alter cross-linguistic rankings.

This raises three important questions for future research. **First**, what in transformers is *causing* this bias? We hypothesize that next-word prediction favors languages with rigid word order: due to the morphology-word order trade-off (Koplenig et al., 2017), morphologically rich languages tend to have more flexible word order, making next-token prediction harder. **Second**, is this a property of language or of our technology? Answering this requires experiments across architecture types beyond transformer decoders.[13] Only convergent evidence across model types would support a stronger claim about linguistic difficulty per se. **Third**, does such bias matter, since we have good LMs in, e.g., Chinese? We believe so, not because functional NLP systems are impossible for non-Indo-European languages, but because of opportunity cost: our architectural defaults make progress inherently harder for languages like Turkish and Arabic, which already have fewer NLP resources, compounding the underinvestment in NLP technology for these communities.

Implications for Cognitive Science. The human language system has two important characteristics: (1) it favors certain "possible" languages over "impossible" ones (Bolhuis et al., 2024; Moro et al., 2023; Chomsky et al., 2023), for example, there are no attested languages that use linear count-based, as opposed to hierarchical rules. And (2), all possible languages are favored largely equally. That is, barring extremely small variation, all natural languages are equally easy to learn and process by humans (e.g., Armon-Lotem et al., 2016). Recent work has tested this first property (Kallini et al., 2024), and found that transformers prefer sequences with natural-language-like properties. However, our results and recent other work (Yang et al., 2025; Mielke et al., 2019; Cotterell et al., 2018; Chang et al., 2024) strongly suggest that transformers fail to capture the second property. While it is certainly a remarkable fact that LMs naturally capture human-language-like preferences, our results suggest that equal cross-linguistic abilities are an important desideratum for cognitive modeling, and that the current generation of models lacks cognitive plausibility in this regard.

[12] *z ~ Tokenization * language + VocabSize_k_s * language + data_size_s * language + ModelSize * language + (Tokenization + VocabSize_k_s + data_size_s || task)*

[13] E.g., encoder-only models trained with masked language modeling would provide a direct test of our hypothesis in Q1: if similar biases persist, the cause is architectural rather than a consequence of the autoregressive objective

Limitations

First, in the Discussion & Conclusion section, we argue that option (b) is more likely. However, admittedly, we cannot fully rule out (a), as alternatives such as WordPiece (Devlin et al., 2019) and BoundlessBPE (Schmidt et al., 2025) remain unexplored. Due to computational constraints, our experiments cover only three tokenization algorithms, leaving alternatives such as WordPiece (Devlin et al., 2019) and more recent designs (Gazit et al., 2025; Ahia et al., 2024; Schmidt et al., 2025) unexplored. We also do not manipulate tokenizer normalizers, and the pre-tokenizer follows the GPT-2 default setting. Future work should examine whether alternative tokenization designs alter the cross-linguistic patterns we observe.

Second, we only evaluate GPT-2-style autoregressive transformers. Whether our conclusions generalize to other architectures, such as encoder-only models or state space models, is unknown.

Third, our study covers 10 languages from 6 families, which, while typologically diverse, cannot exhaustively represent the world's linguistic variety. Extending to more language families, particularly those underrepresented in NLP research, remains an important direction.

Fourth, our evaluation relies partly on UD treebanks for UAS and MultiBLiMP, whose annotation quality and coverage vary across languages, potentially introducing evaluation noise. Regarding MultiBLiMP specifically, subject–verb agreement differs substantially across languages: English agreement involves primarily number, whereas Arabic additionally encodes gender, person, and other morphological distinctions. However, we argue that a fair evaluation of Arabic grammatical knowledge should include all agreement categories, as these reflect the actual complexity that a language model must learn to model Arabic correctly.

Finally, the non-Indo-European languages in our study are predominantly lower-resourced, raising questions about data quality. We invited native speakers of Arabic, Chinese, Korean, German, and French to evaluate data quality and found no correlation with model performance, suggesting this is not a major confound. Details see App. B.2

References

Ananth Agarwal, Jasper Jian, Christopher D Manning, and Shikhar Murty. 2025. Mechanisms vs. outcomes: Probing for syntax fails to explain performance on targeted syntactic evaluations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33737–33757, Suzhou, China. Association for Computational Linguistics.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Valentin Hofmann, Tomasz Limisiewicz, Yulia Tsvetkov, and Noah A Smith. 2024. Magnet: Improving the multilingual fairness of language models with adaptive gradient-based tokenization. *Advances in neural information processing systems*, 37:47790–47814.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Buschhoff, Charvi Jain, Alexander Weber, Lena Jurkschat, Hammam Abdelwahab, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Samuel Weinbach, Rafet Sifa, and 2 others. 2024. Tokenizer choice for LLM training: Negligible or crucial? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3907–3924, Mexico City, Mexico. Association for Computational Linguistics.

Mohamed Taher Alrefaie, Nour Eldin Morsy, and Nada Samir. 2024. Exploring tokenization strategies and vocabulary sizes for enhanced arabic language models. *arXiv preprint arXiv:2403.11130*.

Tatsuya Aoyama and Ethan Wilcox. 2025. Language models grow less humanlike beyond phase transition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24938–24958, Vienna, Austria. Association for Computational Linguistics.

Sharon Armon-Lotem, Ewa Haman, Kristine Jensen de López, Magdalena Smoczynska, Kazuko Yatsushiro, Marcin Szczerbinski, Angeliek Van Hout, Ineta Dabašinskienė, Anna Gavarró, Erin Hobbs, and 1 others. 2016. A large-scale cross-linguistic investigation of the acquisition of passive. *Language acquisition*, 23(1):27–56.

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.

Catherine Arnett, Tyler A. Chang, and Benjamin Bergen. 2024a. A bit of a problem: Measurement disparities in dataset sizes across languages. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group*

on Under-resourced Languages @ LREC-COLING 2024, pages 1–9, Torino, Italia. ELRA and ICCL.

Catherine Arnett, Tyler A. Chang, Stella Biderman, and Ben Bergen. 2025. Explaining and mitigating crosslingual tokenizer inequities. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Catherine Arnett, Pamela D. Rivière, Tyler A. Chang, and Sean Trott. 2024b. Different tokenization schemes lead to comparable performance in Spanish number agreement. In *Proceedings of the 21st SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 32–38, Mexico City, Mexico. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

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.

Johan J Bolhuis, Stephen Crain, Sandiway Fong, and Andrea Moro. 2024. Three reasons why ai doesn't model human language. *Nature*, 627(8004):489.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624, Online. Association for Computational Linguistics.

Mika Braginsky. 2022. *jglmm: Generalized Mixed-Effects Models in Julia*. R package version 0.1.0.9001.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2024. When is multilinguality a curse? language modeling for 250 high- and low-resource languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4074–4096, Miami, Florida, USA. Association for Computational Linguistics.

Tyler A Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K Bergen. 2026. Goldfish: Monolingual language models for 350 languages. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, Palma de Mallorca, Spain. ELRA and ICCL.

Pavel Chizhov, Catherine Arnett, Elizaveta Korotkova, and Ivan P. Yamshchikov. 2024. BPE gets picky: Efficient vocabulary refinement during tokenizer training. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16587–16604, Miami, Florida, USA. Association for Computational Linguistics.

Noam Chomsky, Ian Roberts, and Jeffrey Watumull. 2023. Noam chomsky: The false promise of chatgpt. *The New York Times*, 8(4):177–179.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49(2):375–395.

Marco Cognetta, Vilém Zouhar, Sangwhan Moon, and Naoaki Okazaki. 2024. Two counterexamples to tokenization and the noiseless channel. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16897–16906, Torino, Italia. ELRA and ICCL.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

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

Jennifer Culbertson, Julie Franck, Guillaume Braquet, Magda Barrera Navarro, and Inbal Arnon. 2020. A learning bias for word order harmony: Evidence from speakers of non-harmonic languages. *Cognition*, 204:104392.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthias Gallé. 2019. Investigating the effectiveness of BPE: The power of shorter sequences. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1375–1381, Hong Kong, China. Association for Computational Linguistics.

Bar Gazit, Shaltiel Shmidman, Avi Shmidman, and Yuval Pinter. 2025. Splintering nonconcatenative languages for better tokenization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22405–22417, Vienna, Austria. 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.

Lila R Gleitman and Elissa L Newport. 1995. The invention of language by children: Environmental and biological influences on the acquisition of language. *An invitation to cognitive science*, 1:1–24.

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.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Yutong Hu, Quzhe Huang, Mingxu Tao, Chen Zhang, and Yansong Feng. 2024. Can perplexity reflect large language model's ability in long text understanding? In *The Second Tiny Papers Track at ICLR 2024*.

Cliodhna Hughes, Jennifer Culbertson, and Simon Kirby. 2025. Evidence for word order harmony between abstract categories in silent gesture. *Cognition*, 259:106100.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2026. MultiBLiMP 1.0: A massively multilingual benchmark of linguistic minimal pairs. *Transactions of the Association for Computational Linguistics*, 14:193–216.

Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. 2024. Mission: Impossible language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14691–14714, Bangkok, Thailand. Association for Computational Linguistics.

Aditya Khan, Mason Shipton, David Anugraha, Kaiyao Duan, Phuong H. Hoang, Eric Khiu, A. Seza Doğruöz, and En-Shiun Annie Lee. 2025. URIEL+: Enhancing linguistic inclusion and usability in a typological and multilingual knowledge base. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6937–6952, Abu Dhabi, UAE. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86.

Alexander Koplenig, Peter Meyer, Sascha Wolfer, and Carolin Müller-Spitzer. 2017. The statistical trade-off between word order and word structure–large-scale evidence for the principle of least effort. *PloS one*, 12(3):e0173614.

Alexander Koplenig and Sascha Wolfer. 2023. Languages with more speakers tend to be harder to (machine-) learn. *Scientific Reports*, 13(1):18521.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, and 33 others. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

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

Sander Land and Catherine Arnett. 2025. BPE stays on SCRIPT: Structured encoding for robust multilingual pretokenization. In *Tokenization Workshop at ICML*.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André FT Martins, and Hinrich Schütze. 2024. Mala-500: Massive language adaptation of large language models. *arXiv preprint arXiv:2401.13303*.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Alisa Liu, Jonathan Hayase, Valentin Hofmann, Sewoong Oh, Noah A. Smith, and Yejin Choi. 2025. SuperBPE: Space travel for language models. In *Second Conference on Language Modeling*.

Hong Liu, Sang Michael Xie, Zhiyuan Li, and Tengyu Ma. 2023. Same pre-training loss, better downstream: Implicit bias matters for language models. In *International Conference on Machine Learning*, pages 22188–22214. PMLR.

Yikang Liu, Yeting Shen, Hongao Zhu, Lilong Xu, Zhiheng Qian, Siyuan Song, Kejia Zhang, Jialong Tang, Pei Zhang, Baosong Yang, and 1 others. 2026. A systematic assessment of language models with linguistic minimal pairs in chinese. *Transactions of the Association for Computational Linguistics*, 14:755–771.

Jonas F. Lotz, António V. Lopes, Stephan Peitz, Hendra Setiawan, and Leonardo Emili. 2025. Beyond text compression: Evaluating tokenizers across scales. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32155–32173, Vienna, Austria. Association for Computational Linguistics.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163, Reykjavik, Iceland. European Language Resources Association (ELRA).

Clara Meister and Ryan Cotterell. 2021. Language model evaluation beyond perplexity. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5328–5339, Online. Association for Computational Linguistics.

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.

Andrea Moro, Matteo Greco, and Stefano F Cappa. 2023. Large languages, impossible languages and human brains. *Cortex*, 167:82–85.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Yoon Mi Oh and François Pellegrino. 2023. Towards robust complexity indices in linguistic typology: A corpus-based assessment. *Studies in Language*, 47(4):789–829.

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.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. *Advances in neural information processing systems*, 36:36963–36990.

Wessel Poelman and Miryam de Lhoneux. 2026. Form and meaning in intrinsic multilingual evaluations. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2503–2521, Rabat, Morocco. Association for Computational Linguistics.

Krithika Ramesh, Sunayana Sitaram, and Monojit Choudhury. 2023. Fairness in language models beyond English: Gaps and challenges. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2106–2119, Dubrovnik, Croatia. Association for Computational Linguistics.

Varshini Reddy, Craig W Schmidt, Yuval Pinter, and Chris Tanner. 2025. How much is enough? the diminishing returns of tokenization training data. In *Tokenization Workshop at ICML*.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Craig W Schmidt, Varshini Reddy, Chris Tanner, and Yuval Pinter. 2025. Boundless byte pair encoding: Breaking the pre-tokenization barrier. In *Second Conference on Language Modeling*.

Craig W Schmidt, Varshini Reddy, Haoran Zhang, Alec Alameddine, Omri Uzan, Yuval Pinter, and Chris Tanner. 2024. Tokenization is more than compression. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 678–702, Miami, Florida, USA. Association for Computational Linguistics.

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

Chen Shani, Yuval Reif, Nathan Roll, Dan Jurafsky, and Ekaterina Shutova. 2026. The roots of performance disparity in multilingual language models: Intrinsic modeling difficulty or design choices? *arXiv preprint arXiv:2601.07220*.

Sho Takase, Ryokan Ri, Shun Kiyono, and Takuya Kato. 2025. Large vocabulary size improves large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1015–1026, Vienna, Austria. Association for Computational Linguistics.

Chaofan Tao, Qian Liu, Longxu Dou, Niklas Muennighoff, Zhongwei Wan, Ping Luo, Min Lin, and Ngai Wong. 2024. Scaling laws with vocabulary: Larger models deserve larger vocabularies. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Cagri Toraman, Eyup Halit Yilmaz, Furkan Şahinuç, and Oguzhan Ozcelik. 2023. Impact of tokenization on language models: An analysis for turkish. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–21.

Matej Ulčar, Aleš Žagar, Carlos S Armendariz, Andraž Repar, Senja Pollak, Matthew Purver, and Marko Robnik-Šikonja. 2025. Mono-and cross-lingual evaluation of representation language models on less-resourced languages. *Computer Speech & Language*, page 101852.

Omri Uzan, Craig W. Schmidt, Chris Tanner, and Yuval Pinter. 2024. Greed is all you need: An evaluation of tokenizer inference methods. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 813–822, Bangkok, Thailand. Association for Computational Linguistics.

Rik van Noord, Miquel Esplà-Gomis, Malina Chichirau, Gema Ramírez-Sánchez, and Antonio Toral. 2025. Quality beyond a glance: Revealing large quality differences between web-crawled parallel corpora. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1824–1838, Abu Dhabi, UAE. Association for Computational Linguistics.

Saketh Reddy Vemula, Sandipan Dandapat, Dipti Sharma, and Parameswari Krishnamurthy. 2025. Rethinking tokenization for rich morphology: The dominance of unigram over BPE and morphological alignment. In *The 14th International Joint Conference on Natural Language Processing and The 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 232–252, Mumbai, India. Association for Computational Linguistics.

Jennifer C. White and Ryan Cotterell. 2021. Examining the inductive bias of neural language models with artificial languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 454–463, Online. Association for Computational Linguistics.

Xiulin Yang, Tatsuya Aoyama, Yuekun Yao, and Ethan Wilcox. 2025. Anything goes? a crosslinguistic study of (im)possible language learning in LMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26058–26077, Vienna, Austria. Association for Computational Linguistics.

Xiulin Yang, Arianna Bisazza, Nathan Schneider, and Ethan Gotlieb Wilcox. 2026. A unified assessment of the poverty of the stimulus argument for neural language models. *arXiv preprint arXiv:2602.09992*.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

Vilém Zouhar, Clara Meister, Juan Gastaldi, Li Du, Mrinmaya Sachan, and Ryan Cotterell. 2023. Tokenization and the noiseless channel. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5184–5207, Toronto, Canada. Association for Computational Linguistics.

A The Tokenization Problem in Multilingual NLP

With the rise of multilingual NLP, growing research has focused on the linguistic unfairness introduced by tokenization (Ali et al., 2024; Petrov et al., 2023; Arnett and Bergen, 2025). Several new tokenization algorithms have been proposed to address these

issues, including SuperBPE (Liu et al., 2025), PickyBPE (Chizhov et al., 2024), and BoundlessBPE (Schmidt et al., 2025). Here we provide a more detailed review of how pre-tokenization strategies, tokenization algorithms, and vocabulary sizes interact with different languages, complementing the discussion in §2.

Pre-tokenization Most mainstream language models adopt space-based pre-tokenization strategies, which assume that words in all languages are delimited by spaces. However, this assumption does not hold for languages such as Chinese and Japanese, which do not use spaces between words. Furthermore, morphologically rich languages often attach functional elements directly to content words, whereas fusional languages like English represent them as separate tokens. After space-based pre-tokenization, English yields frequent standalone function words that are easily included in the tokenizer vocabulary, while languages like Turkish produce morphologically complex tokens that are harder to represent. Arnett et al. (2025) show that even when training tokenizers on the same amount of parallel data with the same vocabulary size, compression rates differ across languages, and that whitespace pre-tokenization is one of the primary causes of this disparity.

Tokenization algorithm The optimal tokenization algorithm varies across languages. Bostrom and Durrett (2020) find that Unigram better respects morphological boundaries compared to BPE. Ali et al. (2024) show that Germanic languages benefit from BPE while Romance languages perform better with Unigram, though their findings are limited to Indo-European languages. For Japanese, Bostrom and Durrett (2020) find that BPE and Unigram outperform WordPiece regardless of task type. For morphologically rich languages such as Telugu and Hindi, Vemula et al. (2025) find that Unigram generally outperforms other algorithms, though hybrid tokenizers incorporating morphological segmentation improve performance within the BPE framework. Arnett and Bergen (2025) show that morphology-aligned tokenization does not necessarily reduce perplexity for agglutinative languages, and follow-up work (Shani et al., 2026) discuss morphologically rich languages interact differently with tokenizers due to their complex inflectional systems.

Vocabulary size The impact of vocabulary size on performance remains mixed. Tao et al. (2024) and Takase et al. (2025) suggest that larger vocabularies benefit larger models, though their findings are based on English-centric settings. Alrefaie et al. (2024) find that for Arabic, vocabulary size has little effect on downstream performance when model size is fixed. Mielke et al. (2019) show that vocabulary size influences surprisal rankings across languages, and note that setting vocabulary size proportional to word type count leads to disproportionately large vocabularies for morphologically rich languages. For Turkish, Toraman et al. (2023) find that allocating approximately 20% of model parameters to vocabulary parameters yields a reasonable trade-off between model size and performance. For Russian, Reddy et al. (2025) find that compared to English, Russian requires approximately 33% more training data for the tokenizer to reach comparable performance.

B Parallel Multilingual Corpus

The training data sources and split can be found in Tab. 1 and Tab. 2.

B.1 Postprocessing

To guarantee the quality of the parallel corpus, we applied the following filtering steps:

- Exclude web-crawled parallel datasets, as their quality is often suboptimal (van Noord et al., 2025).
- Run `FastText` (Grave et al., 2018) language identification on the non-English side and remove sentences that are predicted to be English.[14]
- Remove sentences with excessive punctuation (punctuation ratio over sentence length > 0.5).
- Remove sentences shorter than four characters.
- For Chinese, Korean, and Arabic (which normally do not use Latin script), exclude sentences in which more than 50% of characters are Latin letters.

[14] We observed that `FastText` frequently produces false positives for Chinese, i.e., it mistakenly classifies Chinese sequences as English. To retain sufficient data, we additionally checked the penultimate character of each Chinese sentence classified as English, keeping those whose character type is Han rather than Latin.

Source	Description	parallel-10		parallel-3	
		Words	Sents	Words	Sents
OpenSubtitles (Lison and Tiedemann, 2016)	Movie/TV show subtitles	17.0M	2.0M	64.9M	8.1M
Bible (Christodouloupoulos and Steedman, 2015)	Bible text	1.5M	57K	1.5M	61.2K
WikiTitle	Titles of Wikipedia articles	167K	42K	848.2K	218.3K
Neulab-Ted	Ted talks	356K	19K	3.7M	201.2K
KDE4	KDE4 localization files	117K	13K	280.7K	30.0K
Ted2020 (Reimers and Gurevych, 2020)	Ted talks	347K	19K	3.1M	171.0K
QED	Subtitles for educational videos	10K	0.6K	73.5K	5.0K
Ubuntu	Ubuntu localization files	<1K	<100	2.5K	<1K
Tatoeba	Translated sentences from Tatoeba	<1K	<100	20.2K	2.9K
GNOME	GNOME localization files	<1K	<100	<1K	<100
UNPC (Ziemski et al., 2016)	United Nation documents	NA	NA	306.0M	10.7M
GlobalVoices	News report	NA	NA	156.0K	7.7K
NewsCommentary (Barrault et al., 2019)	news comments	NA	NA	2.5M	55.6K
Ted2013	Ted talks	NA	NA	983.9K	171.0K
Tanzil	Quran translations	NA	NA	14.2K	1.7K
Overall		19.5M	2.4M	384.2M	19.6M

Table 1: Sources of the parallel-10 and parallel-3 corpora.

Dataset	Language	Train		Dev		Test	
		#sent	#word	#sent	#word	#sent	#word
parallel-10	English	2,338,286	19,089,154	10,000	82,457	40,000	327,724
	Arabic	2,338,286	14,426,353	10,000	61,982	40,000	246,954
	Chinese	2,338,286	30,129,452	10,000	130,205	40,000	516,401
	Turkish	2,338,286	12,540,711	10,000	54,238	40,000	215,345
	German	2,338,286	16,544,740	10,000	71,464	40,000	283,997
	Korean	2,338,286	11,857,382	10,000	51,292	40,000	203,336
	Polish	2,338,286	13,458,472	10,000	57,932	40,000	231,262
	French	2,338,286	17,603,323	10,000	76,087	40,000	302,056
	Russian	2,338,286	14,613,449	10,000	62,951	40,000	251,018
	Finnish	2,338,286	11,975,617	10,000	51,698	40,000	205,766
parallel-3-100M	English	5,100,382	99,999,986	10,000	201,696	40,000	785,561
	Arabic	5,100,382	86,689,681	10,000	175,034	40,000	681,070
	Chinese	5,100,382	176,894,596	10,000	359,597	40,000	1,391,600
parallel-3-380M	English	19,555,763	383,187,466	10,000	198,658	40,000	780,968
	Arabic	19,555,763	332,197,948	10,000	171,796	40,000	676,630
	Chinese	19,555,763	677,898,235	10,000	349,927	40,000	1,380,368

Table 2: Data splits for three experimental settings across ten languages. Word counts are computed using whitespace-separated tokens for all languages except Chinese, for which each character is treated as a word.

- Discard sentence pairs whose target side differs from the source side by only two or fewer unigram tokens, as these are likely misaligned or untranslated.
- Remove all invisible characters

For Chinese, we find there are mixed traditional Chinese characters in our text. We therefore use OpenCC[15] to convert traditional Chinese characters to simple Chinese characters. For Bible, each character is separated by space, we remove the space between characters.

After postprocessing, for the parallel-10 corpus, we collect 19.5M words data for each language; for the parallel-3 corpus, we collect 385M words data for each language.

The sequence length distribution for each language is shown in Fig. 8.

B.2 Data quality checking

We invite native speaker volunteers speaking Arabic, French, German, Korean, and Chinese, to evaluate 200 parallel samples under two criteria: **Strict**, which flags any deviation including meaning changes, missing content, unnatural phrasing, grammatical errors, typos, and machine translation artifacts; and **Loose**, which flags only severe errors including irrelevant translations, grammatical errors, and typos. Error rates are reported in Tab. 3.

[15] https://github.com/BYVoid/OpenCC

Language	Strict	Loose
Arabic	16/200	8/200
French	16/200	6/200
German	18/200	5/200
Korean	21/200	5/200
Chinese	15/200	3/200

Table 3: Error rates under strict and loose evaluation settings.

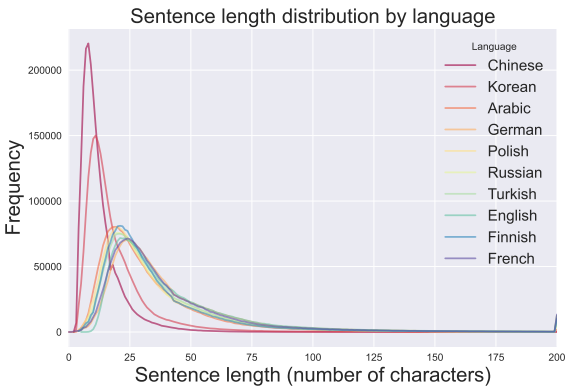


Figure 8: Length distribution (measured by characters) of each language in parallel-10 dataset

Under the loose criterion, error rates are acceptably low across all languages. Crucially, data quality rankings do not correlate with model performance rankings, suggesting that data quality is not a confounding factor in our results.

C Experiment Settings

C.1 Hyperparameter settings

The model architecture experimented is reported in Tab. 4 and hyperparameter settings are listed in Tab. 5.

C.2 TTR & Word Entropy

Following Oh and Pellegrino (2023); Bentz et al. (2016), we define *type-token ratio (TTR)* as:

$$C_{\text{TTR}} = \frac{V}{\sum_{i=1}^{V} f_i} \tag{2}$$

where V is the vocabulary size and f_i the frequency of word type i. We define *word entropy* as

$$H(T) = -\sum_{i=1}^{V} p(w_i) \log_2 p(w_i) \tag{3}$$

where T is the corpus, $N = \sum_{i=1}^{V} f_i$ is the total token count, and $p(w_i) = \frac{f_i}{N}$.

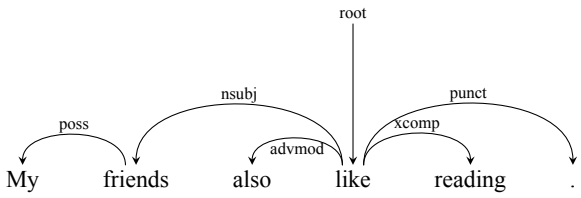
Configuration	GPT2-small	GPT2-medium
Hidden size	768	1024
Number of layers	12	24
Number of heads	12	16

Table 4: Model configurations used in the experiments.

Training	
train_size	10M
num_epoch	10
train_amount	19M
batch_size	128
context_length	128
grad_acc_steps	1
weight_decay	0.1
warmup_steps	10%
lr	5e-4
lr_scheduler	linear
GPU	8*A100 GPUs

Table 5: Hyperparameter Settings

C.3 Evaluation: UAS



In a dependency tree under the UD framework (Nivre et al., 2020), every word (except the root) has exactly one syntactic head. As head and complement are important concept in syntax, a better language model is expected to learn more about the underlying syntactic rules. By investigating on the attention of the models trained on different languages, we aim to evaluate how well the models grasp the concept of head and complement.

We define our evaluation approach as follows. Given a target word x_i in a sentence, we define a head-specific probe $f_{h,l}$ that predicts the parent word of x_i using the attention map a. The prediction is made by selecting the word x_j (where $j < i$ due to the autoregressive training scheme) that receives the highest attention from or to x_i at attention head h and layer l. Due to the autoregressive processing of the model, sometimes the head follows the dependent (for example, *friends* and *like* in the example tree). Therefore, in our evaluation, we do not consider the direction of the dependency but take this dependency relation as symmetrical. This is denoted as:

$$f_{h,l}(x_i) = \arg\max_{x_j < x_i} \left(a_{ij}^{(h,l)} \right). \tag{4}$$

Language	BPE	SuperBPE	Unigram
AR	10,879	7,909	8,495
DE	12,310	6,674	9,527
EN	100,000	6,990	14,196
FR	70,255	6,651	13,582
TR	6,137	5,826	4,271
RU	12,340	8,008	11,703
FI	5,781	5,102	5,388
PL	7,591	6,137	8,177
ZH	3,318	2,757	—
KO	4,169	4,263	5,994

Table 6: Vocabulary sizes by language and tokenization setting.

Since the probe $f_{h,l}$ is head-specific, we evaluate all heads for each dependency relation and report the score for the head–layer pair that yields the highest UAS. The overall UAS is then computed across all dependency relations. Besides, as different languages have different linguistic phenomena which are hard to harmonize by UD framework, we therefore choose only the six core grammatical function cross-linguistically defined in UD. Due to the use of subword tokenization, we convert subtoken-level attention to word-level attention by averaging over the attention scores of all subtokens corresponding to each word.

C.4 Tokenizer-controlled settings

Language	Tokenizer	Vocab Size	CTC	Batch Size
French	BPE	65,536	24,440,482	128
English	Unigram	32,768	22,593,787	118
Korean	Unigram	8,192	22,350,404	117
German	Unigram	32,768	21,058,271	110
Russian	Unigram	65,536	19,747,070	103
Arabic	Unigram	32,768	19,712,087	103
Polish	Unigram	81,920	18,185,007	95
Finnish	BPE	81,920	18,044,106	94
Turkish	Unigram	81,920	16,234,672	85
Chinese	SuperBPE	65,536	14,157,276	74

Table 7: Optimal Tokenization and Batch Size Configuration per Language

D Task-specific results

D.1 Task-specific ranking of Mean z-score

See Fig. 9.

D.2 Correlation between model performance and tokenizer measures

See Figures 10-12.

Language	TTR	Word Entropy	Dist. to EN	Speakers (M)	Markedness
French	0.08	10.92	0.45	333.5	0.59
English	0.04	10.15	0	1500.0	0.60
German	0.06	11.56	0.40	133.4	0.66
Russian	0.09	13.20	0.50	210.3	0.57
Chinese	0.06	11.16	0.49	1200.0	0.58
Arabic	0.11	11.16	0.49	334.9	0.59
Polish	0.18	12.66	0.51	45.5	0.60
Finnish	0.26	13.53	0.47	5.6	0.58
Turkish	0.17	13.56	0.57	93.7	0.57
Korean	0.27	14.51	0.53	82.2	0.56

Table 8: Typological and demographic features of the 10 languages used in this study. TTR = type-token ratio; Dist. to EN = typological distance to English (Khan et al., 2025); Speakers = estimated number of speakers in millions; Markedness = mean markedness score.

E Other results & Analysis

E.1 Main Mixed Effect Model

We report the main mixed-effects model results in Tab. 9 and visualize the leave-one-language-out results in Fig. 19. We additionally conduct an item-level analysis to verify the robustness of our findings. Sentence-level NLL and UAS are computed directly; for UAS, relations absent from a given sentence are excluded from the average. For Multi-BLiMP, we use the NLL difference between correct and incorrect minimal pair sentences as the dependent variable. For the KS metric, we run 1,000 sampling iterations per model and randomly sample 750 observations per task, language, and model configuration to ensure balanced data. Due to the large sample size, we fit the mixed-effects model using `jglmm` (Braginsky, 2022) with the formula:

`value ~ tokenizer + vocab_size_k_s + ttr_s + markedness_s + word_entropy_s + speakers_s + dist_en_s + (0 + tokenizer | language) + (1 | metric) + (0 + tokenizer | metric) + (0 + vocab_size_k_s | metric) + (0 + word_entropy_s | metric) + (0 + ttr_s | metric) + (0 + markedness_s | metric) + (0 + speakers_s | metric) + (0 + dist_en_s | metric)`

Results are shown in Tab. 10, and the corresponding leave-one-language-out results in Fig. 18. The item-level analysis yields patterns consistent with the main analysis.

E.2 Linguistic Feature Results

For markedness measure, we include the following WALS features: 81A, 82A, 83A, 87A, 88A, 89A,

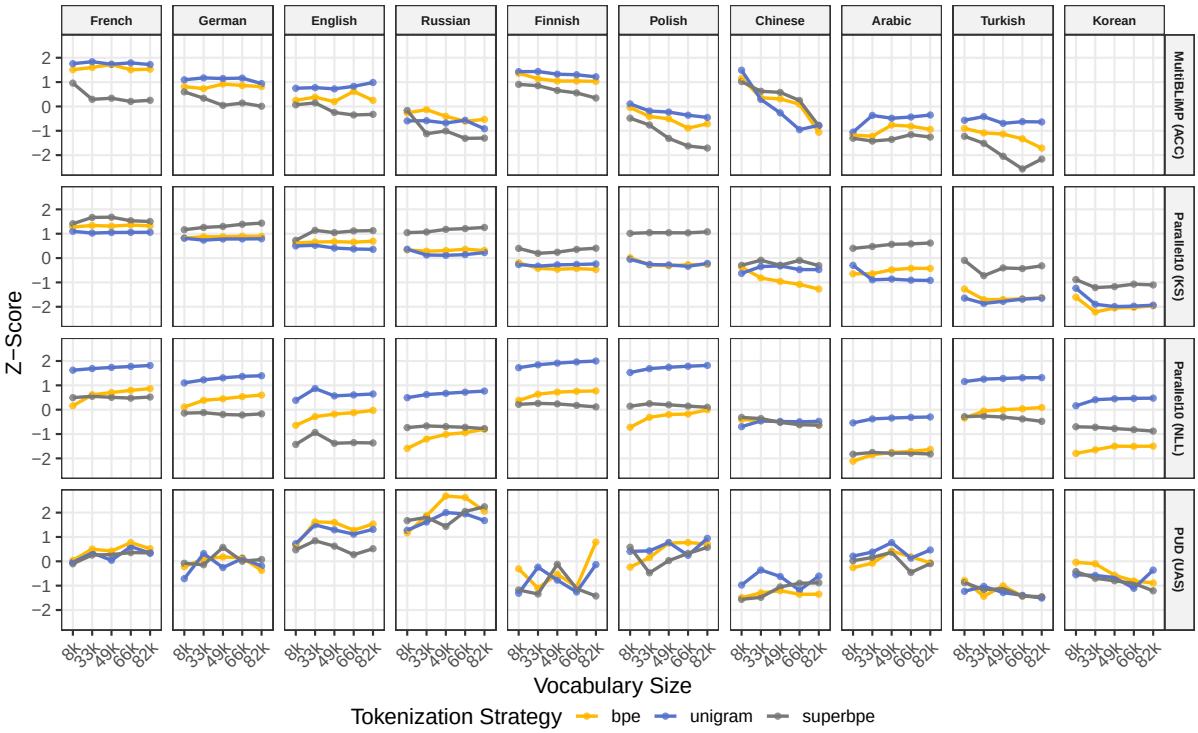


Figure 9: Task-specific ranking across languages

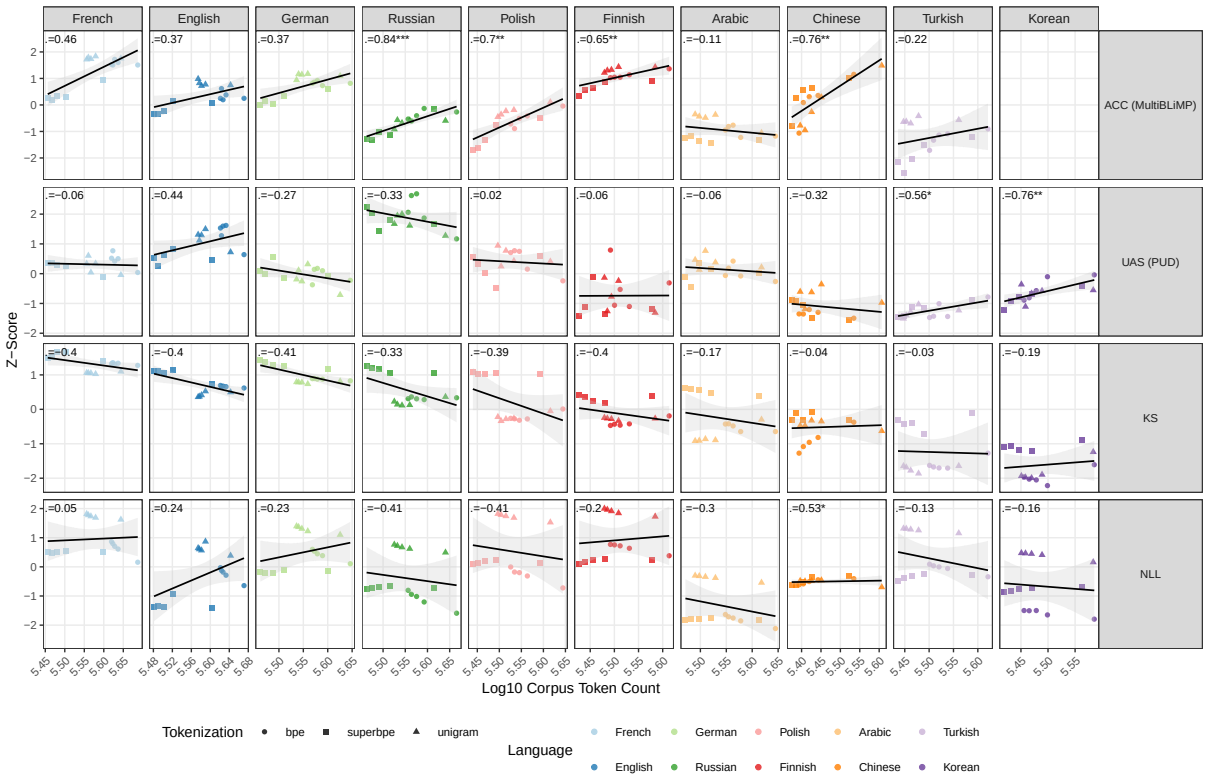


Figure 10: Task-level correlation between CTC and model performance

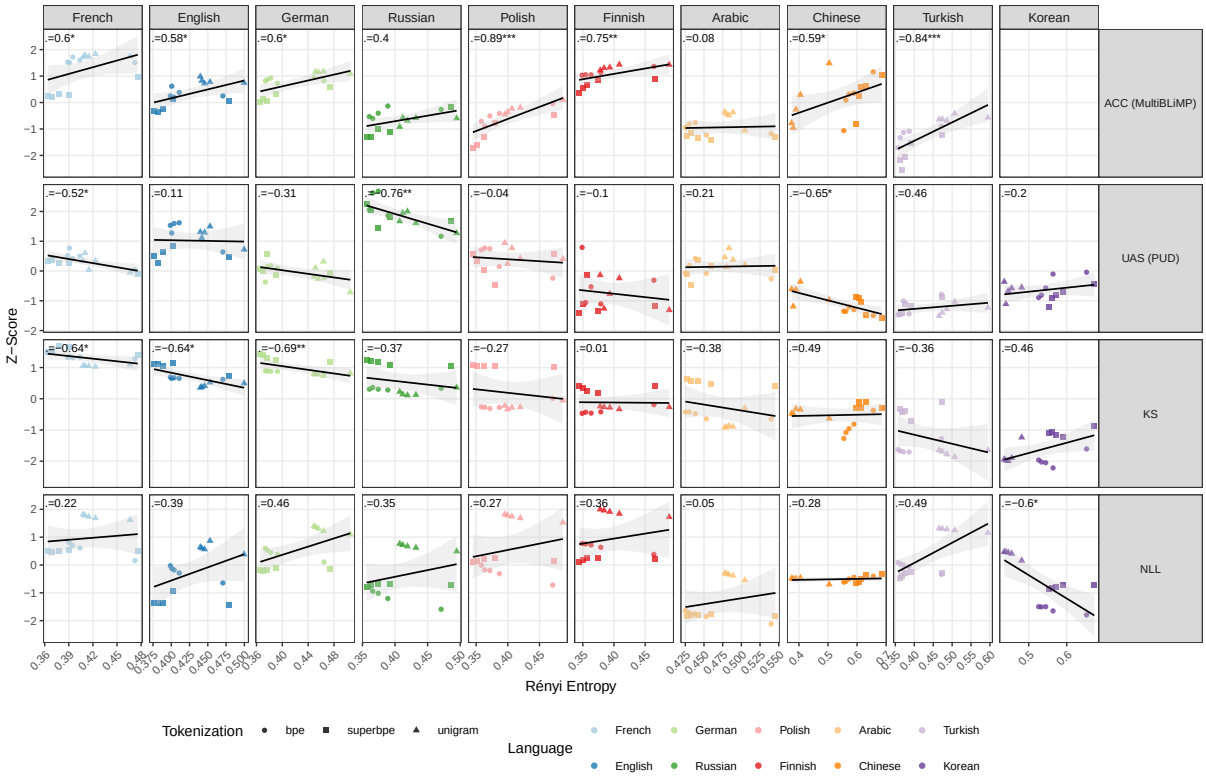


Figure 11: Task-level correlation between Rényi entropy and model performance

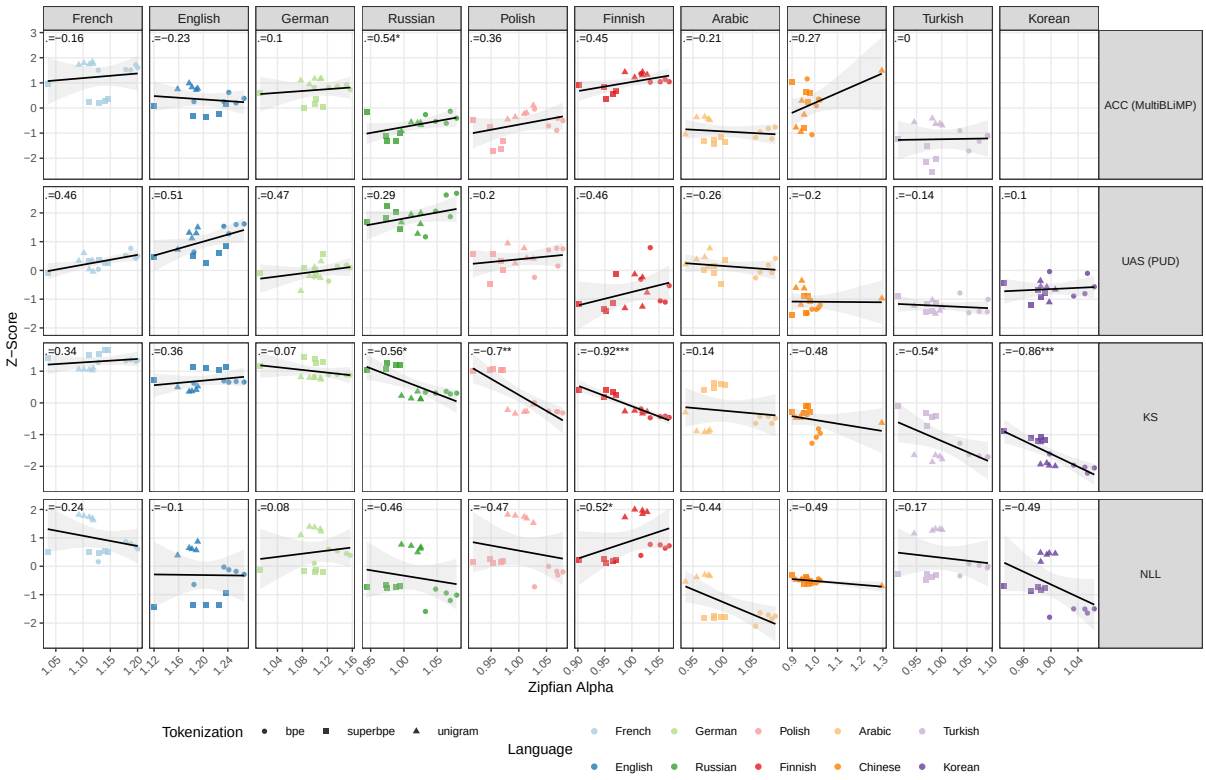


Figure 12: Task-level correlation between Zipfian α and model performance

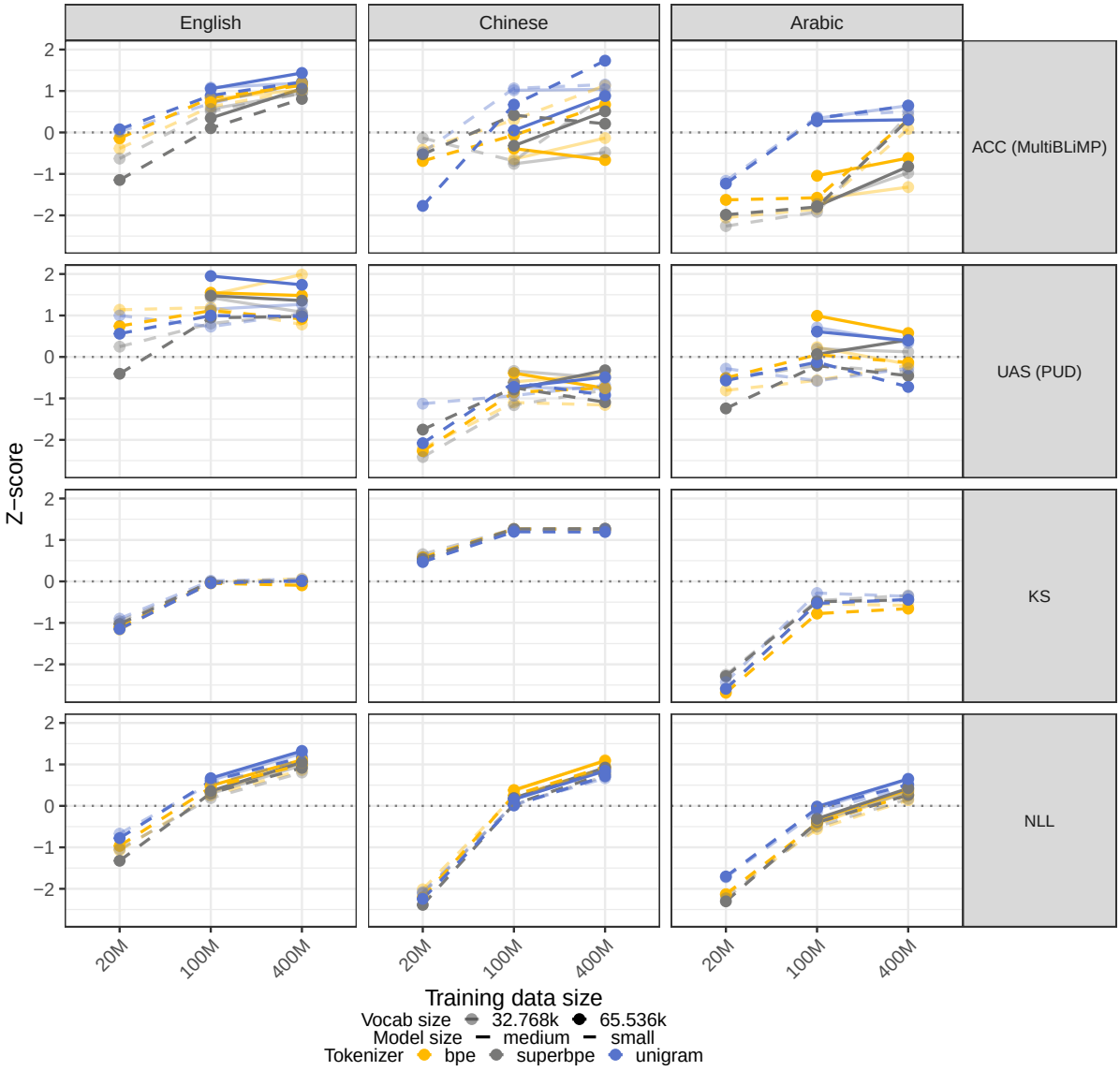


Figure 13: Scaling experiment per task

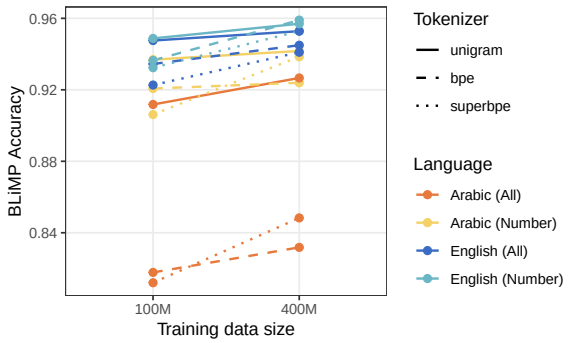


Figure 14: MultiBLiMP accuracy for English and Arabic by GPT2-medium under two evaluation conditions: all agreement categories (*Language All*) and number agreement only (*Language Number*). Restricting to number agreement substantially narrows the gap, though a difference persists.

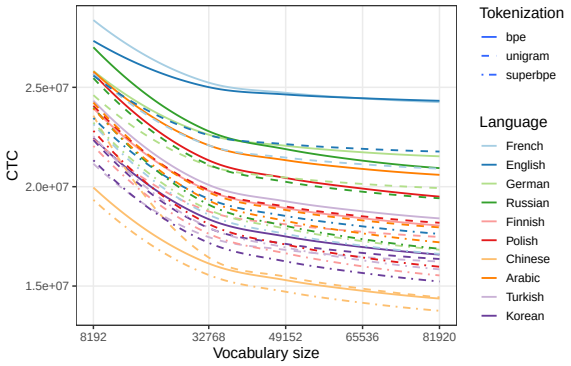


Figure 15: CTC across vocab size, languages, and tokenziation algorithms

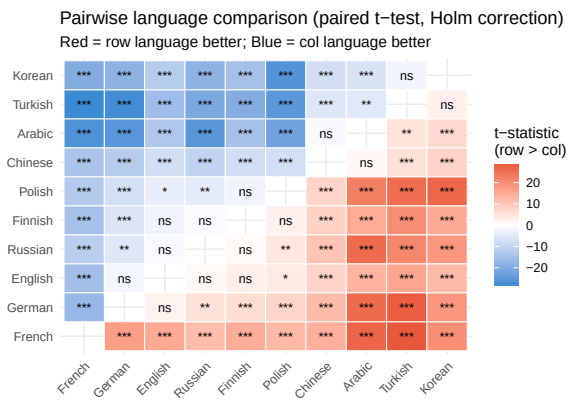


Figure 16: Paired t-test between language pairs (10 languages in total)

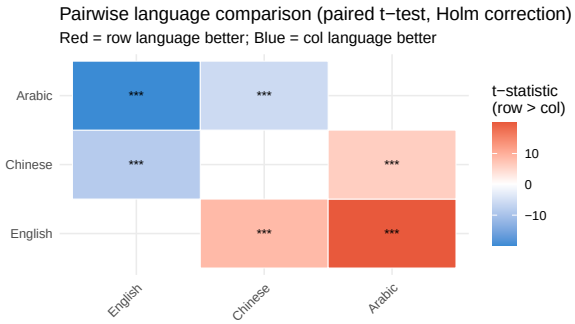


Figure 17: Paired t-test between language pairs (3 languages in total)

Predictor	Estimate	Std. Error	t value	p value
Intercept	-0.118	0.107	-1.106	0.348
Unigram	0.354	0.307	1.154	0.332
SuperBPE	-0.014	0.278	-0.051	0.963
Vocabulary size (log)	-0.011	0.058	-0.193	0.859
TTR (z)	-0.761	0.079	-9.672	$< .001$
Word entropy (z)	-0.189	0.063	-2.988	0.003
Markedness (z)	-0.110	0.103	-1.067	0.350
Distance to English (z)	-0.219	0.061	-3.601	0.029
Speaker population (z)	-0.856	0.145	-5.897	0.003

Table 9: Fixed effects from the mixed-effects model predicting model performance (analysis in the main text).

Predictor	Estimate	Std. Error	z value	p value
Intercept	-0.087	0.063	-1.39	0.164
SuperBPE	0.168	0.209	0.80	0.421
Unigram	0.034	0.026	1.29	0.196
Vocabulary size (z)	-0.044	0.030	-1.47	0.142
TTR (z)	-0.344	0.176	-1.96	0.050
Markedness (z)	-0.061	0.030	-2.05	0.040
Word entropy (z)	-0.188	0.073	-2.58	0.010
Speaker population (z)	-0.385	0.130	-2.96	0.003
Distance to English (z)	-0.090	0.044	-2.03	0.042

Table 10: Fixed effects from the regression model (item-level analysis).

92A, 97A, 116A, 143A, 143E, 143F, 143G, 144A, 26A, 51A, 85A, 86A, 95A, and 91A.

The results of each linguistic feature are reported in Tab. 8

E.3 CTC & Tokenizer

See Fig. 15.

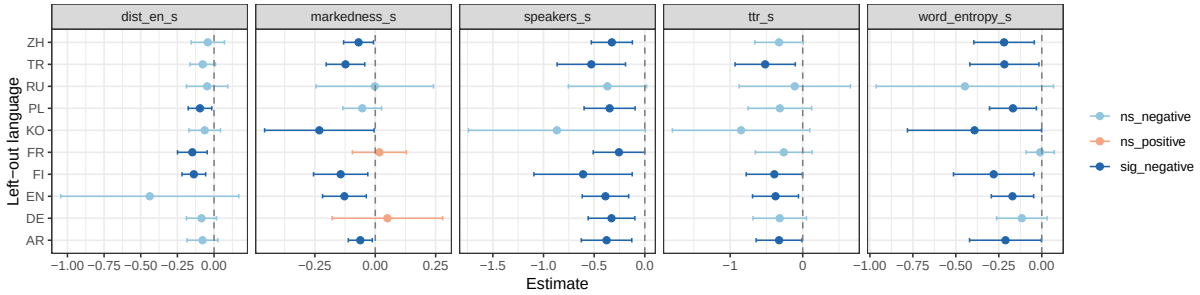


Figure 18: Leave-on-language-out experiment results for item-level analysis

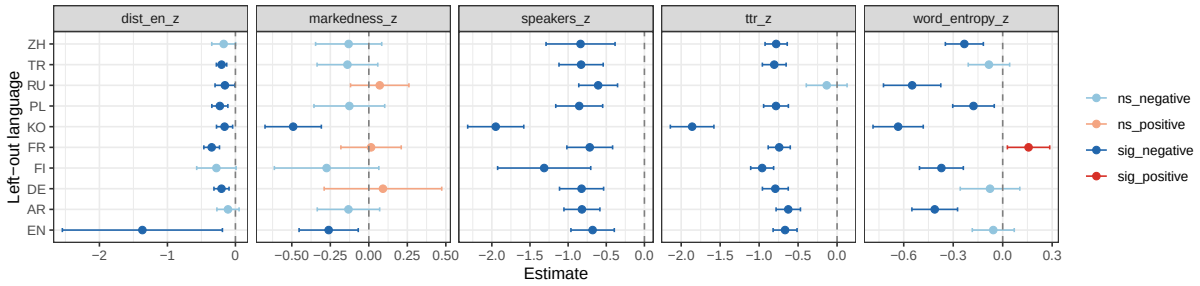


Figure 19: Leave-one-language-out experiment results for different linguistic predictors

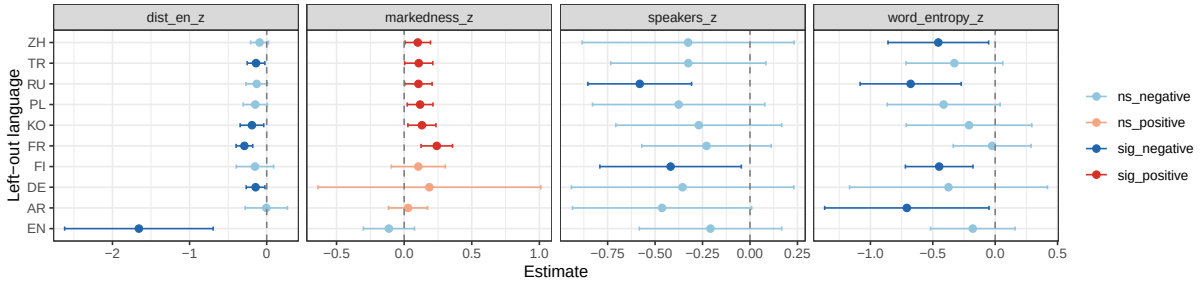


Figure 20: Leave-one-preditor-out (no TTR)

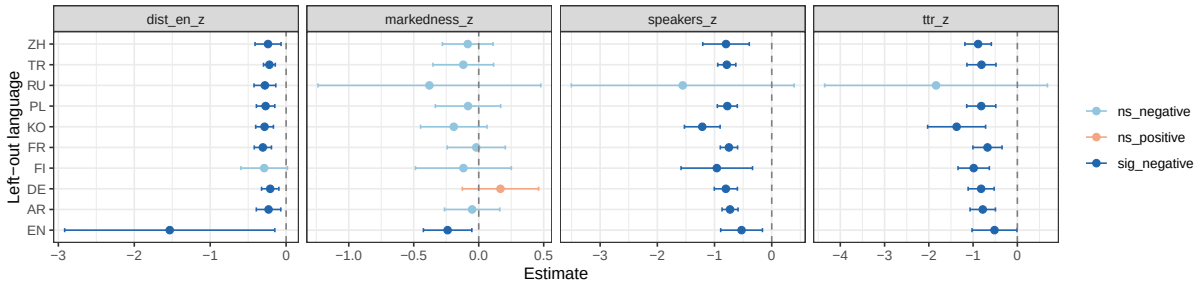


Figure 21: Leave-one-predictor-out (no Word Entropy)

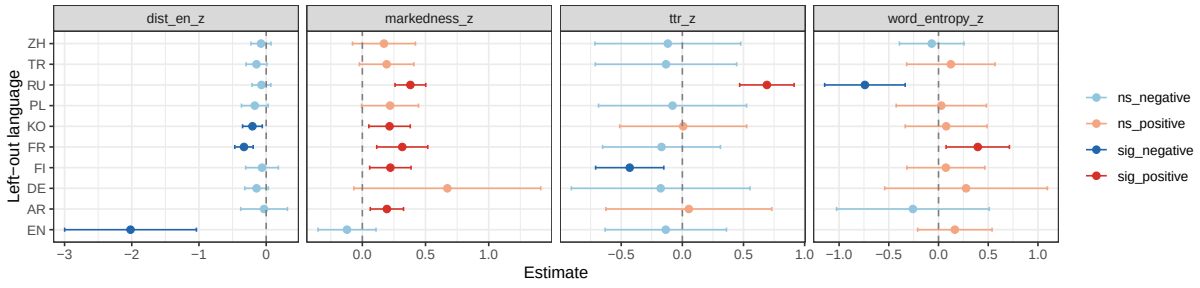


Figure 22: Leave-one-predictor-out (no speaker population)

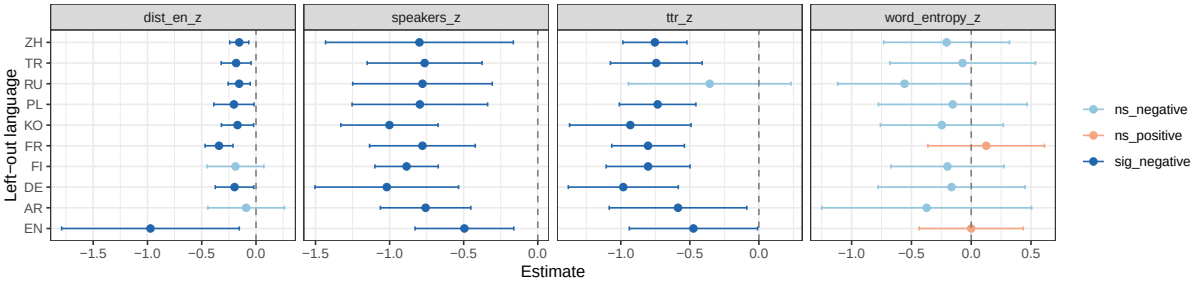


Figure 23: Leave-one-predictor-out (no markedness)

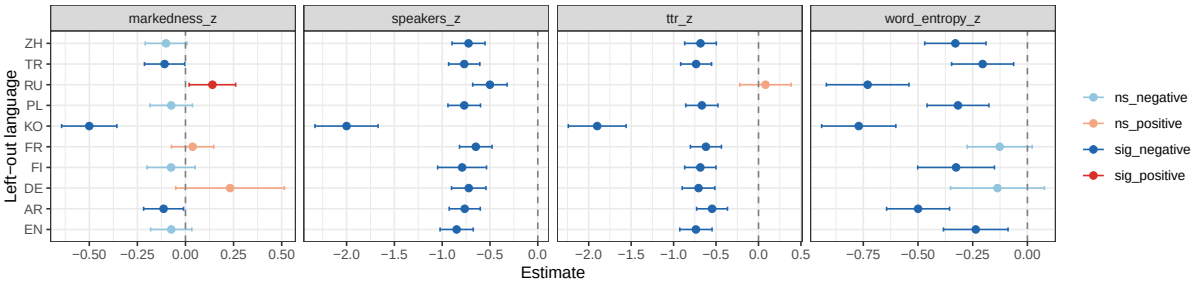


Figure 24: Leave-one-preditor-out (no typological distance to EN)