# FAIRNESS IN REPRESENTATION FOR MULTILINGUAL NLP: INSIGHTS FROM CONTROLLED EXPERIMENTS ON CONDITIONAL LANGUAGE MODELING

**Ada Wan**
University of Zurich
`ada.wan@uzh.ch`

## ABSTRACT

We perform systematically and fairly controlled experiments with the 6-layer Transformer to investigate the hardness in conditional-language-modeling languages which have been traditionally considered morphologically rich (AR and RU) and poor (ZH). We evaluate through statistical comparisons across 30 possible language directions from the 6 languages of the United Nations Parallel Corpus across 5 data sizes on 3 representation levels — character, byte, and word. Results show that performance is relative to the representation granularity of each of the languages, not to the language as a whole. On the character and byte levels, we are able to eliminate statistically significant performance disparity, hence demonstrating that a language cannot be intrinsically hard. The disparity that mirrors the morphological complexity hierarchy is shown to be a byproduct of word segmentation. Evidence from data statistics, along with the fact that word segmentation is qualitatively indeterminate, renders a decades-long debate on morphological complexity (unless it is being intentionally modeled in a word-based, meaning-driven context) irrelevant in the context of computing. The intent of our work is to help effect more objectivity and adequacy in evaluation as well as fairness and inclusivity in experimental setup in the area of language and computing so to uphold diversity in Machine Learning and Artificial Intelligence research. Multilinguality is real and relevant in computing not due to canonical, structural linguistic concepts such as morphology or "words" in our minds, but rather standards related to internationalization and localization, such as character encoding — something which has thus far been sorely overlooked in our discourse and curricula.

## 1 INTRODUCTION

### 1.1 BACKGROUND AND MOTIVATION

Most current work on fairness in Machine Learning (ML) and Natural Language Processing (NLP) focuses on the societal biases encoded in natural language data that are propagated and amplified when they are used at scale for/as Artificial Intelligence (AI) solutions[1]. But little has been said or questioned about the bias, as in, the favoring of certain outcomes, implicit in our theoretical/scientific assumptions that results in the varying performance of different languages in computing.

**Disparity in machine translation results** For instance, results reported in Junczys-Dowmunt et al. (2016) for Phrase-Based Statistical Machine Translation (PBSMT) (Koehn et al., 2003) and Neural MT (Bahdanau et al., 2014) on the 6 official languages[2] of the United Nations (UN) Parallel Corpus (Ziemski et al., 2016) indicate a disparity between EN/ES/FR and AR/RU/ZH in BLEU (Papineni et al., 2002) — translation performance in the latter group is generally worse, regardless of the MT algorithm used. AR and RU are traditionally considered morphologically complex (see e.g. Minkov et al. (2007), Seddah et al. (2010) and proceedings of related workshops in subsequent years for *Statistical Parsing of Morphologically Rich Languages*), and ZH morphologically frugal (for its lacking determiners and plural or tense markers) (Koehn, 2005). While Koehn (2005) found translating into EN to be easier than into morphologically rich languages based on BLEU scores of word tokens from PBSMT systems of 110 language directions from the 11 Europarl languages then, Bugliarello et al. (2020) found it is easier to translate out of EN than into it based on 21 Europarl languages in BPEs (Byte Pair Encodings) (Sennrich et al., 2016) with the Transformer (Vaswani et al., 2017) in a new metric, cross-mutual information.

---

[1] see e.g. work from conference (`https://facctconference.org`) and workshops in previous years on "Fairness, Accountability, and Transparency" (FAccT)

[2] Arabic (AR), English (EN), Spanish (ES), French (FR), Russian (RU), and Chinese (ZH)

**Disparity in language modeling results** Disparate performances across different languages seem to have been implicitly accepted in that it is often believed that some languages are harder to model than others. Bender (2009) advocated the relevance of linguistic typology for the design of language-universal NLP systems due to differences based on crosslinguistic structural notions, such as parts of speech and morphological complexity. Cotterell et al. (2018) studied (monolingual) language models (LMs) on the 21 Europarl languages using a word-level 7-gram standard Kneser & Ney (1995) model and LSTM-LMs (Sundermeyer et al., 2012) with characters and lemmatized forms in information-theoretic terms, and found morphological complexity to be the primary culprit for the differences in performance. Mielke et al. (2019) supplemented with the multilingual Bible corpus (Mayer & Cysouw, 2014) to extend the coverage to 69 languages, tested on RNN-LMs (an implementation of LSTM (Hochreiter & Schmidhuber, 1997)) with characters and BPEs, but concluded that basic data statistics in vocabulary size ($|V|$) and sequence length were the most predictive performance features.

We noticed, however, a discrepancy in the results from Mielke et al. (2019) for ZH — it came out as the least difficult for the character model, but it is the 6th most difficult language for the BPE model. As different input representations have been tested with different architectures with divergent results in different metrics in previous studies, each of them only testing with one data size, we decided to investigate the matter more systematically once again with statistical comparisons of score distributions between languages.

### 1.2 RESEARCH QUESTIONS AND CONTRIBUTIONS

**Research questions** Are there any statistically significant differences in hardness when it comes to Conditional-Language-Modeling (CLMing) languages which have been traditionally considered morphologically rich (AR and RU) and poor (ZH) with the 6-layer Transformer? Is morphological complexity inherent in language? When is the notion of morphological complexity relevant in computing?

**Summary of findings and insights** Based on our bilingual CLMing setup with the UN Parallel Corpus in the data size range from $10^2$ to $10^6$ lines on the character, byte, and word levels, we find:

1. Language has more finer-grained dimensions with different representations and learning patterns. Hardness in modeling is relative to its representation granularity (*representation relativity*).
2. There is neutralization of source language instances, i.e. there are no statistically significant differences between source language pairs. Only pairs of target languages differ significantly.
3. On the character and byte levels, hardness is correlated with statistical properties concerning sequence length and $|V|$ of a language, regardless of its morphological profile. As it is possible to eliminate performance disparity by decomposing sequences into finer-grained units in characters and bytes, we show that morphological complexity is not an intrinsic property of language. Unless word-based methods are used, or unless we implement/model it explicitly, the notion of morphological complexity is irrelevant in computing.
4. On the word level, hardness is correlated with $|V|$, and a complexity hierarchy arises through the manual preprocessing step of word tokenization. This complexity/disparity effected by word segmentation can be improved by subword tokenization but cannot be eliminated due to the fundamental qualitative differences in the definition of a "word" being one that neither holds universally nor is suitable/consistent for fair crosslinguistic comparisons.
5. Representational units of finer granularity can help close the gap in performance disparity.

Orthogonal to our main research questions, we also observed 2 types of sample-wise non-monotonicity — Double Descent (Belkin et al., 2019; Nakkiran et al., 2020) and *erraticity*. For reasons due to length and scope for this paper, we will defer discussions and analyses of these beyond what is addressed in § 5 to future work.

**Outline of the paper** In § 2, we define our method and experimental setup. We present our results and analysis on the primary representations in § 3 and those from secondary set of controls in § 4 in a progressive manner to ease understanding. Meta analysis on performance disparity and other discussions are in § 5.

## 2 METHOD AND DEFINITIONS

**Conditional language modeling** CLMing is the modeling of the probability of the next token, given the history of the preceding tokens and conditioning context. In our case, such conditioning context is a line from the source language. To explicitly focus on modeling the complexities that may or may not be *intrinsic* to the languages, we study the more fundamental process of CLMing without performing any translation. This allows us to eliminate confounds associated with generation and other evaluation metrics. One could think of our setup as estimating conditional probabilities with the Transformer, with a bilingual (one-to-one) setup where the perplexity (PP) of one target language ($l_{\text{trg}}$) is estimated given the parallel data in one source language ($l_{\text{src}}$), where $l_{\text{src}} \neq l_{\text{trg}}$. We focus on the very basics and examine the first step in our pipeline — input representations, holding everything else constant. Instead of measuring absolute cross-entropy scores at one data size, we evaluate the relative differences between development (dev) set score distributions between languages.

**Controlled experiments as basic research for scientific understanding of language data** Using the UN Parallel Corpus, the data from which the MT results in Junczys-Dowmunt et al. (2016) stem, we perform a series of controlled experiments with the Transformer, holding the hyperparameter settings for all 30 one-to-one language directions from the 6 languages constant. We control for size (from $10^2$ to $10^6$ lines) and language with respect to representational granularity. We examine 3 primary representation types/levels — character, byte (UTF-8), and word, and upon encountering some unusual phenomena, we perform a secondary set of controls with 5 alternate representations — on the character level: Pinyin and Wubi (ASCII representations for ZH phones and character strokes, respectively), on the byte level: code page 1256 (for AR) and code page 1251 (for RU), and on the word level: BPE. These symbolic variants allow us to manipulate the statistical properties of the representations, while staying as "faithful" to the language as possible. We adopt this symbolic data-centric[3] approach because we would like to more directly interpret the confounds, if any, that make language data different from other data types. We operate on a smaller data size range as this is more common in traditional language sciences and one of our higher goals is to bridge an understanding between language sciences and engineering (the latter being the dominant focus in NLP), and between traditional symbolic sciences and ML. We run statistical tests to identify the strongest correlates of performance and to assess whether the differences between the mean performance of different groups are indeed significant. We are concerned *not* with the absolute scores, but with the *differences* between score distributions from different languages.

**Fair evaluation with multitexts** Multitexts are multiway parallel corpora. The UN Parallel Corpus is a 6-way parallel corpus consisting of manually translated UN documents from the 25-year period between 1990 and 2014. We use the UN Parallel Corpus because it contains languages conventionally regarded as morphologically rich and poor, has quality and size sufficient for evaluation, and more importantly, it comes as raw texts (untokenized), unlike both of the corpora that Mielke et al. (2019) used. Detokenization (esp. the evaluation thereof) is not a trivial task.

**Fair information-theoretic evaluation metric** Most sequence-to-sequence models are optimized using a cross-entropy loss, defined as:

$$H(\boldsymbol{t}, \boldsymbol{s}) = -\sum_{i=1}^{N} \log_2 p(t_i \mid \boldsymbol{t}_{<i}, \boldsymbol{s}) \tag{1}$$

where $\boldsymbol{t}$ is the sequence of tokens to be predicted, $t_i$ refers to the $i^{th}$ token in that sequence, $\boldsymbol{s}$ is the sequence of tokens conditioned on, and $N = |\boldsymbol{t}|$. It is customary to report scores as PP, which is $2^{\frac{1}{N}H(\boldsymbol{t},\boldsymbol{s})}$, i.e. 2 to the power of the cross-entropy averaged by the number of tokens in the dev

---

[3] Two testing/evaluation approaches — data-centric: hold the algorithm constant and tweak data, vs. algorithm-centric: hold data constant and tweak the algorithm.

data. Cotterell et al. (2018) proposed to use “renormalized” PP to evaluate LMs tokenwise fairly by dividing the overall bits per utterance/sequence by one constant token count in any one arbitrary language (e.g. so to arrive at “bits per character” in one language to evaluate all languages). But we find that it is not necessary to assign a perspective that is centered on any one particular language, when we can evaluate simply by the total number of bits for a larger portion of texts/sequences. This can be a fairer, more general and flexible way of evaluating data that has not been or cannot be perfectly segmented or aligned line by line. We hence used instead *unnormalized* PP, i.e. the total number of bits needed to encode the dev set (3,077 lines per language, after length filtering, in our case). As the implementation we used only reports PP, we transformed it back to entropy as defined above via $H(\boldsymbol{t}, \boldsymbol{s}) = \log_2 PP(\boldsymbol{t}|\boldsymbol{s}) \times N$.

**Disparity/Inequality** In the context of our CLMing experiments, we consider there to be “disparity” or “inequality” between languages $l_1$ and $l_2$ if there are significant differences between the performance distributions of these two languages with respect to each representation. Here, by performance we mean the number of bits required to encode the held-out data using a trained CLM. With 30 directions, there are 15 pairs of source languages ($l_{\text{src}1}, l_{\text{src}2}$) and 15 pairs of target languages ($l_{\text{trg}1}, l_{\text{trg}2}$) possible. We compare the source languages among each other, and the target languages among each other. Each $l_{\text{src}}$ or each $l_{\text{trg}}$ consists of scores from all models trained across various sizes and directions. To assess whether the differences are significant, we perform unpaired two-sided significance tests with the null hypothesis that the score distributions for the two languages are not different. Upon testing for normality with the Shapiro-Wilk test (Shapiro & Wilk, 1965; Royston, 1995), we use the parametric unpaired two-sample Welch’s t-test (Welch, 1947) (when normal) or the non-parametric unpaired Wilcoxon test (Wilcoxon, 1945) (when not normal) for the comparisons. We use the implementation in R (R Core Team, 2014) for these 3 tests. To account for the multiple comparisons we are performing, we correct all p-values using Bonferroni’s correction (Benjamini & Heller, 2008; Dror et al., 2017) and follow Holm’s procedure[4] (Holm, 1979; Dror et al., 2017) to identify the pairs of $l_1$ and $l_2$ with significant differences after correction. We report all 3 levels of significance ($\alpha \leq 0.05, 0.01, 0.001$) for a more comprehensive overview. In contrast to Dror et al. (2017), which aimed to compare the performance of different algorithms, we compare languages (in the context of computing).

**Experimental setup** The systematic, identical treatment we give to our data is described as follows with further preprocessing and hyperparameter details in Appendices A and B, respectively.

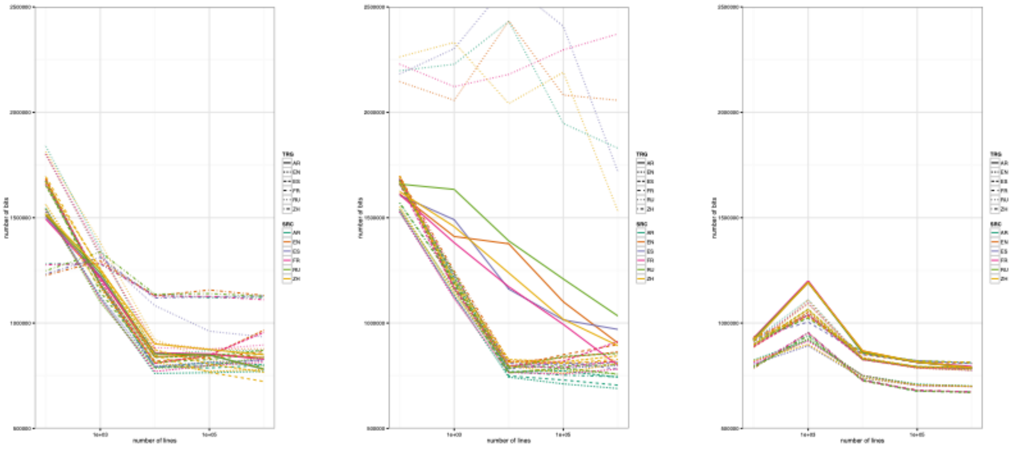
After filtering length to 300 characters maximum per line in parallel for the 6 languages, we made 3 subsets of the data with 1 million lines each — one having lines in the order of the original corpus (dataset A) and two other randomly sampled (without replacement) from the full corpus (datasets B & C). Lines in all datasets are extracted in parallel and remain fully aligned for the 6 languages. For each run and each representation, there are 30 pairwise directions (i.e. one $l_{\text{src}}$ to one $l_{\text{trg}}$) that result from the 6 languages. We trained all 150 (for 5 sizes) 6-layer Transformer models for each run using the SOCKEYE Toolkit (Hieber et al., 2018). We optimize using PP and use early stopping if no PP improvement occurs after 3 checkpoints up to 50 epochs maximum, taking the best checkpoint. Characters and bytes are supposed to mitigate the out-of-vocabulary (OOV) problem on the word level. In order to assess the effect of modeling with finer granularity more precisely, all vocabulary items appearing once in the train set are accounted for (i.e. full vocabulary on train, as in Gerz et al. (2018a;b)). But we allow our system to categorize all unknown items in the dev set to be unknown (UNK) so to measure OOVs (open vocabulary on dev (Jurafsky & Martin, 2009)). To identify correlates of performance, we perform Spearman’s correlation (Spearman, 1904) with some basic statistical properties of the data (e.g. length, $|V|$, type-token-ratio, OOV rate) as metrics — a complete list is provided in App. C. See App. D for sample construction for statistical comparisons.

## 3 EXPERIMENTAL RESULTS OF PRIMARY REPRESENTATIONS

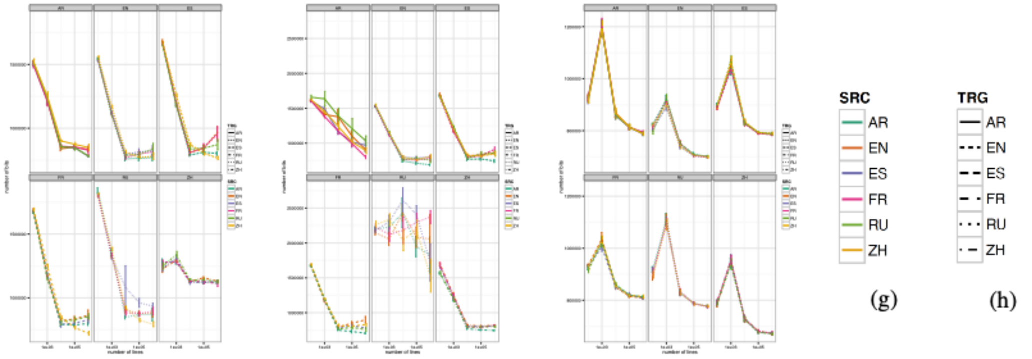
Subfigures 1a, 1b, and 1c present the mean results across 12 runs of the 3 primary representations — character, byte, and word, respectively. The x-axis represents data size in number of lines and y-axis the total conditional cross-entropy, measured in bits (Eq. 1). Each line connects 5 data points corresponding to the number of bits the CLMs (trained with training data of $10^2$, $10^3$, $10^4$, $10^5$, and

---

[4] using implementation from `https://github.com/rtmdrr/replicability-analysis-NLP`



(a) CHAR (b) BYTE (c) WORD



(g) (h)

(d) CHAR by target (e) BYTE by target (f) WORD by target

Figure 1: Number of bits (the lower the better) across data size from $10^2$ to $10^6$ lines plotted for all 30 directions. Subfigures 1a, 1b, and 1c show mean scores across 12 runs. Subfigures 1d, 1e, and 1f depict the corresponding information respectively sorted in 6 facets by target language and with error bars. Legend in Subfigure 1g shows the correspondence between colors and source languages, in Subfigure 1h between line types and target languages. (These figures are also shown enlarged in Appendix F. Please note that results pertinent to our first research question in this paper are summarized in Table 1, figures are a visual aid only. We are not concerned with the absolute scores but the distances between scores (i.e. spaces between the set of lines by $l_{\text{trg}}$). The point here is to show the differences in Transformer's overall learning pattern relative to the representational granularity. Language has more finer-grained dimensions with different representations and learning patterns. Hardness in modeling is relative to its representation granularity (representation relativity).)

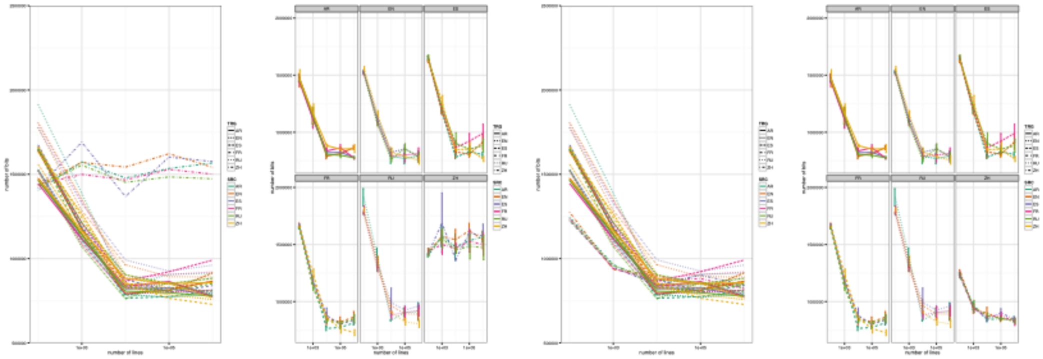
$10^6$ lines) needed to encode the target language dev set given the corresponding text in the source language. These are the same data in the same 30 language directions and 5 sizes with the same training regime, just preprocessed/segmented differently. This confirms **representation relativity** — hardness in modeling is relative to its representation granularity. Languages (or any objects being modeled) need to be evaluated relative to their representation. "One size does not fit all" (Durrani et al., 2019). Our conventional way of referring to "language" (as a socio-cultural product or with traditional word-based approaches, or even for most multilingual tasks and competitions) is too coarse-grained for computing (see also Fisch et al. (2019) and Ponti et al. (2020)).

Subfigures 1d, 1e, and 1f display the corresponding information sorted into facets by target language, source languages represented as line types. Through these we see more clearly that results can be grouped rather neatly by target language — as implicit in the Transformer's architecture, the decoder is unaware of the source language in the encoder. As shown in Table 1 in § 5 summarizing the number of source and target language pairs with significant differences, there are **no significant differences across any source language pairs**. The Transformer neutralizes source language instances. This could explain why transfer learning or multilingual/zero-shot translation (Johnson et al., 2017) is possible at all on a conceptual level.

In general, for character and byte models, most language directions do seem to converge at $10^4$ lines to similar values across all target languages, with few notable exceptions. There are some fluctuations past $10^4$, indicating further tuning of hyperparameters would be beneficial due to our present setting possibly working most favorably at $10^4$. On the character level, target language ZH (ZH$_{\text{trg}}$) shows a different learning pattern throughout. And on the byte level, AR$_{\text{trg}}$ and RU$_{\text{trg}}$ display highly unstable behavior, which we refer to as *erratic*. Word models exhibit Double Descent across the board (note the spike at $10^3$), but overall, difficult/easy languages stay consistent, with AR and RU being the hardest, followed by ES and FR, then EN and ZH. A practical takeaway from this set of experiments: in order to obtain more robust training results, use bytes for ZH (as suggested in Li et al. (2019a)) and characters for AR and RU (e.g. Lee et al. (2017)) — also if one wanted to avoid any "class" problems in performance disparity with words. Performance disparity for these representations is reported in Table 1 under "CHAR", "BYTE", and "WORD". Do note, however, that the intrinsic performance of ZH with word segmentation is not particularly subpar. But this often does not correlate with its poorer downstream tasks results (recall results from Junczys-Dowmunt et al. (2016)). Since the notion of word in ZH is highly contested and ambiguous — i) it is often aimed to align with that in other languages so to accommodate academic theories and manual feature engineering[5], ii) there is great variation among different conventions, and iii) native ZH speakers identify characters as words — there are reasons to rethink this procedure now that fairer and language-independent processing in finer granularity is possible. Li et al. (2019b) questioned the necessity of CWS in Deep Learning (DL)-based ZH NLP and presented evidence in favor of character-based processing, including results from downstream NLP tasks. In Linguistics, Duanmu (2017) presented a summary on the contested nature of wordhood in (Mandarin) ZH in relation to EN. A more native account of ZH, however, despite a couple of dialects/varieties of it being considered a high-resource language, has not yet been fully recognized and accepted in NLP.

## 4 UNDERSTANDING THE PHENOMENA WITH ALTERNATE REPRESENTATIONS

To understand why some languages show different results than others, we carried out a secondary set of controlled experiments with representations targeting the problematic statistical properties of the corresponding target languages.

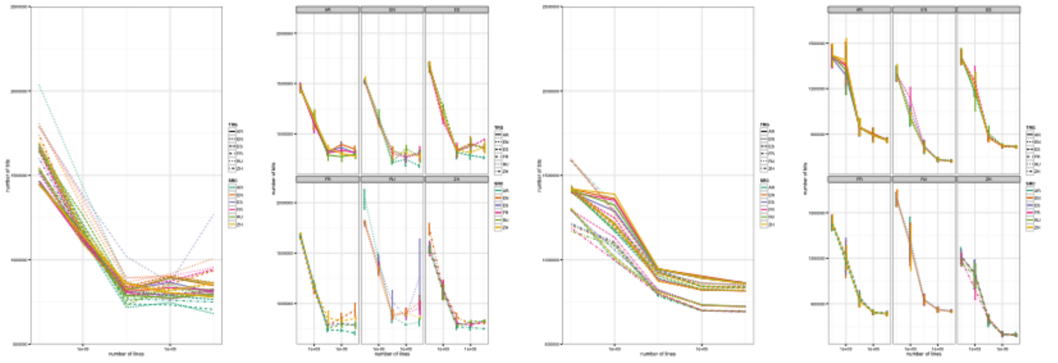


(a) Wubi (b) Wubi by target (c) Pinyin (d) Pinyin by target

Figure 2: Character-level remedies for ZH: Wubi vs. Pinyin.

---

[5] It is a "legacy interpretation" stemmed from a practical compromise from the early days in ZH NLP when the goal was to align with EN words for MT. Chinese word segmentation (CWS) has been a decades-long issue in text processing. But even in EN, for computing, the variability in "word" counts (from the trivial convention of whitespace tokenization) results in different bit counts, affecting file sizes. In NLP, such method of "word" counting brings about a high $|V|$, hence different tokenization schemes have been designed to mitigate this problem. For humans, there is no consensus about the definition of "words". Even for an academic account external of CWS, it is held to be indeterminate (see Haspelmath (2011) and references therein from the past century). Kilgarriff (1997; 2014) pointed out that "words" and "word senses" and the number thereof, in terms of lexical entries for dictionaries, are contextual and arbitrary.



(a) Code page 1256 & 1251 (b) Code page by target (c) BPE (d) BPE by target

Figure 3: Byte-level (Subfigures 3a & 3b) remedies with code page 1256 for target AR and 1251 for target RU, and word-level (Subfigures 3c & 3d) remedy with BPE for all languages.

**Character level** We reduced the high $|V|$ in ZH with representations in ASCII characters — Pinyin and Wubi. We replaced the ZH data in these formats *only on the target side* and reran the experiments involving $\text{ZH}_{\text{trg}}$ on the character level. Results in Figure 2 and Table 1 show that the elimination of disparity on character level is possible if ZH is represented through Pinyin (transliteration), as in Subfigure 2c, though at the cost of the native script information. Models with ZH logographic scripts display a behaviorial tendency unlike those with other (phonetic) alphabetic scripts[6] (Subfigure 2a), suggesting that this script pattern decomposes differently.

**Byte level** Length is the most salient statistical attribute that makes AR and RU outliers. To shorten their sequence lengths, we tested with alternate encodings on $\text{AR}_{\text{trg}}$ and $\text{RU}_{\text{trg}}$ — code page 1256 and 1251, which provide 1-byte encodings specific to AR and RU, respectively. Results are shown in Subfigures 3a and 3b. Not only is erraticity resolved, the number of 15 possible target language pairs with significant differences reduces from 8 with the UTF-8 byte representation to **0** (Table 1 under "$\text{ARRU}_{\text{t}}$"), indicating that we eliminated disparity with this optimization heuristic. Since our heuristic is a lossless and reversible transform, it shows that **a complexity that is intrinsic and necessary in language[7] does not exist** in computing, however diverse they may be, as our 6 are, from the conventional linguistic typological, phylogenetic, historical, or geographical perspectives.

**Word level** The main difference between word and character/byte models is length not being a top contributing factor correlating with performance, but instead $|V|$ is. This is understandable as word segmentation neutralizes sequence lengths. To remedy the OOV problem, we use BPE, which learns a fixed vocabulary of variable-length character sequences (on word level, as it presupposes word segmentation) from the training data. It is more fine-grained than word segmentation and is known to better model subword units for morphologically complex languages. We use the same vocabulary of 30,000 as specified in Junczys-Dowmunt et al. (2016). This reduced our averaged OOV token rate by 89-100% across the 5 sizes. The number of language pairs with significant differences reduced to 7 from 8 for word models. While BPEs are still not as effective as our character/byte variants, their results show how **more fine-grained modeling contributes positively to closing the disparity gap**.

## 5 META RESULTS, ANALYSES, AND DISCUSSION

**Performance disparity** Table 1 lists the number of language pairs with significant differences under the representations studied. Since it is **possible** for our character and byte models to effect no performance disparity for the same languages on the same data, a complexity intrinsic to language does not exist. In fact, the customary expectation that languages ought to perform differently is created through our word segmentation practice. Furthermore, the order of AR/RU > ES/FR > EN/ZH

---

[6] which are sequences with an implicit/explicit pattern made up of consonants and vowels

[7] aside from its statistical properties related to length and vocabulary. To show something is not necessarily true, only 1 counter observation is needed.

Table 1: **Disparity Table** Number of language pairs out of 15 with significant differences, with respective p-values. $\text{ARRU}_\text{t}$ refers to AR & RU being optimized only on the target side; whereas $\text{ARRU}_\text{s,t}$ denotes optimization on both source and target sides (relevant for directions AR-RU and RU-AR).

| | | CHAR | | Pinyin | | Wubi | | BYTE | | $\text{ARRU}_\text{t}$ | | $\text{ARRU}_\text{s,t}$ | | WORD | | BPE | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | p-value | src | trg | src | trg | src | trg | src | trg | src | trg | src | trg | src | trg | src | trg |
| | 0.05 | 0 | 7 | 0 | 4 | 0 | 8 | 0 | 9 | 0 | 4 | 0 | 4 | 0 | 11 | 0 | 10 |
| | 0.01 | 0 | 5 | 0 | 2 | 0 | 6 | 0 | 8 | 0 | 3 | 0 | 4 | 0 | 8 | 0 | 8 |
| ☞ | 0.001 | 0 | 3 | **0** | **0** | 0 | 5 | 0 | 8 | **0** | **0** | 0 | 2 | 0 | 8 | 0 | 7 |

Table 2: Target language pairs with significant differences indicate that the 2 languages are *not* equally/similarly good or equally/similarly bad. 15 (non-directional) language pairs total possible from 30 language directions, p=0.001.

| $\text{LANG}_\text{trg}$ PAIR | CHAR | Pinyin | Wubi | BYTE | $\text{ARRU}_\text{t}$ | $\text{ARRU}_\text{s,t}$ | WORD | BPE |
|---|---|---|---|---|---|---|---|---|
| AR-EN | | | | X | | | X | X |
| AR-ES | | | | | | | | |
| EN-ES | | | | | | | X | |
| AR-FR | | | | X | | | | |
| EN-FR | | | | | | | X | X |
| ES-FR | | | | | | | | |
| AR-RU | | | | X | | | | |
| EN-RU | | | | X | | X | X | X |
| ES-RU | | | | X | | | | |
| FR-RU | | | | X | | | | |
| AR-ZH | X | | X | X | | | X | X |
| EN-ZH | X | | X | | | | | |
| ES-ZH | | | X | | | | X | X |
| FR-ZH | X | | X | | | | X | X |
| RU-ZH | | | X | X | | X | X | X |

(Figure 1c) resembles the idea of morphological complexity. Considering there are character-internal meaningful units in languages with logographic script such as ZH (cf. Zhang & Komachi (2018)) that are rarely captured, studied, or referred to as "morphemes", this goes to show that linguistic morphology, along with its complexity, as it is practiced today[8] and that which has occurred in the NLP discourse thus far, has only been relevant on the "word" level, conceptually constrained by unstandardizable units such as "words" (and "sentences"). The definition of word, however, has been recognized as problematic for a very long time in the language sciences (cf. Footnote 5).

While the lack of significant differences between pairs of source languages would signify neutralization of source language instances, it does not mean that source languages have no effect on target. For our byte solutions with code pages, we experimented also with source side optimization in the directions that involve AR/RU as source. This affected the distribution of the disparity results for that representation — with 2 pairs being significantly different (see Table 1 under "$\text{ARRU}_\text{s,t}$"). We defer further investigation on the nature of source language neutralization to future work.

Target language pairs with significant differences are summarized in Table 2. We show that morphological complexity can be empirically eliminated in this one-setting-for-all configuration with a 6-layer network, no hyperparameter tuning, and a maximum line length of 300 characters (and its corresponding equivalence in other representations) as constrained by our hardware and compute time listed in App. A and current data availability. A more analytical solution can be obtained through data statistics (see App. E). A conceptual solution lies in the definition of "words" and morphology.

**Sample-wise Double Descent (DD)** Sample-wise non-monotonicity/DD (Nakkiran et al., 2020) denotes a degradation followed by an improvement in performance with increasing data size. We notice word models and character models with $\text{ZH}_{\text{trg}}$, i.e. models with high target $|V|$, are prone to exhibit a spike at $10^3$. A common pattern for these is the ratio of target training token count to number of parameters falls into O($10^{-4}$) for $10^2$ lines, O($10^{-3}$) at $10^3$, O($10^{-2}$) at $10^4$, and O($10^{-1}$) for $10^5$ lines and so on. But for more atomic units such as alphabetic (not logographic) characters (may it be Latin, Cyrillic, or Abjad) and for bytes, this progression instead begins at O($10^{-3}$) at $10^2$ lines. Instead of thinking this spike of $10^3$ as irregular, we may instead want to think of this learning curve as shifted by 1 order of magnitude to the right for characters and bytes and/or the performance at $10^2$ lines for words and ZH-characters due to being overparameterized and hence abnormal. This would also fit in with the findings by Belkin et al. (2019) and Nakkiran et al. (2020) attributing DD to overparameterization. While almost all work attribute DD to algorithmic reasons, findings from Chen et al. (2020) corroborate our observation and confirms that DD arises due to "the interaction between the properties of the data and the inductive biases of learning algorithms". Other related work on the DD phenomenon and its development can also be found in their work.

---

[8] But there are no reasons why we cannot adopt a statistical science of language in finer granularities beyond/without "words", with standardized units (characters/bytes) and/or continuous representations. Resources, e.g. quality parallel data or contrast sets, can serve both data science and ML interpretation and evaluation well.

**Erraticity** We observe another type of sample-wise non-monotonicity, one that signals irregular and unstable performance across data sizes and runs. Within one run, erraticity can be observed directly as changes in direction on the y-axis. Across runs, large variance can be observed, even with the same dataset. Erraticity can also be observed indirectly through a negative correlation between data size and performance. Many work on length bias in NMT have focused on solutions related to search, e.g. Murray & Chiang (2018). Our experiments show that a kind of length bias can surface already with CLMing, without generation taking place.

**Additional related work** That basic data statistics are the driver of success in performance in multilingual modeling has so far only been explicitly argued for in Mielke et al. (2019). We go beyond their work in monolingual LMs to study CLMs and evaluate also in relation to data size, representational granularity, and quantitative and qualitative fairness. To the best of our knowledge, there has been no prior work on demonstrating the neutralization of source language instances through statistical comparisons, a numerical analysis on DD for sequence-to-sequence models, the meta phenomenon of a sample-wise non-monotonicity (erraticity) being related to length.

## 6 CONCLUSION

**Summary** We investigate whether the performance disparity between languages which have been traditionally considered morphologically rich (AR and RU) and poor (ZH) in the 6-layer Transformer CLM is justified due to morphological complexity and find that it is not. Performance disparity can be explained by data statistics and in the context of computing, it can be eliminated by optimization on length and $|V|$ through character/byte representations. In fact, morphological complexity is not a necessary concept in computing because "word" is not a necessary concept in computing, unless we make it so through word segmentation. A morphological complexity hierarchy can result simply through word segmentation. Furthermore, there are many possible interpretations to "words" for humans and since morphology is defined with the concept of "word", there is no stable ground for assessing this complexity. Representational units of finer granularity were shown to help eliminate performance disparity though at the cost of longer sequence length, which can have a negative impact on robustness. In addition, we found all word models and character models with $\text{ZH}_{\text{trg}}$ to behave similarly in their being prone to exhibit a peak (as sample-wise DD) around $10^3$ lines in our setting.

**Outlook** ML has enabled greater diversity in NLP (Joshi et al., 2020). Fairness, in the elimination of disparity, does not require big data. This paper made a pioneering attempt to bridge research in NNs/DL, language sciences, and language engineering through a data-centric perspective. Multilinguality is real and relevant in computing not due to canonical, structural linguistic concepts such as morphology or "words" in our minds, but rather standards related to internationalization and localization, such as character encoding — something which has thus far been sorely overlooked in our discourse and curricula. We also believe that a more fine-grained statistical data science can well complement algorithmic analyses with a view that is more empirically robust (i.e. experimentally verifiable) and more relevant to machine processing, contributing to a more generalizable and interpretable pool of knowledge for ML/NNs/DL. A more comprehensive study can lead us not only to new scientific frontiers, but also to better designs and evaluation, benefitting the development of a more general, diverse and inclusive AI.

## ETHICS STATEMENT: FAIRNESS CONCERNS FOR MULTILINGUALITY

**Clearer nomenclature** If/When the intent is *not* to explicitly model linguistic morphology in computing, one can simply describe languages and their statistical profiles with respect to their representational granularity in characters or bytes (which are and can be exhaustively standardized in computing), or refer to sequences as longer/shorter or having a higher/lower vocabulary size when comparing them with each other, rather than "richer"/"poorer" based on concepts (e.g. "words", "sentences") that can be ambiguous, contested, and inaccessible to many.

**Accessibility** Language communities who are unfamiliar with languages similar to dominant languages or those who are reluctant to conform to one structurally similar interpretation should not have to feel inadequate in processing their own language "from scratch", if they so choose to. As technologists, we can help take equitable measures to make fairer data representations and infrastructure available. A "word"-free view of language preempts linguistic/cultural hegemony and such an interpretation would also help make the analyses of language data more objective and clearer.

**Scarcity of quality and/or multiway data for science, evaluation, and culture** With the rise of multilingual models comes an alleged decrease in reliance on parallel corpora for MT. But parallel data, esp. high-quality, *multiway* and/or multi-modal parallel data, are important for scientific research and as a form of cultural and language documentation. And as this paper shows, they can also serve as evaluation data for ML models for better understanding and interpretation. As challenge sets, data for machine processing would need to be statistically diverse and challenging as well. Parallel data from previous years have often come in the form of bitexts (2-way parallel text data), usually "word"-tokenized where real length information has been compromised. (In fact, the Bible data from Mayer & Cysouw (2014) came with another confound — the ZH numeral '一' ("one") is recognized as a dash (punctuation) and hence tokenized with surrounding whitespaces.) At the time of our present study (2019-present), the UN Parallel Corpus was the only unperturbed, fully multiway parallel data sufficient for reliable evaluation for our size range. Data for science, evaluation, and culture require a long-term, stable platform and support. There are many forms of data science. But in terms of having a sustainable practice to collect and curate good data, and exercising enough of a science with that data so to improve collective intelligence and mutual understanding, instead of looking at data from an utilitarian point of view only for consumer application purposes, there seems to be room for improvement still. We hope our work could help effect a positive change in this direction.[9]

### ACKNOWLEDGMENTS

The author thanks all anonymous reviewers and PCs as well as the following individuals for their informative comments on previous version(s) of this paper (names in alphabetic order by last name): Benjamin Börschinger, Jordan Boyd-Graber, Christian Buck, Kyunghyun Cho, Kenneth Church, Miryam de Lhoneux, Rotem Dror, San Duanmu, Chris Dyer, Roman Flury, John Goldsmith, Yannic Kilcher, Sandra Kübler, Thomas McColgan, Jason Naradowsky, Ryokan Ri, Anna Rogers, Mark Rowan, Nikunj Saunshi, Rico Sennrich, Ekaterina Vylomova, Jason Weston, and Kie Zuraw.

We also thank Kevin Duh, Daniela Gerz, Kenneth Heafield, Felix Hieber, Marcin Junczys-Dowmunt, Gal Kaplun, Sabrina Mielke, Mathias Müller, Nakkiran Preetum, and Annette Rios for their kind responses in correspondence concerning related work and software. We thank Ismail Moukadiri for his help with the Arabic language.

Some initial experiments were supported by the computing resources at the Department of Informatics and Computational Linguistics at UZH and the Institute of Neuroinformatics at UZH and ETH Zurich (for which we especially thank Richard H.R. Hahnloser, Nikola I. Nikolov, Yuhuang Hu, and Pawel Pyk).

## REFERENCES

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv e-prints*, abs/1409.0473, September 2014. URL `https:`

---

[9] Please see further discussion in Wan (2022).

//arxiv.org/abs/1409.0473.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1903070116. URL `https://www.pnas.org/content/116/32/15849`.

Emily M. Bender. Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pp. 26–32, Athens, Greece, March 2009. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/W09-0106`.

Yoav Benjamini and Ruth Heller. Screening for partial conjunction hypotheses. *Biometrics*, 64(4): 1215–1222, 2008. ISSN 0006341X, 15410420. URL `http://www.jstor.org/stable/25502204`.

Emanuele Bugliarello, Sabrina J. Mielke, Antonios Anastasopoulos, Ryan Cotterell, and Naoaki Okazaki. It's easier to translate out of English than into it: Measuring neural translation difficulty by cross-mutual information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1640–1649, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.149. URL `https://www.aclweb.org/anthology/2020.acl-main.149`.

Lin Chen, Yifei Min, Mikhail Belkin, and Amin Karbasi. Multiple descent: Design your own generalization curve, 2020.

Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv preprint arXiv:1512.01274*, 2015.

Colin Cherry, George Foster, Ankur Bapna, Orhan Firat, and Wolfgang Macherey. Revisiting character-based neural machine translation with capacity and compression. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4295–4305. Association for Computational Linguistics, 2018. URL `http://aclweb.org/anthology/D18-1461`.

Ryan Cotterell, Sabrina J. Mielke, Jason Eisner, and Brian Roark. 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)*, pp. 536–541, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2085. URL `https://www.aclweb.org/anthology/N18-2085`.

Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. Replicability analysis for natural language processing: Testing significance with multiple datasets. *Transactions of the Association for Computational Linguistics*, 5:471–486, 2017. URL `http://aclweb.org/anthology/Q17-1033`.

San Duanmu. Word and wordhood, modern. In Rint Sybesma (ed.), *Encyclopedia of Chinese Language and Linguistics*, pp. 543–549. Brill, 2017.

Nadir Durrani, Fahim Dalvi, Hassan Sajjad, Yonatan Belinkov, and Preslav Nakov. One size does not fit all: Comparing NMT representations of different granularities. 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)*, pp. 1504–1516, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1154. URL `https://www.aclweb.org/anthology/N19-1154`.

Adam Fisch, Jiang Guo, and Regina Barzilay. Working hard or hardly working: Challenges of integrating typology into neural dependency parsers. 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)*, pp. 5713–5719, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1574. URL `https://www.aclweb.org/anthology/D19-1574`.

Daniela Gerz, Ivan Vulić, Edoardo Ponti, Jason Naradowsky, Roi Reichart, and Anna Korhonen. Language modeling for morphologically rich languages: Character-aware modeling for word-level prediction. *Transactions of the Association for Computational Linguistics*, 6:451–465, 2018a. doi: 10.1162/tacl_a_00032. URL `https://www.aclweb.org/anthology/Q18-1032`.

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 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*, pp. 316–327, Brussels, Belgium, October-November 2018b. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/D18-1029`.

Martin Haspelmath. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica*, 2011.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. The Sockeye neural machine translation toolkit at AMTA 2018. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pp. 200–207. Association for Machine Translation in the Americas, 2018. URL `http://aclweb.org/anthology/W18-1820`.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL `http://dx.doi.org/10.1162/neco.1997.9.8.1735`.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. ISSN 03036898, 14679469. URL `http://www.jstor.org/stable/4615733`.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017. doi: 10.1162/tacl_a_00065. URL `https://www.aclweb.org/anthology/Q17-1024`.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 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*, pp. 6282–6293, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. URL `https://www.aclweb.org/anthology/2020.acl-main.560`.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. Is neural machine translation ready for deployment? A case study on 30 translation directions. In *IWSLT 2016, Seattle*, October 2016. URL `https://www.microsoft.com/en-us/research/publication/neural-machine-translation-ready-deployment-case-study-30-\translation-directions/`.

Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Pearson Prentice Hall, second edition, 2009.

Adam Kilgarriff. "I don't believe in word senses". *CoRR*, cmp-lg/9712006, 1997. URL `http://arxiv.org/abs/cmp-lg/9712006`.

Adam Kilgarriff. "How many words are there?". 2014. URL `https://www.sketchengine.eu/wp-content/uploads/How_Many_Words_2014.pdf`.

Reinhard Kneser and Hermann Ney. Improved backing-off for m-gram language modeling. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pp. 181–184. IEEE, 1995.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pp. 79–86, Phuket, Thailand, September 13-15 2005. URL `https://aclanthology.org/2005.mtsummit-papers.11`.

Philipp Koehn, Franz J. Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 127–133, 2003. URL `https://www.aclweb.org/anthology/N03-1017`.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pp. 177–180. Association for Computational Linguistics, 2007. URL `http://aclweb.org/anthology/P07-2045`.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5: 365–378, December 2017. doi: 10.1162/tacl_a_00067. URL `https://www.aclweb.org/anthology/Q17-1026`.

Bo Li, Yu Zhang, Tara Sainath, Yonghui Wu, and William Chan. Bytes are all you need: End-to-end multilingual speech recognition and synthesis with bytes. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5621–5625. IEEE, 2019a.

Xiaoya Li, Yuxian Meng, Xiaofei Sun, Qinghong Han, Arianna Yuan, and Jiwei Li. Is word segmentation necessary for deep learning of Chinese representations? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3242–3252, Florence, Italy, July 2019b. Association for Computational Linguistics. doi: 10.18653/v1/P19-1314. URL `https://www.aclweb.org/anthology/P19-1314`.

Ken Lunde. *CJKV Information Processing*. O'Reilly Media, Inc., 2nd edition, 2008. ISBN 0596514476, 9780596514471.

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

Sabrina J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. What kind of language is hard to language-model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4975–4989, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1491. URL `https://www.aclweb.org/anthology/P19-1491`.

Einat Minkov, Kristina Toutanova, and Hisami Suzuki. Generating complex morphology for machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 128–135, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/P07-1017`.

Kenton Murray and David Chiang. Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 212–223, Belgium, Brussels, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6322. URL `https://www.aclweb.org/anthology/W18-6322`.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=B1g5sA4twr`.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for*

*Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL `https://www.aclweb.org/anthology/P02-1040`.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. Modeling language variation and universals: A survey on typological linguistics for natural language processing, 2020.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.R-project.org/`.

Patrick Royston. Remark as r94: A remark on algorithm as 181: The w-test for normality. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 44(4):547–551, 1995. ISSN 00359254, 14679876. URL `http://www.jstor.org/stable/2986146`.

Djame Seddah, Sandra Kuebler, and Reut Tsarfaty (eds.). *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically Rich Languages*, Los Angeles, CA, USA, June 2010. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/W10-1400`.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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)*, pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL `https://www.aclweb.org/anthology/P16-1162`.

S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples)†. *Biometrika*, 52(3-4):591–611, 12 1965. ISSN 0006-3444. doi: 10.1093/biomet/52.3-4.591. URL `https://doi.org/10.1093/biomet/52.3-4.591`.

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. ISSN 00029556. URL `http://www.jstor.org/stable/1412159`.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. Lstm neural networks for language modeling. In *INTERSPEECH*, 2012.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. Curran Associates, Inc., 2017. URL `http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf`.

Ada Wan. A transdisciplinary approach to scientific insights in finer granularity: ml-inspired progress and opportunities. 2022.

B. L. Welch. The Generalization of 'Student's' Problem when Several Different Population Variances are Involved. *Biometrika*, 34(1-2):28–35, 01 1947. ISSN 0006-3444. doi: 10.1093/biomet/34.1-2.28. URL `https://doi.org/10.1093/biomet/34.1-2.28`.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945. ISSN 00994987. URL `http://www.jstor.org/stable/3001968`.

Longtu Zhang and Mamoru Komachi. Neural machine translation of logographic language using sub-character level information. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 17–25, Belgium, Brussels, October 2018. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/W18-6303`.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. The United Nations Parallel Corpus v1.0. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

# APPENDICES

**A DATA SELECTION AND PREPROCESSING DETAILS** 16

**B HYPERPARAMETER SETTING** 17

**C CORRELATION STATISTICS** 18

**D STATISTICAL COMPARISONS** 19

**E DATA STATISTICS** 20

**F ENLARGED FIGURES FOR ALL 30 LANGUAGE DIRECTIONS (AGGREGATE RESULTS FROM ALL RUNS)** 24

# A DATA SELECTION AND PREPROCESSING DETAILS

The UN Parallel Corpus v1.0 (Ziemski et al., 2016) consists of manually translated UN documents from 1990 to 2014 in the 6 official UN languages. Therein is a subcorpus that is fully aligned by line, comprising the 6-way parallel corpus we use. We tried to have as little preprocessing or filtering as necessary to eliminate possible confounds. But as the initial runs of our experiment failed due to insufficient memory on a single GPU with 12 GB VRAM[10], we filtered out lines with more than 300 characters in any language in lockstep with one another for all the 6 languages such that the subcorpora would remain parallel, thereby keeping the material of each language semantically equivalent to one another. 8,944,859 lines for each language were retained as our training data which cover up to the 75th percentile in line length for all 6 languages. In order to monitor the effect of data size, we made subcorpora of each language in 5 sizes by `head`ing the first $10^2, 10^3, 10^4, 10^5, 10^6$ lines[11]. We refer to this as dataset A. In addition, to better understand and verify the consistency of the phenomena observed, we made 2 supplemental datasets by shuffling the 8,944,859 lines two different times randomly and `head`ing the number of lines in our 5 sizes for each language, again in lockstep with one another (datasets B and C).

The systematic training regime that we gave to our language directions is identical for all and we controlled also for seeds. For each of the 3 primary representations — character, byte, and word, we performed:

- 5 runs in 5 sizes ($10^2 - 10^6$ lines): A0 (seed=13), B0 (13), C0 (9948), A1 (9948), A2 (265), and
- 7 more runs in 4 sizes ($10^2 - 10^5$ lines): A3 (777), A4 (42), A5 (340589), A6 (1000), A7 (83146), B1 (9948), & C1 (13).

Figure 1 shows results from all 12 runs in all sizes for the primary representations.

For the alternate/secondary representations, we performed 3 runs each in 5 sizes ($10^2$-$10^6$ lines) (A0, B0, & C0). Due to limitations in computing resources, we were not able to perform as many runs as the primary representations. But important for our statistical comparisons is that we evaluate based on an equal number of runs and on the same data for all candidates. Tables 1 & 2 are the results.

For each run and each size, there are 30 pairwise directions (i.e. 1 source language to 1 target language, e.g. AR-EN for Arabic to English) that result from the 6 languages. We trained all 150 jobs (30 directions x 5 sizes) for each run and representation using the Transformer model (Vaswani et al., 2017) as supported by the SOCKEYE Toolkit (Hieber et al., 2018) (version 1.18.85), based on MXNet (Chen et al., 2015). A detailed description of the architecture of the Transformer can be found in (Vaswani et al., 2017). The same set of hyperparameters applies to all and its values are listed in Appendix B.

For character modeling, we used a dummy symbol to denote each whitespace. For byte, we turned each UTF-8-encoded character into a byte string in decimal value, such that each token is a number between 0 and 255, inclusive. For word, we followed (Junczys-Dowmunt et al., 2016) and used the Moses tokenizer (Koehn et al., 2007) as is standard in NMT practice when word tokenization is applied and Jieba[12] for segmentation in ZH.

Pinyin is a romanization of ZH characters based on their pronunciations and Wubi an input algorithm that decomposes character-internal information into stroke shape and ordering and matches these to 5 classes of radicals (Lunde, 2008). For Pinyin, we used the implementation from `https://github.com/lxyu/pinyin` in the numerical format such that each character/syllable is followed by a single digit indicating its lexical tone in Mandarin. For Wubi, we used the dictionary from the implementation from `https://github.com/arcsecw/wubi`.

---

[10] GPUs used for experiments in this paper range from a NVIDIA TITAN RTX (24 GB), NVIDIA GeForce RTX 2080 Ti (11 GB), a GTX Titan X (12 GB), to a GTX 1080 (8 GB). All jobs were run on a single GPU setting. Some word-level experiments involving $\text{AR}_{\text{trg}}$ or $\text{RU}_{\text{trg}}$ at $10^6$ had to be run on a CPU as 24 GB VRAM were not sufficient. Models with higher maximum sequence lengths (e.g. byte models) were trained with 24 GB VRAM. Difference in equipment does not necessarily lead to degradation/improvement in scores.

[11] The terms "line" and "sentence" have been used interchangeably in the NLP literature. We use "line" to denote a sequence that ends with a newline character and "sentence" as one with an ending punctuation. Most parallel corpora, such as ours, are aligned by line, as a line may be part of a sentence or without an ending punctuation (e.g. a header/title). Using a standardized unit such as "line" would also be a fairer measure to linguae/scriptiones continuae (languages/scripts with no explicit punctuation).

[12] `https://github.com/fxsjy/jieba`

We have implemented all representations such that they would be reversible even when the sequence contains code-mixing.

We used the official dev set as provided in (Ziemski et al., 2016), 3,077 lines per language remained from 4,000 after filtering line length to 300 characters. Data statistics is provided in Appendix E for reference.

**Notes on training time** Each run of 30 directions in 5 sizes took approximately 8-12 days for character and byte models. Byte models generally took longer — hence training time is positively correlated with length (concurring with observations by Cherry et al. (2018) as they compared character with BPE models). A maximum length of 300 characters entails a maximum length of *at least* 300 bytes in UTF-8. Each run of word models (30 directions, 5 sizes) took about 6 days (excluding the training of some 7-9 directions out of 30 per run involving $\mathrm{AR}_{\mathrm{trg}}$ or $\mathrm{RU}_{\mathrm{trg}}$ at $10^6$ on word level which took about 12-18 hours *each direction* to train on a CPU as these required more space and would run out of memory (OOM) on our GPUs otherwise). These figures do not include the additional probing experiments described in § 4.

## B HYPERPARAMETER SETTING

- encoder transformer;
- decoder transformer;
- num-layers 6:6;
- num-embed 512:512;
- transformer-model-size 512;
- transformer-attention-heads 8;
- transformer-feed-forward-num-hidden 2048;
- transformer-activation-type relu;
- transformer-positional-embedding-type fixed;
- transformer-preprocess d; transformer-postprocess drn;
- transformer-dropout-attention 0.1;
- transformer-dropout-act 0.1;
- transformer-dropout-prepost 0.1;
- batch-size 15;
- batch-type sentence;
- max-num-checkpoint-not-improved 3;
- max-num-epochs 50;
- optimizer adam;
- optimized-metric perplexity;
- optimizer-params epsilon: 0.000000001, beta1: 0.9, beta2: 0.98;
- label-smoothing 0.0;
- learning-rate-reduce-num-not-improved 4;
- learning-rate-reduce-factor 0.001;
- loss-normalization-type valid;
- max-seq-len 300 for character, word, and BPE, 672 for all bytes, 688 for Wubi, 680 for Pinyin;
- checkpoint-frequency/interval 4000.
  (For smaller datasets, the end of 50 epochs is often reached before the first checkpoint. Since SOCKEYE only outputs scores at checkpoints, we adjusted the checkpoint frequency as follows to get a score outputted by the end of 50 epochs: 1000 for 100 lines for all character & byte instances, 400 for 100 lines for word and 500 for 100 lines BPE, 3450 for 1000 lines for word & BPE. For the very few cases that this default does not suffice due to bucketing of similar length sequences, we manually set the checkpoint frequency to the last batch.)

## C CORRELATION STATISTICS

Best correlating metrics, i.e. the union of top 3 metrics for all representations.
For each representation, the **top 3 metrics** are boldfaced.
All correlations are **highly significant** ($p < 10^{-30}$), except for min source length for WORD ($p \approx 0.0001$) and min target length for WORD ($p \approx 0.3861$).

| Metric | CHAR | Pinyin | Wubi | BYTE | $\text{ARRU}_{\text{t}}$ | $\text{ARRU}_{\text{s,t}}$ | WORD | BPE |
|---|---|---|---|---|---|---|---|---|
| minimum length (target) | **0.84** | **0.85** | **0.86** | **0.60** | **0.84** | **0.84** | −0.02 | 0.65 |
| minimum length (source) | **0.82** | **0.84** | **0.85** | 0.57 | **0.84** | **0.84** | 0.10 | 0.64 |
| number of tokens (source) | −0.78 | −0.81 | −0.82 | **−0.60** | **−0.81** | **−0.81** | −0.59 | **−0.83** |
| TTR (target) | **0.83** | **0.83** | **0.84** | 0.48 | 0.81 | 0.81 | 0.61 | 0.83 |
| $\|V\|$ (source) | −0.54 | −0.51 | −0.51 | −0.50 | −0.67 | −0.68 | **−0.63** | **−0.86** |
| data size in lines | −0.80 | −0.83 | −0.83 | −0.59 | −0.81 | −0.81 | −0.62 | **−0.86** |
| OOV token rate (target) | 0.69 | 0.66 | 0.66 | 0.47 | 0.67 | 0.68 | **0.66** | 0.62 |
| OOV type rate (target) | 0.70 | 0.71 | 0.72 | 0.47 | 0.69 | 0.70 | **0.65** | 0.62 |
| TTR (source) | 0.67 | 0.71 | 0.71 | **0.60** | 0.81 | 0.81 | 0.56 | 0.82 |

The full list of metrics used for the correlation analysis is:

1. minimum length (source),
2. minimum length (target),
3. maximum length (source),
4. maximum length (target),
5. median length (source),
6. median length (target),
7. mean length (source),
8. mean length (target),
9. length std (source),
10. length std (target),
11. data size in lines,
12. number of parameters,
13. number of types ($|V|$) (source),
14. number of types ($|V|$) (target),
15. number of tokens (source),
16. number of tokens (target),
17. type-token-ratio (TTR) (source),
18. type-token-ratio (TTR) (target),
19. OOV type rate (source),
20. OOV type rate (target),
21. OOV token rate (source),
22. OOV token rate (target),
23. token ratio,
24. target type-to-parameter ratio,
25. target token-to-parameter ratio,
26. distance between the TTRs of source and target = $(1 - \text{TTR}_{src}/\text{TTR}_{trg})^2$,
27. token-to-parameter ratio (i) = (median length source * median length target * num_lines) / num_parameters,
28. token-to-parameter ratio (ii) = (num_source_tokens * num_target_tokens) / num_parameters.

# D STATISTICAL COMPARISONS

Recall the definition and method for our **Disparity/Inequality** assessment from § 2:

> In the context of our CLMing experiments, we consider there to be "disparity" or "inequality" between languages $l_1$ and $l_2$ if there are significant differences between the performance distributions of these two languages with respect to each representation. Here, by performance we mean the number of bits required to encode the held-out data using a trained CLM. With 30 directions, there are 15 pairs of source languages $(l_{\text{src}1}, l_{\text{src}2})$ and 15 pairs of target languages $(l_{\text{trg}1}, l_{\text{trg}2})$ possible. We compare the source languages among each other, and the target languages among each other. Each $l_{\text{src}}$ or each $l_{\text{trg}}$ consists of scores from all models trained across various sizes and directions. To assess whether the differences are significant, we perform unpaired two-sided significance tests with the null hypothesis that the score distributions for the two languages are not different. Upon testing for normality with the Shapiro-Wilk test (Shapiro & Wilk, 1965; Royston, 1995), we use the parametric unpaired two-sample Welch's t-test (Welch, 1947) (when normal) or the non-parametric unpaired Wilcoxon test (Wilcoxon, 1945) (when not normal) for the comparisons. We use the implementation in R (R Core Team, 2014) for these 3 tests. To account for the multiple comparisons we are performing, we correct all p-values using Bonferroni's correction (Benjamini & Heller, 2008; Dror et al., 2017) and follow Holm's procedure[13] (Holm, 1979; Dror et al., 2017) to identify the pairs of $l_1$ and $l_2$ with significant differences after correction. We report all 3 levels of significance ($\alpha \leq 0.05, 0.01, 0.001$) for a more comprehensive overview. In contrast to Dror et al. (2017), which aimed to compare the performance of different algorithms, we compare languages (in the context of computing).

**To get samples for the statistical comparison results for the Disparity Table (Table 1):**

For each representation, we used 3 runs (A0, B0, C0) in 5 sizes ($10^2$-$10^6$ lines) for each $l_{\text{src}}$ and each $l_{\text{trg}}$. There are:

6 $l_{\text{src}}$ ($\text{AR}_{\text{src}}$, $\text{EN}_{\text{src}}$, $\text{ES}_{\text{src}}$, $\text{FR}_{\text{src}}$, $\text{RU}_{\text{src}}$, $\text{ZH}_{\text{src}}$) and
6 $l_{\text{trg}}$ ($\text{AR}_{\text{trg}}$, $\text{EN}_{\text{trg}}$, $\text{ES}_{\text{trg}}$, $\text{FR}_{\text{trg}}$, $\text{RU}_{\text{trg}}$, $\text{ZH}_{\text{trg}}$).

We compare pairwise among the $l_{\text{src}}$. The 15 pairs are:

$\text{AR}_{\text{src}}$-$\text{EN}_{\text{src}}$, $\text{AR}_{\text{src}}$-$\text{ES}_{\text{src}}$, $\text{AR}_{\text{src}}$-$\text{FR}_{\text{src}}$, $\text{AR}_{\text{src}}$-$\text{RU}_{\text{src}}$, $\text{AR}_{\text{src}}$-$\text{ZH}_{\text{src}}$,
$\text{EN}_{\text{src}}$-$\text{ES}_{\text{src}}$, $\text{EN}_{\text{src}}$-$\text{FR}_{\text{src}}$, $\text{EN}_{\text{src}}$-$\text{RU}_{\text{src}}$, $\text{EN}_{\text{src}}$-$\text{ZH}_{\text{src}}$,
$\text{ES}_{\text{src}}$-$\text{FR}_{\text{src}}$, $\text{ES}_{\text{src}}$-$\text{RU}_{\text{src}}$, $\text{ES}_{\text{src}}$-$\text{ZH}_{\text{src}}$,
$\text{FR}_{\text{src}}$-$\text{RU}_{\text{src}}$, $\text{FR}_{\text{src}}$-$\text{ZH}_{\text{src}}$,
$\text{RU}_{\text{src}}$-$\text{ZH}_{\text{src}}$.

Likewise 15 pairs among the $l_{\text{trg}}$.

**For example**, for the character (primary) representation to compare between $\text{AR}_{\text{src}}$ and $\text{EN}_{\text{src}}$, we construct the sample for $\text{AR}_{\text{src}}$ ($\texttt{sample}_{\texttt{ARsrc}}$) and the sample for $\text{EN}_{\text{src}}$ ($\texttt{sample}_{\texttt{ENsrc}}$) as follows:

Out of the 30 CHAR directions, there are 5 directions involving $\text{AR}_{\text{src}}$ trained for each run and data size (i.e. the directions: AR-EN, AR-ES, AR-FR, AR-RU, AR-ZH).

For each direction, there are 15 models trained (3 runs x 5 sizes). We take all 75 CHAR models (15 models x 5 directions) involving $\text{AR}_{\text{src}}$ as $\texttt{sample}_{\texttt{ARsrc}}$. That's a sample of size 75.

Likewise, for $\text{EN}_{\text{src}}$ (5 directions: EN-AR, EN-ES, EN-FR, EN-RU, EN-ZH), we also have 75 data points for $\texttt{sample}_{\texttt{ENsrc}}$. (Likewise also for all 6 $l_{\text{src}}$ and all 6 $l_{\text{trg}}$.)

For the comparisons, we compare pairwise, i.e. with two samples each time, but with *unpaired* two-sample Welch's t-test (when normal) or the non-parametric *unpaired* Wilcoxon test (when not normal) because $\texttt{sample}_{\texttt{ARsrc}}$ and $\texttt{sample}_{\texttt{ENsrc}}$ have one direction that is not paired: AR-EN and EN-AR. Other directions can be seen as paired, e.g. AR-ES and EN-ES as both having the same $l_{\text{trg}}$.

---

[13] using implementation from `https://github.com/rtmdrr/replicability-analysis-NLP`

# E DATA STATISTICS

- Number of types, i.e. vocabulary size ( $|V|$). Note that Sockeye adds for its calculation 4 additional types: <pad>, <s>, </s>, <unk>.
- Number of tokens. This excludes the 1 EOS/BOS (end-/beginning-of-sentence) marker added by Sockeye to each line.
- Out-of-vocabulary (OOV) type rate (in %), i.e. the fraction of the types in the dev data that is not covered by the types in the training data.
- OOV token rate (in %), i.e. the fraction of tokens in the dev data that is treated as UNKnowns.
- Type-token-ratio (in %), i.e. the ratio between the number of types and tokens in the data. This is a rough proxy for lexical diversity in that a value of 1 would indicate that no type is ever seen twice, and a value very close to 0 would indicate that very few distinct types account for almost all of the data.
- Line length (excl. EOS/BOS marker): mean±standard deviation, and the 0/25/50/75/100-th percentile.

## Statistics for dataset A

| Representation | CHAR | | | | | BYTE | | | | | WORD | | | | | BPE | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Number of lines | 100 | 1,000 | 10,000 | 100,000 | 1,000,000 | 100 | 1,000 | 10,000 | 100,000 | 1,000,000 | 100 | 1,000 | 10,000 | 100,000 | 1,000,000 | 100 | 1,000 | 10,000 | 100,000 | 1,000,000 |
| **Number of TYPES** | | | | | | | | | | | | | | | | | | | | |
| AR | 82 | 107 | 149 | 176 | 239 | 73 | 95 | 130 | 145 | 162 | 866 | 5,865 | 26,105 | 97,994 | 304,975 | 700 | 4,171 | 18,202 | 29,706 | 29,961 |
| EN | 75 | 81 | 122 | 167 | 244 | 76 | 82 | 127 | 151 | 166 | 679 | 3,289 | 12,011 | 37,261 | 122,454 | 645 | 2,856 | 10,225 | 26,184 | 29,068 |
| ES | 84 | 90 | 114 | 173 | 254 | 85 | 91 | 118 | 151 | 166 | 707 | 3,836 | 15,113 | 49,496 | 142,806 | 661 | 3,232 | 12,083 | 27,908 | 29,445 |
| FR | 86 | 94 | 130 | 175 | 256 | 87 | 96 | 130 | 153 | 172 | 742 | 3,878 | 13,995 | 42,867 | 118,201 | 723 | 3,193 | 11,731 | 27,393 | 29,333 |
| RU | 95 | 108 | 160 | 187 | 260 | 97 | 110 | 148 | 157 | 166 | 883 | 5,430 | 23,400 | 81,339 | 236,085 | 842 | 4,319 | 17,995 | 29,239 | 29,807 |
| ZH | 602 | 1,236 | 2,267 | 3,400 | 4,885 | 137 | 143 | 153 | 164 | 193 | 627 | 3,178 | 13,134 | 46,938 | 134,489 | 800 | 2,849 | 10,656 | 27,284 | 29,148 |
| ZH_pinyin | 79 | 90 | 104 | 154 | 564 | | | | | | | | | | | | | | | |
| ZH_wubi | 100 | 106 | 123 | 170 | 542 | | | | | | | | | | | | | | | |
| AR_cp1256 | | | | | | 82 | 107 | 149 | 176 | 239 | | | | | | | | | | |
| RU_cp1251 | | | | | | 95 | 108 | 160 | 187 | 260 | | | | | | | | | | |
| **Number of TOKENS** | | | | | | | | | | | | | | | | | | | | |
| AR | 9,079 | 123,832 | 1,083,517 | 10,625,047 | 102,064,230 | 16,655 | 227,163 | 1,985,014 | 19,487,689 | 186,171,180 | 1,776 | 23,460 | 206,549 | 2,035,190 | 19,410,502 | 2,907 | 28,227 | 223,510 | 2,218,044 | 21,344,900 |
| EN | 11,730 | 159,444 | 1,344,001 | 13,132,862 | 123,491,871 | 11,731 | 159,449 | 1,345,771 | 13,158,948 | 123,705,128 | 2,071 | 27,398 | 236,569 | 2,339,109 | 21,943,139 | 2,980 | 29,778 | 243,628 | 2,369,723 | 22,351,059 |
| ES | 12,374 | 171,104 | 1,484,804 | 14,549,703 | 138,596,036 | 12,629 | 175,286 | 1,513,782 | 14,821,495 | 141,276,766 | 2,232 | 29,461 | 263,024 | 2,588,791 | 24,654,449 | 3,180 | 32,323 | 271,995 | 2,644,055 | 25,264,439 |
| FR | 12,456 | 179,048 | 1,490,983 | 14,528,593 | 138,049,189 | 12,875 | 185,227 | 1,542,105 | 15,055,657 | 143,495,667 | 2,298 | 32,011 | 273,195 | 2,684,982 | 25,595,487 | 3,245 | 34,801 | 280,928 | 2,727,386 | 26,077,553 |
| RU | 11,980 | 168,156 | 1,436,078 | 14,151,728 | 134,706,120 | 21,751 | 309,279 | 2,636,591 | 25,990,263 | 247,098,758 | 1,854 | 24,746 | 216,638 | 2,150,746 | 20,421,965 | 3,110 | 29,106 | 231,293 | 2,307,656 | 22,066,945 |
| ZH | 3,318 | 42,572 | 372,003 | 3,659,617 | 34,672,612 | 8,559 | 116,667 | 1,019,969 | 9,990,046 | 94,268,840 | 1,751 | 23,568 | 207,714 | 2,038,639 | 19,361,101 | 2,223 | 25,459 | 215,179 | 2,083,632 | 19,885,789 |
| ZH_pinyin | 10,811 | 148,083 | 1,279,586 | 12,530,567 | 118,471,229 | | | | | | | | | | | | | | | |
| ZH_wubi | 9,245 | 127,190 | 1,104,529 | 10,885,508 | 102,908,155 | | | | | | | | | | | | | | | |
| AR_cp1256 | | | | | | 9,079 | 123,832 | 1,083,517 | 10,625,047 | 102,064,230 | | | | | | | | | | |
| RU_cp1251 | | | | | | 11,980 | 168,156 | 1,436,078 | 14,151,728 | 134,706,120 | | | | | | | | | | |
| **OOV type rate (%)** | | | | | | | | | | | | | | | | | | | | |
| AR | 40.77 | 23.08 | 2.31 | 0.00 | 0.00 | 40.65 | 22.76 | 2.44 | 0.00 | 0.00 | 95.80 | 79.80 | 48.02 | 24.19 | 8.14 | 10.17 | 1.39 | 0.09 | 0.00 | 0.00 |
| EN | 23.71 | 20.62 | 11.34 | 4.12 | 2.06 | 26.47 | 23.53 | 6.86 | 0.98 | 0.00 | 92.07 | 69.08 | 33.56 | 15.27 | 7.38 | 4.73 | 1.01 | 0.18 | 0.05 | 0.01 |
| ES | 25.23 | 20.72 | 16.22 | 8.11 | 7.21 | 23.64 | 19.09 | 11.82 | 4.55 | 0.91 | 93.32 | 71.90 | 34.72 | 13.24 | 6.12 | 5.44 | 1.06 | 0.27 | 0.10 | 0.08 |
| FR | 25.66 | 19.47 | 11.50 | 4.42 | 2.65 | 27.12 | 20.34 | 7.63 | 2.54 | 0.00 | 92.50 | 70.25 | 32.16 | 12.31 | 5.62 | 5.11 | 0.99 | 0.26 | 0.05 | 0.02 |
| RU | 34.93 | 26.71 | 4.11 | 2.05 | 0.00 | 32.64 | 24.31 | 3.47 | 1.39 | 0.00 | 94.52 | 75.84 | 39.14 | 14.54 | 5.69 | 6.38 | 1.27 | 0.06 | 0.02 | 0.00 |
| ZH | 69.79 | 39.73 | 10.58 | 1.87 | 0.56 | 10.46 | 7.19 | 3.27 | 1.31 | 0.00 | 93.19 | 73.29 | 38.59 | 15.63 | 5.65 | 76.09 | 40.95 | 8.52 | 0.89 | 0.21 |
| ZH_pinyin | 20.41 | 15.31 | 8.16 | 4.08 | 0.00 | | | | | | | | | | | | | | | |
| ZH_wubi | 17.50 | 13.33 | 5.83 | 3.33 | 0.00 | | | | | | | | | | | | | | | |
| AR_cp1256 | | | | | | 40.77 | 23.08 | 2.31 | 0.00 | 0.00 | | | | | | | | | | |
| RU_cp1251 | | | | | | 34.93 | 26.71 | 4.11 | 2.05 | 0.00 | | | | | | | | | | |
| **OOV token rate (%)** | | | | | | | | | | | | | | | | | | | | |
| AR | 0.75 | 0.35 | 0.00 | 0.00 | 0.00 | 0.41 | 0.19 | 0.00 | 0.00 | 0.00 | 60.86 | 40.28 | 17.94 | 8.62 | 2.10 | 2.57 | 1.99 | 0.11 | 0.00 | 0.00 |
| EN | 0.05 | 0.02 | 0.00 | 0.00 | 0.00 | 0.06 | 0.02 | 0.00 | 0.00 | 0.00 | 39.04 | 18.57 | 6.16 | 2.70 | 1.01 | 0.35 | 0.08 | 0.02 | 0.01 | 0.00 |
| ES | 0.07 | 0.01 | 0.01 | 0.01 | 0.01 | 0.07 | 0.02 | 0.01 | 0.01 | 0.00 | 36.21 | 18.28 | 6.28 | 2.34 | 0.86 | 0.28 | 0.13 | 0.05 | 0.03 | 0.03 |
| FR | 0.19 | 0.10 | 0.07 | 0.00 | 0.00 | 0.21 | 0.11 | 0.07 | 0.00 | 0.00 | 35.44 | 17.47 | 6.04 | 2.13 | 0.76 | 0.64 | 0.52 | 0.36 | 0.01 | 0.00 |
| RU | 0.30 | 0.24 | 0.16 | 0.00 | 0.00 | 0.16 | 0.13 | 0.08 | 0.00 | 0.00 | 53.22 | 30.94 | 12.59 | 4.17 | 1.28 | 0.88 | 1.07 | 0.84 | 0.01 | 0.00 |
| ZH | 17.83 | 3.64 | 0.38 | 0.04 | 0.01 | 0.05 | 0.02 | 0.00 | 0.00 | 0.00 | 45.48 | 22.08 | 8.34 | 3.20 | 0.83 | 31.67 | 11.76 | 1.66 | 0.14 | 0.03 |
| ZH_pinyin | 0.06 | 0.03 | 0.00 | 0.00 | 0.00 | | | | | | | | | | | | | | | |
| ZH_wubi | 0.07 | 0.04 | 0.00 | 0.00 | 0.00 | | | | | | | | | | | | | | | |
| AR_cp1256 | | | | | | 0.75 | 0.35 | 0.00 | 0.00 | 0.00 | | | | | | | | | | |
| RU_cp1251 | | | | | | 0.30 | 0.24 | 0.16 | 0.00 | 0.00 | | | | | | | | | | |
| **TTR (%)** | | | | | | | | | | | | | | | | | | | | |
| AR | 0.90 | 0.09 | 0.01 | 0.00 | 0.00 | 0.44 | 0.04 | 0.01 | 0.00 | 0.00 | 48.76 | 25.00 | 12.64 | 4.81 | 1.57 | 24.08 | 14.78 | 8.14 | 1.34 | 0.14 |
| EN | 0.64 | 0.05 | 0.01 | 0.00 | 0.00 | 0.65 | 0.05 | 0.01 | 0.00 | 0.00 | 32.79 | 12.00 | 5.08 | 1.59 | 0.56 | 21.64 | 9.59 | 4.20 | 1.10 | 0.13 |
| ES | 0.68 | 0.05 | 0.01 | 0.00 | 0.00 | 0.67 | 0.05 | 0.01 | 0.00 | 0.00 | 31.68 | 13.02 | 5.75 | 1.91 | 0.58 | 20.79 | 10.00 | 4.44 | 1.06 | 0.12 |
| FR | 0.69 | 0.05 | 0.01 | 0.00 | 0.00 | 0.68 | 0.05 | 0.01 | 0.00 | 0.00 | 32.29 | 12.11 | 5.12 | 1.60 | 0.46 | 22.28 | 9.18 | 4.18 | 1.00 | 0.11 |
| RU | 0.79 | 0.06 | 0.01 | 0.00 | 0.00 | 0.45 | 0.04 | 0.01 | 0.00 | 0.00 | 47.63 | 21.94 | 10.80 | 3.78 | 1.16 | 27.07 | 14.84 | 7.78 | 1.27 | 0.14 |
| ZH | 18.14 | 2.90 | 0.61 | 0.09 | 0.01 | 1.60 | 0.12 | 0.02 | 0.00 | 0.00 | 35.81 | 13.48 | 6.32 | 2.30 | 0.69 | 35.99 | 11.19 | 4.95 | 1.31 | 0.15 |
| ZH_pinyin | 0.73 | 0.06 | 0.01 | 0.00 | 0.00 | | | | | | | | | | | | | | | |
| ZH_wubi | 1.08 | 0.08 | 0.01 | 0.00 | 0.00 | | | | | | | | | | | | | | | |
| AR_cp1256 | | | | | | 0.90 | 0.09 | 0.01 | 0.00 | 0.00 | | | | | | | | | | |
| RU_cp1251 | | | | | | 0.79 | 0.06 | 0.01 | 0.00 | 0.00 | | | | | | | | | | |
| **Mean line length±std 0/25/50/75/100-th** | | | | | | | | | | | | | | | | | | | | |
| AR | 90.79 ± 54.45<br>8/33/91/136/223 | 123.83 ± 47.09<br>5/94/126/155/263 | 108.35 ± 53.79<br>1/69/109/147/299 | 106.25 ± 54.88<br>1/64/107/147/299 | 102.06 ± 57.25<br>1/54/102/145/300 | 166.55 ± 99.88<br>16/62/166/247/404 | 227.16 ± 86.09<br>9/172/230/285/479 | 198.50 ± 99.02<br>1/126/200/269/554 | 194.88 ± 101.03<br>1/117/196/270/561 | 186.17 ± 104.86<br>1/98/187/265/565 | 17.76 ± 10.78<br>1/7/17/27/48 | 23.46 ± 9.12<br>1/18/24/29/56 | 20.65 ± 10.24<br>1/13/20/28/104 | 20.35 ± 10.46<br>1/13/20/28/107 | 19.41 ± 10.83<br>1/11/19/27/137 | 29.07 ± 17.48<br>3/14/27/43/75 | 28.23 ± 11.33<br>1/21/28/35/71 | 22.35 ± 11.30<br>1/14/22/30/113 | 22.18 ± 11.66<br>1/14/22/30/122 | 21.34 ± 12.13<br>1/12/21/30/173 |
| EN | 117.30 ± 72.51<br>8/49/112/170/265 | 159.44 ± 58.88<br>8/117/163/203/285 | 134.40 ± 66.79<br>1/86/137/185/300 | 131.33 ± 68.25<br>1/79/134/183/300 | 123.49 ± 69.53<br>1/65/125/177/300 | 117.31 ± 72.50<br>8/49/112/170/265 | 159.45 ± 58.88<br>8/117/163/203/285 | 134.58 ± 66.80<br>1/87/137/185/300 | 131.59 ± 68.34<br>1/80/134/184/300 | 123.71 ± 69.64<br>1/65/125/177/311 | 20.71 ± 12.16<br>1/9/19/30/46 | 27.40 ± 10.35<br>1/20/28/35/60 | 23.66 ± 11.68<br>1/15/24/32/93 | 23.39 ± 12.02<br>1/15/24/32/106 | 21.94 ± 12.29<br>1/12/22/31/136 | 29.80 ± 17.18<br>3/16/27/42/70 | 29.78 ± 11.32<br>1/22/30/37/68 | 24.36 ± 12.09<br>1/16/24/33/101 | 23.70 ± 12.20<br>1/15/24/32/106 | 22.35 ± 12.53<br>1/12/22/32/138 |
| ES | 123.74 ± 78.91<br>6/42/119/186/296 | 171.10 ± 64.03<br>6/127/175/217/297 | 148.48 ± 73.38<br>1/95/150/205/300 | 145.50 ± 75.32<br>1/88/148/204/300 | 138.60 ± 77.56<br>1/73/140/200/300 | 126.29 ± 80.21<br>7/44/121/191/302 | 175.29 ± 65.44<br>7/130/179/222/306 | 151.38 ± 74.70<br>1/97/153/209/307 | 148.21 ± 76.67<br>1/89/150/208/312 | 141.28 ± 78.96<br>1/74/143/203/318 | 22.32 ± 13.97<br>1/7/22/32/52 | 29.46 ± 11.35<br>1/22/29/37/63 | 26.30 ± 13.03<br>1/17/26/36/104 | 25.89 ± 13.35<br>1/16/26/36/107 | 24.65 ± 13.78<br>1/13/25/35/139 | 31.80 ± 19.39<br>4/16/31/43/79 | 32.32 ± 12.60<br>1/24/32/41/72 | 27.20 ± 13.57<br>1/18/27/37/106 | 26.44 ± 13.67<br>1/16/27/37/107 | 25.26 ± 14.14<br>1/14/25/36/140 |
| FR | 124.56 ± 79.38<br>5/45/116/190/287 | 179.05 ± 66.44<br>5/133/183/228/300 | 149.10 ± 74.19<br>1/93/152/206/300 | 145.29 ± 75.75<br>1/87/148/204/300 | 138.05 ± 77.90<br>1/72/140/199/300 | 128.75 ± 82.16<br>5/47/121/196/297 | 185.23 ± 68.52<br>5/138/190/236/312 | 154.21 ± 76.66<br>1/96/157/213/322 | 150.56 ± 78.42<br>1/90/153/212/322 | 143.50 ± 80.99<br>1/75/145/207/331 | 22.98 ± 14.16<br>1/9/22/34/50 | 32.01 ± 11.99<br>1/24/32/41/70 | 27.32 ± 13.53<br>1/17/28/37/107 | 26.85 ± 13.89<br>1/16/27/37/110 | 25.60 ± 14.45<br>1/14/26/37/138 | 32.45 ± 19.52<br>2/16/30/46/76 | 34.80 ± 13.09<br>2/27/35/44/77 | 28.09 ± 13.96<br>1/18/28/38/107 | 27.27 ± 14.13<br>1/17/28/38/110 | 26.08 ± 14.71<br>1/14/26/37/138 |
| RU | 119.80 ± 72.09<br>10/44/119/183/260 | 168.16 ± 62.11<br>5/126/174/213/298 | 143.61 ± 71.55<br>1/91/146/197/300 | 141.52 ± 73.95<br>1/85/144/199/300 | 134.71 ± 76.20<br>1/70/136/194/300 | 217.51 ± 134.51<br>20/77/215/332/489 | 309.28 ± 114.98<br>10/231/318/392/554 | 263.66 ± 132.77<br>1/167/268/363/562 | 259.90 ± 137.10<br>1/155/264/365/566 | 247.10 ± 141.14<br>1/127/249/357/567 | 18.54 ± 10.65<br>1/8/18/27/39 | 24.75 ± 9.53<br>1/19/25/31/58 | 21.66 ± 10.81<br>1/14/22/29/104 | 21.51 ± 11.12<br>1/13/22/30/107 | 20.42 ± 11.44<br>1/11/20/29/136 | 31.10 ± 17.42<br>4/18/29/45/69 | 29.11 ± 11.51<br>1/21/29/37/68 | 23.13 ± 11.71<br>1/15/23/31/104 | 23.08 ± 12.17<br>1/14/23/32/110 | 22.07 ± 12.57<br>1/12/22/31/166 |
| ZH | 33.18 ± 18.38<br>3/19/30/45/88 | 42.57 ± 16.91<br>2/32/42/53/127 | 37.20 ± 18.46<br>1/24/37/49/165 | 36.60 ± 19.05<br>1/23/36/49/241 | 34.67 ± 19.29<br>1/19/34/48/281 | 85.59 ± 51.18<br>7/35/81/127/192 | 116.67 ± 43.82<br>6/88/117/148/243 | 102.00 ± 50.61<br>1/66/103/139/297 | 99.90 ± 51.49<br>1/61/101/138/337 | 94.27 ± 52.80<br>1/51/95/134/594 | 17.51 ± 10.68<br>2/7/17/25/46 | 23.57 ± 9.52<br>1/18/23/29/76 | 20.77 ± 10.48<br>1/13/20/28/108 | 20.39 ± 10.60<br>1/13/20/28/111 | 19.36 ± 10.87<br>1/11/19/27/167 | 22.23 ± 12.33<br>2/13/20/31/58 | 25.46 ± 10.29<br>1/19/25/32/79 | 21.52 ± 10.94<br>1/14/21/29/108 | 20.84 ± 10.92<br>1/13/21/28/111 | 19.89 ± 11.23<br>1/11/20/28/167 |
| ZH_pinyin | 108.11 ± 65.51<br>7/43/98/161/255 | 148.08 ± 55.57<br>7/109/149/188/312 | 127.96 ± 63.96<br>1/82/129/175/369 | 125.31 ± 65.25<br>1/76/126/174/407 | 118.47 ± 66.88<br>1/63/119/169/645 | | | | | | | | | | | | | | | |
| ZH_wubi | 92.45 ± 55.26<br>8/39/86/137/218 | 127.19 ± 48.00<br>7/94/129/158/266 | 110.45 ± 54.71<br>1/71/111/150/327 | 108.86 ± 56.04<br>1/67/110/150/364 | 102.91 ± 57.41<br>1/55/103/146/627 | | | | | | | | | | | | | | | |
| AR_cp1256 | | | | | | 90.79 ± 54.45<br>8/33/91/136/223 | 123.83 ± 47.09<br>5/94/126/155/263 | 108.35 ± 53.79<br>1/69/109/147/299 | 106.25 ± 54.88<br>1/64/107/147/299 | 102.06 ± 57.25<br>1/54/102/145/300 | | | | | | | | | | |
| RU_cp1251 | | | | | | 119.80 ± 72.09<br>10/44/119/183/260 | 168.16 ± 62.11<br>5/126/174/213/298 | 143.61 ± 71.55<br>1/91/146/197/300 | 141.52 ± 73.95<br>1/85/144/199/300 | 134.71 ± 76.20<br>1/70/136/194/300 | | | | | | | | | | |

## Statistics for dataset B

| Representation | CHAR | | | | | BYTE | | | | | WORD | | | | | BPE | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Number of lines | 100 | 1,000 | 10,000 | 100,000 | 1,000,000 | 100 | 1,000 | 10,000 | 100,000 | 1,000,000 | 100 | 1,000 | 10,000 | 100,000 | 1,000,000 | 100 | 1,000 | 10,000 | 100,000 | 1,000,000 |
| **Number of TYPES** | | | | | | | | | | | | | | | | | | | | |
| AR | 87 | 134 | 168 | 251 | 431 | 88 | 124 | 147 | 159 | 181 | 1,129 | 6,719 | 28,222 | 97,467 | 311,355 | 810 | 4,408 | 18,800 | 29,680 | 29,959 |
| EN | 76 | 99 | 131 | 189 | 343 | 79 | 103 | 135 | 163 | 181 | 876 | 4,148 | 13,660 | 42,844 | 140,912 | 784 | 3,408 | 10,756 | 26,820 | 29,178 |
| ES | 86 | 104 | 134 | 189 | 366 | 89 | 108 | 133 | 161 | 185 | 970 | 4,775 | 16,735 | 51,366 | 154,482 | 814 | 3,705 | 12,529 | 28,068 | 29,471 |
| FR | 89 | 112 | 147 | 202 | 352 | 91 | 120 | 142 | 162 | 185 | 955 | 4,636 | 15,639 | 45,348 | 132,881 | 812 | 3,670 | 12,258 | 27,744 | 29,370 |
| RU | 109 | 142 | 173 | 204 | 320 | 112 | 141 | 159 | 161 | 180 | 1,142 | 6,520 | 25,959 | 83,937 | 251,355 | 993 | 4,829 | 18,837 | 29,261 | 29,811 |
| ZH | 761 | 1,575 | 2,402 | 3,465 | 4,705 | 123 | 153 | 165 | 175 | 188 | 897 | 4,162 | 14,714 | 46,788 | 140,188 | 1,068 | 3,717 | 11,069 | 27,409 | 29,107 |
| ZH_pinyin | 71 | 104 | 140 | 241 | 430 | | | | | | | | | | | | | | | |
| ZH_wubi | 87 | 124 | 154 | 246 | 421 | | | | | | | | | | | | | | | |
| AR_cp1256 | | | | | | 87 | 134 | 168 | 251 | 431 | | | | | | | | | | |
| RU_cp1251 | | | | | | 109 | 142 | 173 | 204 | 320 | | | | | | | | | | |
| **Number of TOKENS** | | | | | | | | | | | | | | | | | | | | |
| AR | 9,798 | 100,599 | 1,019,696 | 10,248,976 | 102,481,816 | 17,656 | 181,866 | 1,842,138 | 18,517,669 | 185,172,618 | 1,825 | 18,760 | 190,465 | 1,915,746 | 19,138,724 | 3,387 | 25,234 | 210,622 | 2,090,403 | 20,986,116 |
| EN | 11,816 | 119,796 | 1,201,839 | 12,087,922 | 120,969,163 | 11,820 | 119,839 | 1,202,217 | 12,091,602 | 121,006,803 | 2,047 | 21,058 | 211,029 | 2,122,365 | 21,234,437 | 3,262 | 24,634 | 220,390 | 2,159,566 | 21,669,769 |
| ES | 13,543 | 135,106 | 1,363,454 | 13,709,493 | 137,180,542 | 13,790 | 137,613 | 1,389,315 | 13,967,072 | 139,753,263 | 2,390 | 23,858 | 241,789 | 2,431,579 | 24,326,055 | 3,702 | 28,179 | 253,075 | 2,485,292 | 24,933,987 |
| FR | 13,173 | 134,702 | 1,358,567 | 13,662,024 | 136,811,870 | 13,617 | 139,250 | 1,404,001 | 14,121,629 | 141,403,257 | 2,362 | 24,442 | 247,027 | 2,485,203 | 24,874,355 | 3,664 | 28,541 | 257,305 | 2,527,889 | 25,374,629 |
| RU | 13,106 | 130,289 | 1,325,793 | 13,356,717 | 133,699,473 | 24,059 | 239,145 | 2,433,154 | 24,512,949 | 245,396,332 | 1,960 | 19,642 | 200,188 | 2,014,889 | 20,153,124 | 3,760 | 26,136 | 218,325 | 2,170,410 | 21,792,216 |
| ZH | 3,329 | 33,768 | 340,692 | 3,419,803 | 34,206,106 | 9,122 | 92,689 | 935,172 | 9,402,743 | 94,064,353 | 1,866 | 18,697 | 189,773 | 1,906,704 | 19,073,133 | 2,563 | 21,630 | 199,028 | 1,946,543 | 19,574,572 |
| ZH_pinyin | 11,391 | 115,577 | 1,166,227 | 11,732,264 | 117,367,349 | | | | | | | | | | | | | | | |
| ZH_wubi | 10,035 | 101,049 | 1,019,113 | 10,242,540 | 102,460,573 | | | | | | | | | | | | | | | |
| AR_cp1256 | | | | | | 9,798 | 100,599 | 1,019,696 | 10,248,976 | 102,481,816 | | | | | | | | | | |
| RU_cp1251 | | | | | | 13,106 | 130,289 | 1,325,793 | 13,356,717 | 133,699,473 | | | | | | | | | | |
| **OOV type rate (%)** | | | | | | | | | | | | | | | | | | | | |
| AR | 35.38 | 5.38 | 0.77 | 0.00 | 0.00 | 28.46 | 4.88 | 0.81 | 0.00 | 0.00 | 93.61 | 72.07 | 35.51 | 13.07 | 4.82 | 7.86 | 0.30 | 0.02 | 0.00 | 0.00 |
| EN | 25.77 | 11.34 | 8.25 | 2.06 | 0.00 | 26.47 | 11.76 | 5.88 | 0.00 | 0.00 | 89.58 | 61.04 | 26.28 | 10.25 | 4.18 | 4.00 | 0.44 | 0.15 | 0.04 | 0.00 |
| ES | 26.13 | 12.61 | 11.71 | 3.60 | 0.00 | 24.55 | 10.91 | 8.18 | 0.91 | 0.00 | 90.59 | 64.27 | 27.73 | 9.61 | 3.95 | 4.08 | 0.44 | 0.17 | 0.06 | 0.00 |
| FR | 23.01 | 11.50 | 7.96 | 2.65 | 0.00 | 24.58 | 7.63 | 5.08 | 0.85 | 0.00 | 90.23 | 62.61 | 25.79 | 8.71 | 3.36 | 3.87 | 0.47 | 0.12 | 0.05 | 0.00 |
| RU | 26.71 | 6.16 | 2.74 | 0.00 | 0.00 | 23.61 | 4.86 | 2.08 | 0.00 | 0.00 | 93.20 | 69.75 | 32.20 | 11.10 | 3.91 | 4.39 | 0.23 | 0.03 | 0.01 | 0.00 |
| ZH | 61.84 | 26.97 | 6.68 | 1.01 | 0.15 | 20.26 | 2.61 | 0.65 | 0.00 | 0.00 | 89.73 | 64.10 | 29.54 | 10.44 | 3.84 | 67.89 | 27.00 | 5.27 | 0.67 | 0.04 |
| ZH_pinyin | 27.55 | 8.16 | 2.04 | 0.00 | 0.00 | | | | | | | | | | | | | | | |
| ZH_wubi | 28.33 | 4.17 | 0.83 | 0.00 | 0.00 | | | | | | | | | | | | | | | |
| AR_cp1256 | | | | | | 35.38 | 5.38 | 0.77 | 0.00 | 0.00 | | | | | | | | | | |
| RU_cp1251 | | | | | | 26.71 | 6.16 | 2.74 | 0.00 | 0.00 | | | | | | | | | | |
| **OOV token rate (%)** | | | | | | | | | | | | | | | | | | | | |
| AR | 0.31 | 0.01 | 0.00 | 0.00 | 0.00 | 0.11 | 0.00 | 0.00 | 0.00 | 0.00 | 53.94 | 25.89 | 9.26 | 3.14 | 1.15 | 0.82 | 0.08 | 0.00 | 0.00 | 0.00 |
| EN | 0.05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.00 | 0.00 | 0.00 | 0.00 | 34.11 | 12.18 | 3.50 | 1.22 | 0.50 | 0.30 | 0.02 | 0.01 | 0.01 | 0.00 |
| ES | 0.04 | 0.01 | 0.01 | 0.00 | 0.00 | 0.04 | 0.01 | 0.01 | 0.00 | 0.00 | 31.52 | 12.57 | 3.72 | 1.15 | 0.46 | 0.19 | 0.03 | 0.04 | 0.02 | 0.00 |
| FR | 0.07 | 0.01 | 0.00 | 0.00 | 0.00 | 0.08 | 0.00 | 0.00 | 0.00 | 0.00 | 30.96 | 11.91 | 3.38 | 0.99 | 0.39 | 0.22 | 0.05 | 0.01 | 0.01 | 0.00 |
| RU | 0.22 | 0.00 | 0.00 | 0.00 | 0.00 | 0.11 | 0.00 | 0.00 | 0.00 | 0.00 | 49.05 | 22.75 | 7.55 | 2.36 | 0.81 | 0.71 | 0.02 | 0.00 | 0.00 | 0.00 |
| ZH | 10.44 | 1.21 | 0.16 | 0.02 | 0.00 | 0.10 | 0.00 | 0.00 | 0.00 | 0.00 | 37.81 | 13.74 | 4.42 | 1.42 | 0.53 | 22.02 | 4.99 | 0.79 | 0.10 | 0.00 |
| ZH_pinyin | 0.03 | 0.00 | 0.00 | 0.00 | 0.00 | | | | | | | | | | | | | | | |
| ZH_wubi | 0.09 | 0.00 | 0.00 | 0.00 | 0.00 | | | | | | | | | | | | | | | |
| AR_cp1256 | | | | | | 0.31 | 0.01 | 0.00 | 0.00 | 0.00 | | | | | | | | | | |
| RU_cp1251 | | | | | | 0.22 | 0.00 | 0.00 | 0.00 | 0.00 | | | | | | | | | | |
| **TTR (%)** | | | | | | | | | | | | | | | | | | | | |
| AR | 0.89 | 0.13 | 0.02 | 0.00 | 0.00 | 0.50 | 0.07 | 0.01 | 0.00 | 0.00 | 61.86 | 35.82 | 14.82 | 5.09 | 1.63 | 23.91 | 17.47 | 8.93 | 1.42 | 0.14 |
| EN | 0.64 | 0.08 | 0.01 | 0.00 | 0.00 | 0.67 | 0.09 | 0.01 | 0.00 | 0.00 | 42.79 | 19.70 | 6.47 | 2.02 | 0.66 | 24.03 | 13.83 | 4.88 | 1.24 | 0.13 |
| ES | 0.64 | 0.08 | 0.01 | 0.00 | 0.00 | 0.65 | 0.08 | 0.01 | 0.00 | 0.00 | 40.59 | 20.01 | 6.92 | 2.11 | 0.64 | 21.99 | 13.15 | 4.95 | 1.13 | 0.12 |
| FR | 0.68 | 0.08 | 0.01 | 0.00 | 0.00 | 0.67 | 0.09 | 0.01 | 0.00 | 0.00 | 40.43 | 18.97 | 6.33 | 1.82 | 0.53 | 22.16 | 12.86 | 4.76 | 1.10 | 0.12 |
| RU | 0.83 | 0.11 | 0.01 | 0.00 | 0.00 | 0.47 | 0.06 | 0.01 | 0.00 | 0.00 | 58.27 | 33.19 | 12.97 | 4.17 | 1.25 | 26.41 | 18.48 | 8.63 | 1.35 | 0.14 |
| ZH | 22.86 | 4.66 | 0.71 | 0.10 | 0.01 | 1.35 | 0.17 | 0.02 | 0.00 | 0.00 | 48.07 | 22.26 | 7.75 | 2.45 | 0.74 | 41.67 | 17.18 | 5.56 | 1.41 | 0.15 |
| ZH_pinyin | 0.62 | 0.09 | 0.01 | 0.00 | 0.00 | | | | | | | | | | | | | | | |
| ZH_wubi | 0.87 | 0.12 | 0.02 | 0.00 | 0.00 | | | | | | | | | | | | | | | |
| AR_cp1256 | | | | | | 0.89 | 0.13 | 0.02 | 0.00 | 0.00 | | | | | | | | | | |
| RU_cp1251 | | | | | | 0.83 | 0.11 | 0.01 | 0.00 | 0.00 | | | | | | | | | | |
| **Mean line length±std 0/25/50/75/100-th** | | | | | | | | | | | | | | | | | | | | |
| AR | 97.98 ± 55.47<br>9/53/94/147/234 | 100.60 ± 59.60<br>2/49/101/147/274 | 101.97 ± 59.10<br>1/51/101/147/300 | 102.49 ± 58.36<br>1/53/103/147/300 | 102.48 ± 58.39<br>1/53/102/147/300 | 176.56 ± 101.97<br>14/98/170/266/432 | 181.87 ± 108.68<br>2/88/182/266/501 | 184.21 ± 107.49<br>1/90/184/267/551 | 185.18 ± 106.22<br>1/95/185/266/552 | 185.17 ± 106.29<br>1/95/185/266/561 | 18.25 ± 10.12<br>2/10/17/27/43 | 18.76 ± 11.10<br>1/9/18/27/73 | 19.05 ± 10.95<br>1/10/19/27/84 | 19.16 ± 10.83<br>1/10/19/27/143 | 19.14 ± 10.82<br>1/10/19/27/129 | 33.87 ± 18.38<br>2/20/32/50/73 | 25.23 ± 15.26<br>1/13/24/36/96 | 21.06 ± 12.27<br>1/11/21/30/95 | 20.90 ± 12.02<br>1/11/21/29/146 | 20.99 ± 12.08<br>1/11/21/29/165 |
| EN | 118.16 ± 68.05<br>7/64/114/176/256 | 119.80 ± 71.15<br>2/58/122/174/288 | 120.18 ± 70.00<br>1/60/121/176/299 | 120.88 ± 69.43<br>1/62/122/175/300 | 120.97 ± 69.54<br>1/62/122/175/300 | 118.20 ± 68.05<br>7/64/114/176/256 | 119.84 ± 71.17<br>2/58/122/175/288 | 120.22 ± 70.02<br>1/60/121/176/299 | 120.92 ± 69.44<br>1/62/122/175/300 | 121.01 ± 69.55<br>1/62/122/175/303 | 20.47 ± 11.02<br>1/12/19/31/42 | 21.06 ± 12.60<br>1/11/21/31/74 | 21.10 ± 12.25<br>1/11/21/30/86 | 21.22 ± 12.14<br>1/11/21/30/117 | 21.23 ± 12.15<br>1/11/21/30/127 | 32.62 ± 17.93<br>2/19/31/47/81 | 24.63 ± 14.90<br>1/13/24/35/85 | 22.04 ± 12.78<br>1/11/22/31/89 | 21.60 ± 12.34<br>1/12/22/31/117 | 21.67 ± 12.40<br>1/12/22/31/129 |
| ES | 135.43 ± 77.54<br>11/80/133/200/286 | 135.11 ± 79.66<br>2/65/137/200/300 | 136.35 ± 79.00<br>1/67/138/200/300 | 137.09 ± 78.22<br>1/70/139/199/300 | 137.18 ± 78.45<br>1/70/139/200/300 | 137.90 ± 78.93<br>11/81/134/202/288 | 137.61 ± 81.03<br>2/65/140/205/305 | 138.93 ± 80.43<br>1/69/140/204/309 | 139.67 ± 79.61<br>1/71/141/203/311 | 139.75 ± 79.84<br>1/71/141/203/318 | 23.90 ± 13.29<br>1/14/24/34/49 | 23.86 ± 14.12<br>1/12/24/35/80 | 24.18 ± 13.94<br>1/12/24/35/87 | 24.32 ± 13.83<br>1/13/24/35/119 | 24.33 ± 13.86<br>1/13/24/35/130 | 37.02 ± 20.68<br>2/22/36/54/81 | 28.18 ± 16.96<br>1/14/27/41/93 | 25.31 ± 14.60<br>1/13/25/37/91 | 24.85 ± 14.13<br>1/13/25/36/119 | 24.93 ± 14.21<br>1/13/25/36/133 |
| FR | 131.73 ± 75.95<br>8/74/128/196/289 | 134.70 ± 80.48<br>2/64/135/202/300 | 135.86 ± 79.51<br>1/65/137/199/300 | 136.62 ± 78.71<br>1/69/138/199/300 | 136.81 ± 78.90<br>1/69/138/200/300 | 136.17 ± 78.50<br>8/77/132/203/296 | 139.25 ± 83.09<br>2/65/140/207/313 | 140.40 ± 82.18<br>1/68/142/206/315 | 141.22 ± 81.35<br>1/71/143/206/322 | 141.40 ± 81.55<br>1/71/143/206/325 | 23.62 ± 13.40<br>1/13/23/34/53 | 24.44 ± 14.69<br>1/12/24/36/78 | 24.70 ± 14.41<br>1/12/25/36/92 | 24.85 ± 14.27<br>1/13/25/36/118 | 24.87 ± 14.30<br>1/13/25/36/132 | 36.64 ± 21.18<br>2/21/35/54/83 | 28.54 ± 17.31<br>1/15/27/42/94 | 25.73 ± 14.96<br>1/13/26/37/97 | 25.28 ± 14.49<br>1/13/25/36/118 | 25.37 ± 14.57<br>1/13/26/36/134 |
| RU | 131.06 ± 76.89<br>10/71/123/195/271 | 130.29 ± 77.80<br>2/62/130/193/299 | 132.58 ± 77.80<br>1/65/133/195/300 | 133.57 ± 77.09<br>1/67/134/194/300 | 133.70 ± 77.19<br>1/67/134/195/300 | 240.59 ± 144.25<br>12/125/229/363/505 | 239.15 ± 144.35<br>2/113/238/352/558 | 243.32 ± 144.14<br>1/117/245/358/565 | 245.13 ± 142.89<br>1/122/246/357/566 | 245.40 ± 143.08<br>1/122/247/358/569 | 19.60 ± 10.58<br>1/12/20/28/43 | 19.64 ± 11.58<br>1/10/20/29/72 | 20.02 ± 11.63<br>1/10/20/29/85 | 20.15 ± 11.50<br>1/11/20/29/117 | 20.15 ± 11.51<br>1/11/20/29/128 | 37.60 ± 21.30<br>2/21/33/53/91 | 26.14 ± 15.95<br>1/13/25/38/89 | 21.83 ± 12.79<br>1/11/22/31/87 | 21.70 ± 12.55<br>1/11/22/31/120 | 21.79 ± 12.63<br>1/11/22/31/162 |
| ZH | 33.29 ± 17.41<br>4/19/32/47/76 | 33.77 ± 19.77<br>2/18/33/48/148 | 34.07 ± 19.50<br>1/18/34/48/152 | 34.20 ± 19.21<br>1/19/34/48/201 | 34.21 ± 19.26<br>1/19/34/48/289 | 91.22 ± 50.88<br>8/53/90/134/209 | 92.69 ± 55.28<br>2/45/92/133/290 | 93.52 ± 54.29<br>1/48/93/135/283 | 94.03 ± 53.67<br>1/49/94/135/369 | 94.06 ± 53.73<br>1/49/94/135/667 | 18.66 ± 10.28<br>2/11/18/26/48 | 18.70 ± 11.18<br>1/10/18/27/68 | 18.98 ± 11.05<br>1/10/19/27/99 | 19.07 ± 10.88<br>1/10/19/27/116 | 19.07 ± 10.90<br>1/10/19/27/167 | 25.63 ± 13.65<br>2/16/24/34/63 | 21.63 ± 13.19<br>1/11/21/31/86 | 19.90 ± 11.61<br>1/10/20/28/101 | 19.47 ± 11.13<br>1/11/19/27/116 | 19.57 ± 11.24<br>1/11/19/28/167 |
| ZH_pinyin | 113.91 ± 64.46<br>8/66/109/174/265 | 115.58 ± 69.18<br>2/55/115/167/325 | 116.62 ± 68.13<br>1/58/117/169/357 | 117.32 ± 67.41<br>1/60/118/169/394 | 117.37 ± 67.50<br>1/60/118/169/680 | | | | | | | | | | | | | | | |
| ZH_wubi | 100.35 ± 55.70<br>8/59/99/150/226 | 101.05 ± 60.13<br>2/48/100/146/325 | 101.91 ± 58.94<br>1/52/102/147/318 | 102.43 ± 58.21<br>1/53/102/147/381 | 102.46 ± 58.25<br>1/54/103/146/688 | | | | | | | | | | | | | | | |
| AR_cp1256 | | | | | | 97.98 ± 55.47<br>9/53/94/147/234 | 100.60 ± 59.60<br>2/49/101/147/274 | 101.97 ± 59.10<br>1/51/101/147/300 | 102.49 ± 58.36<br>1/53/103/147/300 | 102.48 ± 58.39<br>1/53/102/147/300 | | | | | | | | | | |
| RU_cp1251 | | | | | | 131.06 ± 76.89<br>10/71/123/195/271 | 130.29 ± 77.80<br>2/62/130/193/299 | 132.58 ± 77.80<br>1/65/133/195/300 | 133.57 ± 77.09<br>1/67/134/194/300 | 133.70 ± 77.19<br>1/67/134/195/300 | | | | | | | | | | |

# Statistics for dataset C

| Representation | CHAR | | | | | BYTE | | | | | WORD | | | | | BPE | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Number of lines | 100 | 1,000 | 10,000 | 100,000 | 1,000,000 | 100 | 1,000 | 10,000 | 100,000 | 1,000,000 | 100 | 1,000 | 10,000 | 100,000 | 1,000,000 | 100 | 1,000 | 10,000 | 100,000 | 1,000,000 |
| **Number of TYPES** | | | | | | | | | | | | | | | | | | | | |
| AR | 97 | 136 | 162 | 261 | 427 | 92 | 123 | 144 | 160 | 176 | 1,165 | 6,823 | 28,296 | 97,380 | 310,632 | 864 | 4,530 | 18,735 | 29,673 | 29,954 |
| EN | 78 | 98 | 126 | 186 | 310 | 81 | 103 | 131 | 159 | 178 | 906 | 4,173 | 13,658 | 42,677 | 141,163 | 802 | 3,484 | 10,800 | 26,766 | 29,215 |
| ES | 84 | 108 | 133 | 186 | 306 | 85 | 112 | 133 | 160 | 175 | 970 | 4,734 | 16,850 | 51,281 | 154,649 | 803 | 3,686 | 12,470 | 28,134 | 29,470 |
| FR | 87 | 113 | 139 | 195 | 330 | 91 | 117 | 134 | 159 | 179 | 991 | 4,603 | 15,586 | 45,235 | 132,493 | 853 | 3,693 | 12,170 | 27,731 | 29,355 |
| RU | 98 | 141 | 167 | 211 | 287 | 100 | 140 | 154 | 166 | 177 | 1,165 | 6,490 | 26,112 | 84,137 | 250,778 | 1,002 | 4,858 | 19,029 | 29,322 | 29,814 |
| ZH | 783 | 1,550 | 2,443 | 3,477 | 4,698 | 128 | 152 | 162 | 177 | 189 | 882 | 4,160 | 14,730 | 46,667 | 139,895 | 1,084 | 3,716 | 11,214 | 27,443 | 29,110 |
| ZH_pinyin | 74 | 107 | 139 | 240 | 413 | | | | | | | | | | | | | | | |
| ZH_wubi | 95 | 129 | 154 | 242 | 408 | | | | | | | | | | | | | | | |
| AR_cp1256 | | | | | | 97 | 136 | 162 | 261 | 427 | | | | | | | | | | |
| RU_cp1251 | | | | | | 98 | 141 | 167 | 211 | 287 | | | | | | | | | | |
| **Number of TOKENS** | | | | | | | | | | | | | | | | | | | | |
| AR | 10,212 | 104,144 | 1,021,136 | 10,231,081 | 102,448,339 | 18,448 | 188,507 | 1,845,136 | 18,482,689 | 185,105,498 | 1,889 | 19,281 | 190,738 | 1,912,669 | 19,137,240 | 3,518 | 25,601 | 210,738 | 2,087,431 | 20,981,760 |
| EN | 12,300 | 123,646 | 1,208,827 | 12,089,449 | 120,933,962 | 12,305 | 123,681 | 1,209,187 | 12,093,117 | 120,971,460 | 2,116 | 21,545 | 212,034 | 2,123,153 | 21,231,465 | 3,369 | 25,111 | 221,264 | 2,160,162 | 21,666,482 |
| ES | 14,125 | 139,563 | 1,369,779 | 13,701,387 | 137,133,440 | 14,387 | 142,237 | 1,395,305 | 13,958,221 | 139,702,853 | 2,472 | 24,600 | 242,672 | 2,430,586 | 24,319,760 | 3,784 | 28,726 | 254,092 | 2,483,995 | 24,925,209 |
| FR | 13,797 | 138,528 | 1,365,360 | 13,661,150 | 136,754,360 | 14,243 | 143,139 | 1,411,451 | 14,119,815 | 141,343,675 | 2,453 | 24,950 | 248,488 | 2,485,118 | 24,864,449 | 3,755 | 28,907 | 258,550 | 2,527,957 | 25,362,724 |
| RU | 13,763 | 136,674 | 1,339,286 | 13,369,690 | 133,671,322 | 25,345 | 251,231 | 2,457,938 | 24,535,771 | 245,338,139 | 2,024 | 20,474 | 201,618 | 2,016,451 | 20,156,467 | 3,813 | 26,772 | 219,794 | 2,172,616 | 21,787,539 |
| ZH | 3,371 | 34,429 | 341,536 | 3,419,648 | 34,200,543 | 9,235 | 95,100 | 938,805 | 9,404,386 | 94,052,881 | 1,824 | 19,120 | 190,485 | 1,907,643 | 19,073,463 | 2,559 | 22,067 | 199,706 | 1,947,300 | 19,573,372 |
| ZH_pinyin | 11,605 | 119,258 | 1,171,605 | 11,732,010 | 117,356,880 | | | | | | | | | | | | | | | |
| ZH_wubi | 10,071 | 103,944 | 1,022,870 | 10,245,716 | 102,450,581 | | | | | | | | | | | | | | | |
| AR_cp1256 | | | | | | 10,212 | 104,144 | 1,021,136 | 10,231,081 | 102,448,339 | | | | | | | | | | |
| RU_cp1251 | | | | | | 13,763 | 136,674 | 1,339,286 | 13,369,690 | 133,671,322 | | | | | | | | | | |
| **OOV type rate (%)** | | | | | | | | | | | | | | | | | | | | |
| AR | 28.46 | 5.38 | 0.00 | 0.00 | 0.00 | 26.02 | 4.88 | 0.00 | 0.00 | 0.00 | 93.72 | 71.13 | 35.70 | 13.44 | 5.02 | 6.06 | 0.34 | 0.00 | 0.00 | 0.00 |
| EN | 21.65 | 12.37 | 8.25 | 2.06 | 0.00 | 22.55 | 11.76 | 4.90 | 0.98 | 0.00 | 89.47 | 60.56 | 26.27 | 10.43 | 4.32 | 3.52 | 0.49 | 0.18 | 0.04 | 0.00 |
| ES | 25.23 | 14.41 | 10.81 | 5.41 | 0.90 | 23.64 | 10.91 | 6.36 | 0.91 | 0.00 | 90.71 | 63.96 | 27.18 | 9.71 | 3.88 | 4.44 | 0.53 | 0.19 | 0.08 | 0.00 |
| FR | 23.01 | 9.73 | 7.08 | 1.77 | 0.00 | 22.88 | 9.32 | 5.08 | 0.85 | 0.00 | 89.92 | 62.27 | 25.96 | 8.89 | 3.48 | 3.96 | 0.49 | 0.15 | 0.02 | 0.00 |
| RU | 33.56 | 7.53 | 2.05 | 0.00 | 0.00 | 31.25 | 5.56 | 1.39 | 0.00 | 0.00 | 93.11 | 69.66 | 32.15 | 10.91 | 3.81 | 4.82 | 0.29 | 0.02 | 0.01 | 0.00 |
| ZH | 61.23 | 26.87 | 6.73 | 1.42 | 0.35 | 16.99 | 3.92 | 0.65 | 0.00 | 0.00 | 90.34 | 63.48 | 29.89 | 10.31 | 4.09 | 67.98 | 26.22 | 5.03 | 0.55 | 0.10 |
| ZH_pinyin | 27.55 | 8.16 | 1.02 | 0.00 | 0.00 | | | | | | | | | | | | | | | |
| ZH_wubi | 23.33 | 4.17 | 0.83 | 0.00 | 0.00 | | | | | | | | | | | | | | | |
| AR_cp1256 | | | | | | 28.46 | 5.38 | 0.00 | 0.00 | 0.00 | | | | | | | | | | |
| RU_cp1251 | | | | | | 33.56 | 7.53 | 2.05 | 0.00 | 0.00 | | | | | | | | | | |
| **OOV token rate (%)** | | | | | | | | | | | | | | | | | | | | |
| AR | 0.13 | 0.01 | 0.00 | 0.00 | 0.00 | 0.05 | 0.00 | 0.00 | 0.00 | 0.00 | 53.40 | 25.24 | 9.31 | 3.25 | 1.18 | 0.73 | 0.12 | 0.00 | 0.00 | 0.00 |
| EN | 0.03 | 0.00 | 0.00 | 0.00 | 0.00 | 0.03 | 0.00 | 0.00 | 0.00 | 0.00 | 34.26 | 11.63 | 3.41 | 1.26 | 0.51 | 0.20 | 0.03 | 0.02 | 0.00 | 0.00 |
| ES | 0.04 | 0.01 | 0.01 | 0.00 | 0.00 | 0.05 | 0.01 | 0.01 | 0.00 | 0.00 | 32.17 | 12.05 | 3.56 | 1.16 | 0.46 | 0.29 | 0.09 | 0.04 | 0.02 | 0.00 |
| FR | 0.14 | 0.00 | 0.00 | 0.00 | 0.00 | 0.12 | 0.00 | 0.00 | 0.00 | 0.00 | 31.39 | 11.66 | 3.30 | 1.03 | 0.41 | 0.51 | 0.04 | 0.02 | 0.00 | 0.00 |
| RU | 0.15 | 0.00 | 0.00 | 0.00 | 0.00 | 0.08 | 0.00 | 0.00 | 0.00 | 0.00 | 49.48 | 22.44 | 7.53 | 2.33 | 0.79 | 0.38 | 0.04 | 0.00 | 0.00 | 0.00 |
| ZH | 11.15 | 1.25 | 0.18 | 0.04 | 0.01 | 0.10 | 0.00 | 0.00 | 0.00 | 0.00 | 38.65 | 13.96 | 4.52 | 1.42 | 0.56 | 22.01 | 4.75 | 0.81 | 0.09 | 0.01 |
| ZH_pinyin | 0.07 | 0.00 | 0.00 | 0.00 | 0.00 | | | | | | | | | | | | | | | |
| ZH_wubi | 0.09 | 0.00 | 0.00 | 0.00 | 0.00 | | | | | | | | | | | | | | | |
| AR_cp1256 | | | | | | 0.13 | 0.01 | 0.00 | 0.00 | 0.00 | | | | | | | | | | |
| RU_cp1251 | | | | | | 0.15 | 0.00 | 0.00 | 0.00 | 0.00 | | | | | | | | | | |
| **TTR (%)** | | | | | | | | | | | | | | | | | | | | |
| AR | 0.95 | 0.13 | 0.02 | 0.00 | 0.00 | 0.50 | 0.07 | 0.01 | 0.00 | 0.00 | 61.67 | 35.39 | 14.84 | 5.09 | 1.62 | 24.56 | 17.69 | 8.89 | 1.42 | 0.14 |
| EN | 0.63 | 0.08 | 0.01 | 0.00 | 0.00 | 0.66 | 0.08 | 0.01 | 0.00 | 0.00 | 42.82 | 19.37 | 6.44 | 2.01 | 0.66 | 23.81 | 13.87 | 4.88 | 1.24 | 0.13 |
| ES | 0.59 | 0.08 | 0.01 | 0.00 | 0.00 | 0.59 | 0.08 | 0.01 | 0.00 | 0.00 | 39.24 | 19.24 | 6.94 | 2.11 | 0.64 | 21.22 | 12.83 | 4.91 | 1.13 | 0.12 |
| FR | 0.63 | 0.08 | 0.01 | 0.00 | 0.00 | 0.64 | 0.08 | 0.01 | 0.00 | 0.00 | 40.40 | 18.45 | 6.27 | 1.82 | 0.53 | 22.72 | 12.78 | 4.71 | 1.10 | 0.12 |
| RU | 0.71 | 0.10 | 0.01 | 0.00 | 0.00 | 0.39 | 0.06 | 0.01 | 0.00 | 0.00 | 57.56 | 31.70 | 12.95 | 4.17 | 1.24 | 26.28 | 18.15 | 8.66 | 1.35 | 0.14 |
| ZH | 23.23 | 4.50 | 0.72 | 0.10 | 0.01 | 1.39 | 0.16 | 0.02 | 0.00 | 0.00 | 48.36 | 21.76 | 7.73 | 2.45 | 0.73 | 42.36 | 16.84 | 5.62 | 1.41 | 0.15 |
| ZH_pinyin | 0.64 | 0.09 | 0.01 | 0.00 | 0.00 | | | | | | | | | | | | | | | |
| ZH_wubi | 0.94 | 0.12 | 0.02 | 0.00 | 0.00 | | | | | | | | | | | | | | | |
| AR_cp1256 | | | | | | 0.95 | 0.13 | 0.02 | 0.00 | 0.00 | | | | | | | | | | |
| RU_cp1251 | | | | | | 0.71 | 0.10 | 0.01 | 0.00 | 0.00 | | | | | | | | | | |
| **Mean line length±std 0/25/50/75/100-th** | | | | | | | | | | | | | | | | | | | | |
| AR | 102.12 ± 57.40<br>5/57/96/152/213 | 104.14 ± 57.83<br>4/54/106/149/264 | 102.11 ± 58.06<br>3/53/102/147/293 | 102.31 ± 58.33<br>1/53/102/147/298 | 102.45 ± 58.46<br>1/53/102/147/300 | 184.48 ± 105.32<br>9/101/173/280/383 | 188.51 ± 105.63<br>6/98/191/270/487 | 184.51 ± 105.78<br>3/94/184/266/533 | 184.83 ± 106.15<br>1/95/184/266/551 | 185.11 ± 106.40<br>1/95/185/266/559 | 18.89 ± 10.27<br>2/11/18/27/43 | 19.28 ± 10.59<br>1/10/20/27/50 | 19.07 ± 10.72<br>1/10/19/27/82 | 19.13 ± 10.82<br>1/10/19/27/120 | 19.14 ± 10.84<br>1/10/19/27/143 | 35.18 ± 18.92<br>4/18/34/50/73 | 25.60 ± 14.42<br>1/14/26/36/79 | 21.07 ± 12.01<br>1/11/21/30/86 | 20.87 ± 12.00<br>1/11/21/29/123 | 20.98 ± 12.10<br>1/11/21/29/147 |
| EN | 123.00 ± 72.30<br>9/63/116/187/268 | 123.65 ± 71.12<br>3/60/128/178/300 | 120.88 ± 69.62<br>1/61/122/175/300 | 120.89 ± 69.58<br>1/62/122/175/300 | 120.93 ± 69.60<br>1/61/122/175/300 | 123.05 ± 72.35<br>9/63/116/187/268 | 123.68 ± 71.15<br>3/60/128/178/300 | 120.92 ± 69.64<br>1/61/122/175/300 | 120.93 ± 69.60<br>1/62/122/175/300 | 120.97 ± 69.62<br>1/61/122/175/304 | 21.16 ± 12.24<br>2/10/21/31/48 | 21.55 ± 12.31<br>1/11/22/31/55 | 21.20 ± 12.15<br>1/11/21/30/83 | 21.23 ± 12.17<br>1/11/21/30/118 | 21.23 ± 12.17<br>1/11/21/30/136 | 33.69 ± 19.32<br>3/15/33/49/78 | 25.11 ± 14.32<br>1/13/26/36/86 | 22.13 ± 12.67<br>1/12/22/32/88 | 21.60 ± 12.36<br>1/12/22/31/120 | 21.67 ± 12.42<br>1/12/22/31/138 |
| ES | 141.25 ± 82.83<br>11/59/134/215/292 | 139.56 ± 79.59<br>3/67/145/201/300 | 136.98 ± 78.35<br>1/70/138/200/300 | 137.01 ± 78.45<br>1/70/138/199/300 | 137.13 ± 78.47<br>1/70/139/199/300 | 143.87 ± 84.37<br>11/60/136/218/298 | 142.24 ± 81.05<br>3/69/149/206/306 | 139.53 ± 79.72<br>1/71/140/204/308 | 139.58 ± 79.84<br>1/71/141/203/312 | 139.70 ± 79.87<br>1/71/141/203/316 | 24.72 ± 14.39<br>2/12/24/37/56 | 24.60 ± 13.91<br>1/12/25/36/60 | 24.27 ± 13.80<br>1/13/24/35/84 | 24.31 ± 13.87<br>1/13/24/35/119 | 24.32 ± 13.87<br>1/13/24/35/139 | 37.84 ± 21.74<br>4/17/35/56/86 | 28.73 ± 16.27<br>1/14/30/41/90 | 25.41 ± 14.46<br>1/13/25/36/90 | 24.84 ± 14.17<br>1/13/25/36/121 | 24.93 ± 14.22<br>1/13/25/36/140 |
| FR | 137.97 ± 80.12<br>6/62/133/215/293 | 138.53 ± 79.49<br>3/68/141/203/300 | 136.54 ± 78.80<br>1/68/137/200/300 | 136.61 ± 78.92<br>1/69/138/199/300 | 136.75 ± 78.96<br>1/69/138/199/300 | 142.43 ± 82.97<br>6/64/137/219/303 | 143.14 ± 82.18<br>3/70/145/209/315 | 141.15 ± 81.47<br>1/71/142/206/319 | 141.20 ± 81.56<br>1/71/142/206/323 | 141.34 ± 81.61<br>1/71/143/206/330 | 24.53 ± 13.79<br>2/13/24/37/50 | 24.95 ± 14.19<br>1/12/26/36/62 | 24.85 ± 14.28<br>1/13/25/36/84 | 24.85 ± 14.31<br>1/13/25/36/118 | 24.86 ± 14.31<br>1/13/25/36/138 | 37.55 ± 20.97<br>3/20/37/56/80 | 28.91 ± 16.41<br>1/14/30/42/86 | 25.86 ± 14.85<br>1/13/26/37/88 | 25.28 ± 14.53<br>1/13/25/36/120 | 25.36 ± 14.58<br>1/13/26/36/138 |
| RU | 137.63 ± 80.80<br>11/71/130/212/296 | 136.67 ± 78.77<br>3/63/139/200/300 | 133.93 ± 77.25<br>1/67/135/196/300 | 133.70 ± 77.26<br>1/67/134/195/300 | 133.67 ± 77.25<br>1/67/134/195/300 | 253.45 ± 150.90<br>21/126/242/392/550 | 251.23 ± 146.21<br>4/118/258/366/561 | 245.79 ± 143.31<br>1/121/247/360/562 | 245.36 ± 143.18<br>1/122/247/358/566 | 245.34 ± 143.17<br>1/122/247/358/569 | 20.24 ± 11.27<br>2/11/20/29/47 | 20.47 ± 11.61<br>1/10/21/29/55 | 20.16 ± 11.44<br>1/11/20/29/83 | 20.16 ± 11.53<br>1/11/20/29/117 | 20.16 ± 11.53<br>1/11/20/29/136 | 38.13 ± 21.10<br>3/18/38/56/81 | 26.77 ± 15.26<br>1/14/27/39/78 | 21.98 ± 12.61<br>1/12/22/31/87 | 21.73 ± 12.59<br>1/11/22/31/121 | 21.79 ± 12.64<br>1/11/22/31/138 |
| ZH | 33.71 ± 18.03<br>4/19/33/48/70 | 34.43 ± 19.40<br>2/18/35/49/127 | 34.15 ± 19.06<br>1/19/34/48/133 | 34.20 ± 19.27<br>1/19/34/48/220 | 34.20 ± 19.28<br>1/19/34/48/272 | 92.35 ± 52.42<br>12/47/93/132/198 | 95.10 ± 54.29<br>3/50/96/138/248 | 93.88 ± 53.42<br>3/49/94/135/291 | 94.04 ± 53.75<br>1/49/94/135/378 | 94.05 ± 53.79<br>1/48/94/135/594 | 18.24 ± 9.97<br>2/10/17/26/39 | 19.12 ± 10.87<br>1/10/20/27/55 | 19.05 ± 10.86<br>1/10/19/27/87 | 19.08 ± 10.91<br>1/10/19/27/113 | 19.07 ± 10.92<br>1/10/19/27/248 | 25.59 ± 13.84<br>2/14/26/37/52 | 22.07 ± 12.68<br>1/12/22/31/90 | 19.97 ± 11.43<br>1/11/20/28/92 | 19.47 ± 11.17<br>1/11/19/27/115 | 19.57 ± 11.26<br>1/11/19/28/249 |
| ZH_pinyin | 116.05 ± 67.54<br>11/58/117/167/263 | 119.26 ± 68.54<br>3/61/122/173/320 | 117.16 ± 67.25<br>3/61/117/169/335 | 117.32 ± 67.49<br>1/60/118/169/376 | 117.36 ± 67.55<br>1/60/118/169/584 | | | | | | | | | | | | | | | |
| ZH_wubi | 100.71 ± 57.05<br>13/54/101/139/217 | 103.94 ± 59.48<br>3/54/105/150/296 | 102.29 ± 57.92<br>3/54/102/147/297 | 102.46 ± 58.29<br>1/53/102/146/398 | 102.45 ± 58.31<br>1/53/103/146/627 | | | | | | | | | | | | | | | |
| AR_cp1256 | | | | | | 102.12 ± 57.40<br>5/57/96/152/213 | 104.14 ± 57.83<br>4/54/106/149/264 | 102.11 ± 58.06<br>3/53/102/147/293 | 102.31 ± 58.33<br>1/53/102/147/298 | 102.45 ± 58.46<br>1/53/102/147/300 | | | | | | | | | | |
| RU_cp1251 | | | | | | 137.63 ± 80.80<br>11/71/130/212/296 | 136.67 ± 78.77<br>3/63/139/200/300 | 133.93 ± 77.25<br>1/67/135/196/300 | 133.70 ± 77.26<br>1/67/134/195/300 | 133.67 ± 77.25<br>1/67/134/195/300 | | | | | | | | | | |

# Statistics for development (dev) set

As a different set of vocabulary is learned from each training dataset and data size, BPE has a distinct dev set for each.

| Representation | CHAR | BYTE | WORD | | | BPE_A | | | | | BPE_B | | | | | BPE_C | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Number of lines in train set | | | | 100 | 1, 000 | 10, 000 | 100, 000 | 1, 000, 000 | 100 | 1, 000 | 10, 000 | 100, 000 | 1, 000, 000 | 100 | 1, 000 | 10, 000 | 100, 000 | 1, 000, 000 |
| **Number of TYPES** | | | | | | | | | | | | | | | | | | |
| AR | 130 | 123 | 13, 836 | 708 | 3, 232 | 8, 994 | 11, 473 | 12, 430 | 865 | 3, 991 | 10, 735 | 12, 954 | 13, 090 | 908 | 4, 119 | 10, 681 | 12, 934 | 13, 093 |
| EN | 97 | 102 | 7, 199 | 656 | 2, 567 | 5, 998 | 7, 461 | 7, 556 | 801 | 3, 149 | 6, 518 | 7, 557 | 7, 578 | 823 | 3, 245 | 6, 551 | 7, 524 | 7, 573 |
| ES | 111 | 110 | 8, 551 | 680 | 2, 832 | 6, 919 | 8, 752 | 8, 871 | 833 | 3, 426 | 7, 528 | 8, 869 | 8, 956 | 834 | 3, 424 | 7, 527 | 8, 871 | 8, 939 |
| FR | 113 | 118 | 8, 312 | 744 | 2, 821 | 6, 925 | 8, 526 | 8, 666 | 827 | 3, 407 | 7, 489 | 8, 665 | 8, 705 | 883 | 3, 463 | 7, 431 | 8, 679 | 8, 708 |
| RU | 146 | 144 | 12, 819 | 878 | 3, 769 | 10, 085 | 12, 505 | 12, 788 | 1, 024 | 4, 438 | 11, 052 | 12, 883 | 12, 954 | 1, 038 | 4, 509 | 11, 012 | 12, 882 | 12, 958 |
| ZH | 1, 976 | 153 | 7, 413 | 3, 224 | 4, 215 | 6, 386 | 7, 541 | 7, 654 | 3, 261 | 4, 481 | 6, 735 | 7, 719 | 7, 702 | 3, 260 | 4, 524 | 6, 785 | 7, 654 | 7, 721 |
| ZH_pinyin | 98 | | | | | | | | | | | | | | | | | |
| ZH_wubi | 120 | | | | | | | | | | | | | | | | | |
| AR_cp1256 | | 130 | | | | | | | | | | | | | | | | |
| RU_cp1251 | | 146 | | | | | | | | | | | | | | | | |
| **Number of TOKENS** | | | | | | | | | | | | | | | | | | |
| AR | 334, 358 | 605, 516 | 61, 371 | 167, 574 | 115, 693 | 83, 001 | 76, 284 | 70, 527 | 149, 689 | 97, 843 | 73, 314 | 68, 538 | 68, 270 | 149, 623 | 97, 231 | 73, 541 | 68, 579 | 68, 278 |
| EN | 391, 222 | 391, 260 | 67, 629 | 155, 826 | 101, 782 | 77, 089 | 71, 276 | 70, 339 | 140, 256 | 90, 871 | 73, 531 | 69, 633 | 69, 348 | 140, 377 | 90, 360 | 73, 468 | 69, 633 | 69, 341 |
| ES | 443, 958 | 452, 190 | 78, 087 | 170, 433 | 113, 983 | 88, 634 | 82, 281 | 81, 341 | 155, 687 | 103, 748 | 84, 914 | 80, 569 | 80, 337 | 154, 746 | 103, 172 | 84, 870 | 80, 579 | 80, 371 |
| FR | 438, 083 | 452, 556 | 78, 745 | 166, 280 | 114, 694 | 88, 726 | 82, 448 | 81, 559 | 156, 256 | 104, 071 | 85, 115 | 80, 844 | 80, 671 | 153, 745 | 104, 091 | 85, 125 | 80, 912 | 80, 654 |
| RU | 431, 538 | 793, 214 | 64, 180 | 177, 818 | 113, 628 | 79, 763 | 72, 480 | 71, 081 | 163, 319 | 100, 294 | 75, 051 | 70, 200 | 69, 991 | 163, 806 | 99, 978 | 75, 096 | 70, 196 | 69, 982 |
| ZH | 107, 990 | 301, 085 | 60, 013 | 96, 745 | 80, 231 | 68, 129 | 63, 614 | 62, 810 | 93, 775 | 75, 636 | 65, 212 | 61, 867 | 61, 832 | 94, 127 | 75, 718 | 65, 203 | 61, 916 | 61, 823 |
| ZH_pinyin | 376, 979 | | | | | | | | | | | | | | | | | |
| ZH_wubi | 330, 734 | | | | | | | | | | | | | | | | | |
| AR_cp1256 | | 334, 358 | | | | | | | | | | | | | | | | |
| RU_cp1251 | | 431, 538 | | | | | | | | | | | | | | | | |
| **TTR (%)** | | | | | | | | | | | | | | | | | | |
| AR | 0.04 | 0.02 | 22.54 | 0.42 | 2.79 | 10.84 | 15.04 | 17.62 | 0.58 | 4.08 | 14.64 | 18.90 | 19.17 | 0.61 | 4.24 | 14.52 | 18.86 | 19.18 |
| EN | 0.02 | 0.03 | 10.64 | 0.42 | 2.52 | 7.78 | 10.47 | 10.74 | 0.57 | 3.47 | 8.86 | 10.85 | 10.93 | 0.59 | 3.59 | 8.92 | 10.81 | 10.92 |
| ES | 0.03 | 0.02 | 10.95 | 0.40 | 2.48 | 7.81 | 10.64 | 10.91 | 0.54 | 3.30 | 8.87 | 11.01 | 11.15 | 0.54 | 3.32 | 8.87 | 11.01 | 11.12 |
| FR | 0.03 | 0.03 | 10.56 | 0.45 | 2.46 | 7.80 | 10.34 | 10.63 | 0.53 | 3.27 | 8.80 | 10.72 | 10.79 | 0.57 | 3.33 | 8.73 | 10.73 | 10.80 |
| RU | 0.03 | 0.02 | 19.97 | 0.49 | 3.32 | 12.64 | 17.25 | 17.99 | 0.63 | 4.42 | 14.73 | 18.35 | 18.51 | 0.63 | 4.51 | 14.66 | 18.35 | 18.52 |
| ZH | 1.83 | 0.05 | 12.35 | 3.33 | 5.25 | 9.37 | 11.85 | 12.19 | 3.48 | 5.92 | 10.33 | 12.48 | 12.46 | 3.46 | 5.97 | 10.41 | 12.36 | 12.49 |
| ZH_pinyin | 0.03 | | | | | | | | | | | | | | | | | |
| ZH_wubi | 0.04 | | | | | | | | | | | | | | | | | |
| AR_cp1256 | | 0.04 | | | | | | | | | | | | | | | | |
| RU_cp1251 | | 0.03 | | | | | | | | | | | | | | | | |
| **Mean line length±std 0/25/50/75/100-th** | | | | | | | | | | | | | | | | | | |
| AR | 108.66 ± 58.01<br>3/60/110/153/277 | 196.79 ± 105.85<br>6/107/199/277/503 | 19.95 ± 10.58<br>1/12/20/27/58 | 54.46 ± 29.51<br>1/30/54/76/152 | 37.60 ± 20.51<br>1/22/37/52/125 | 26.97 ± 14.80<br>1/16/27/37/95 | 24.79 ± 13.59<br>1/14/24/34/86 | 22.92 ± 12.56<br>1/14/22/32/75 | 48.65 ± 26.35<br>1/27/49/68/156 | 31.80 ± 17.76<br>1/18/32/44/110 | 23.83 ± 13.15<br>1/14/24/33/80 | 22.27 ± 12.12<br>1/13/22/31/71 | 22.19 ± 12.07<br>1/13/22/31/71 | 48.63 ± 26.21<br>1/28/49/67/145 | 31.60 ± 17.53<br>1/18/31/44/110 | 23.90 ± 13.20<br>1/14/24/33/87 | 22.29 ± 12.14<br>1/13/22/31/73 | 22.19 ± 12.06<br>1/13/22/31/72 |
| EN | 127.14 ± 68.64<br>6/68/130/181/299 | 127.16 ± 68.65<br>6/68/130/181/299 | 21.98 ± 11.81<br>1/13/22/31/61 | 50.64 ± 27.44<br>1/27/51/72/136 | 33.08 ± 18.07<br>1/19/33/46/110 | 25.05 ± 13.56<br>1/15/25/35/78 | 23.16 ± 12.42<br>1/14/23/32/71 | 22.86 ± 12.27<br>1/13/23/32/71 | 45.58 ± 24.86<br>1/26/46/65/129 | 29.53 ± 16.30<br>1/17/29/41/94 | 23.90 ± 12.95<br>1/14/24/33/74 | 22.63 ± 12.13<br>1/13/23/31/65 | 22.54 ± 12.06<br>1/13/23/31/65 | 45.62 ± 24.67<br>1/26/46/64/122 | 29.37 ± 16.13<br>1/17/29/41/92 | 23.88 ± 12.91<br>1/14/24/33/73 | 22.63 ± 12.12<br>1/13/23/31/65 | 22.54 ± 12.05<br>1/13/23/31/66 |
| ES | 144.28 ± 77.78<br>5/77/145/207/300 | 146.96 ± 79.21<br>5/78/148/211/307 | 25.38 ± 13.56<br>1/15/25/36/63 | 55.39 ± 30.13<br>1/30/56/79/141 | 37.04 ± 20.13<br>1/21/37/52/111 | 28.81 ± 15.53<br>1/17/29/41/85 | 26.74 ± 14.29<br>1/15/27/38/76 | 26.44 ± 14.14<br>1/15/26/37/73 | 50.60 ± 27.66<br>1/27/51/72/133 | 33.72 ± 18.48<br>1/19/34/47/103 | 27.60 ± 14.88<br>1/15/28/39/75 | 26.18 ± 13.98<br>1/15/26/37/71 | 26.11 ± 13.94<br>1/15/26/37/70 | 50.29 ± 27.27<br>1/28/51/71/133 | 33.53 ± 18.38<br>1/19/34/47/95 | 27.58 ± 14.86<br>1/15/28/39/75 | 26.19 ± 13.99<br>1/15/26/37/71 | 26.12 ± 13.94<br>1/15/26/37/71 |
| FR | 142.37 ± 77.81<br>4/74/145/205/300 | 147.08 ± 80.30<br>4/77/150/212/310 | 25.59 ± 13.86<br>1/14/26/36/66 | 54.04 ± 29.55<br>1/29/55/76/139 | 37.27 ± 20.65<br>1/20/38/52/119 | 28.84 ± 15.76<br>1/16/29/40/89 | 26.79 ± 14.47<br>1/15/27/37/74 | 26.51 ± 14.35<br>1/15/27/37/70 | 50.78 ± 27.79<br>1/27/52/72/135 | 33.82 ± 18.74<br>1/18/34/47/108 | 27.66 ± 15.11<br>1/15/28/39/80 | 26.27 ± 14.20<br>1/14/27/37/69 | 26.22 ± 14.16<br>1/14/27/37/69 | 49.97 ± 27.25<br>1/27/51/70/131 | 33.83 ± 18.80<br>1/18/34/47/107 | 27.66 ± 15.09<br>1/15/28/39/84 | 26.30 ± 14.21<br>1/14/27/37/69 | 26.21 ± 14.16<br>1/14/27/37/69 |
| RU | 140.25 ± 76.21<br>5/75/141/200/300 | 257.79 ± 141.53<br>7/136/259/370/569 | 20.86 ± 11.25<br>1/12/21/29/69 | 57.79 ± 32.12<br>1/31/58/82/185 | 36.93 ± 20.73<br>1/21/36/52/185 | 25.92 ± 14.48<br>1/15/26/36/128 | 23.56 ± 12.95<br>1/14/23/33/101 | 23.10 ± 12.71<br>1/13/23/32/95 | 53.08 ± 29.00<br>1/29/53/75/185 | 32.59 ± 18.42<br>1/18/32/46/166 | 24.39 ± 13.59<br>1/14/24/34/112 | 22.81 ± 12.50<br>1/13/23/32/85 | 22.75 ± 12.46<br>1/13/23/32/93 | 53.24 ± 29.18<br>2/30/53/76/185 | 32.49 ± 18.49<br>1/18/32/45/161 | 24.41 ± 13.57<br>1/14/24/34/105 | 22.81 ± 12.50<br>1/13/23/32/90 | 22.74 ± 12.45<br>1/13/23/32/92 |
| ZH | 35.10 ± 18.48<br>2/21/35/49/125 | 97.85 ± 52.10<br>4/53/99/138/268 | 19.50 ± 10.42<br>1/11/20/27/64 | 31.44 ± 16.96<br>1/18/31/44/109 | 26.07 ± 14.19<br>1/15/26/36/93 | 22.14 ± 12.02<br>1/13/22/31/80 | 20.67 ± 11.16<br>1/12/21/29/74 | 20.41 ± 11.05<br>1/12/20/28/74 | 30.48 ± 16.47<br>1/17/30/43/104 | 24.58 ± 13.42<br>1/14/24/34/92 | 21.19 ± 11.49<br>1/12/21/29/71 | 20.11 ± 10.79<br>1/12/20/28/64 | 20.09 ± 10.79<br>1/12/20/28/66 | 30.59 ± 16.45<br>1/18/30/43/101 | 24.61 ± 13.49<br>1/14/24/34/93 | 21.19 ± 11.49<br>1/12/21/29/76 | 20.12 ± 10.81<br>1/12/20/28/64 | 20.09 ± 10.79<br>1/12/20/28/65 |
| ZH_pinyin | 122.52 ± 65.59<br>6/67/125/173/353 | | | | | | | | | | | | | | | | | |
| ZH_wubi | 107.49 ± 56.73<br>4/60/108/151/294 | | | | | | | | | | | | | | | | | |
| AR_cp1256 | | 108.66 ± 58.01<br>3/60/110/153/277 | | | | | | | | | | | | | | | | |
| RU_cp1251 | | 140.25 ± 76.21<br>5/75/141/200/300 | | | | | | | | | | | | | | | | |

# F ENLARGED FIGURES FOR ALL 30 LANGUAGE DIRECTIONS (AGGREGATE RESULTS FROM ALL RUNS)

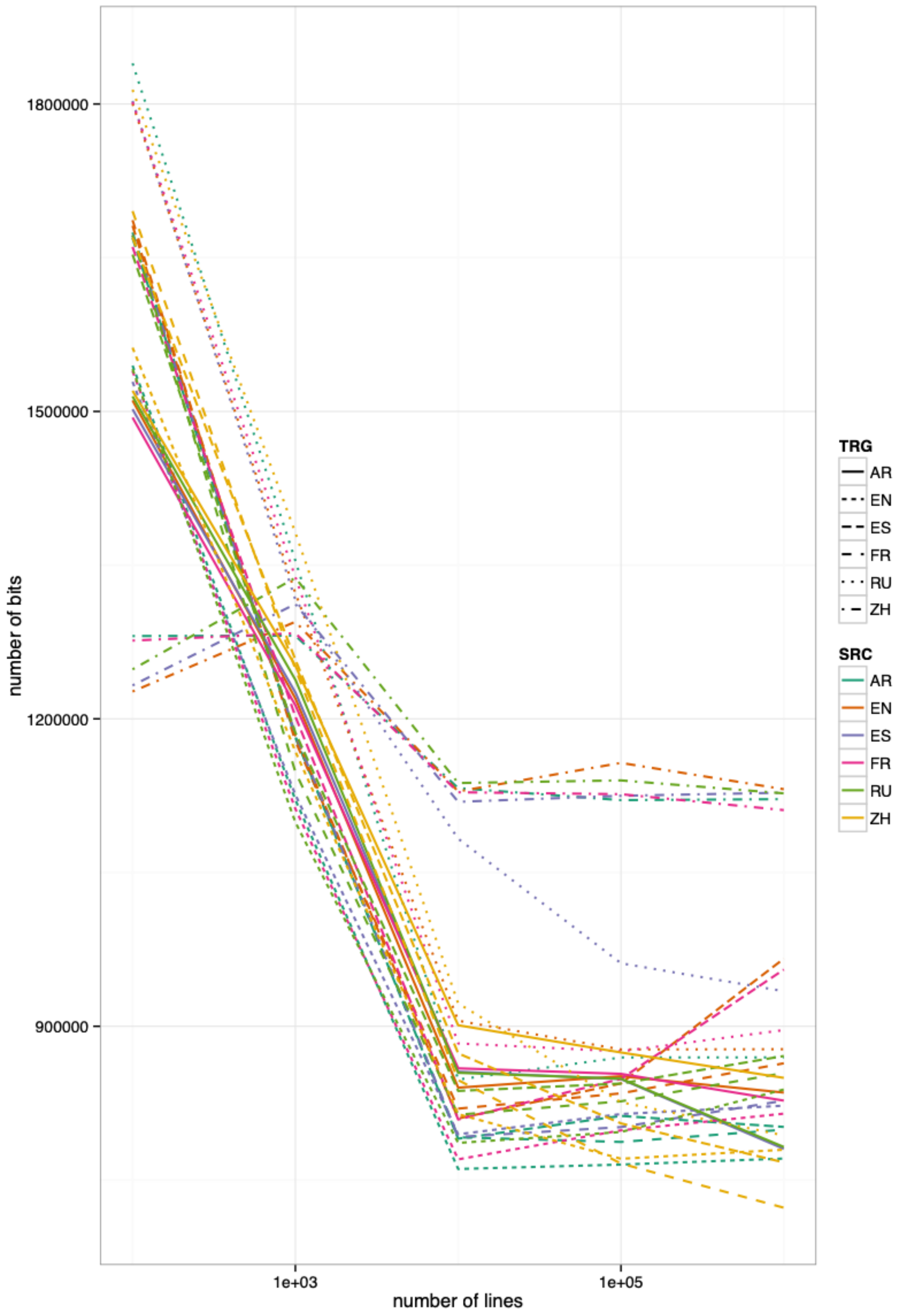


Figure 4: CHAR: character models

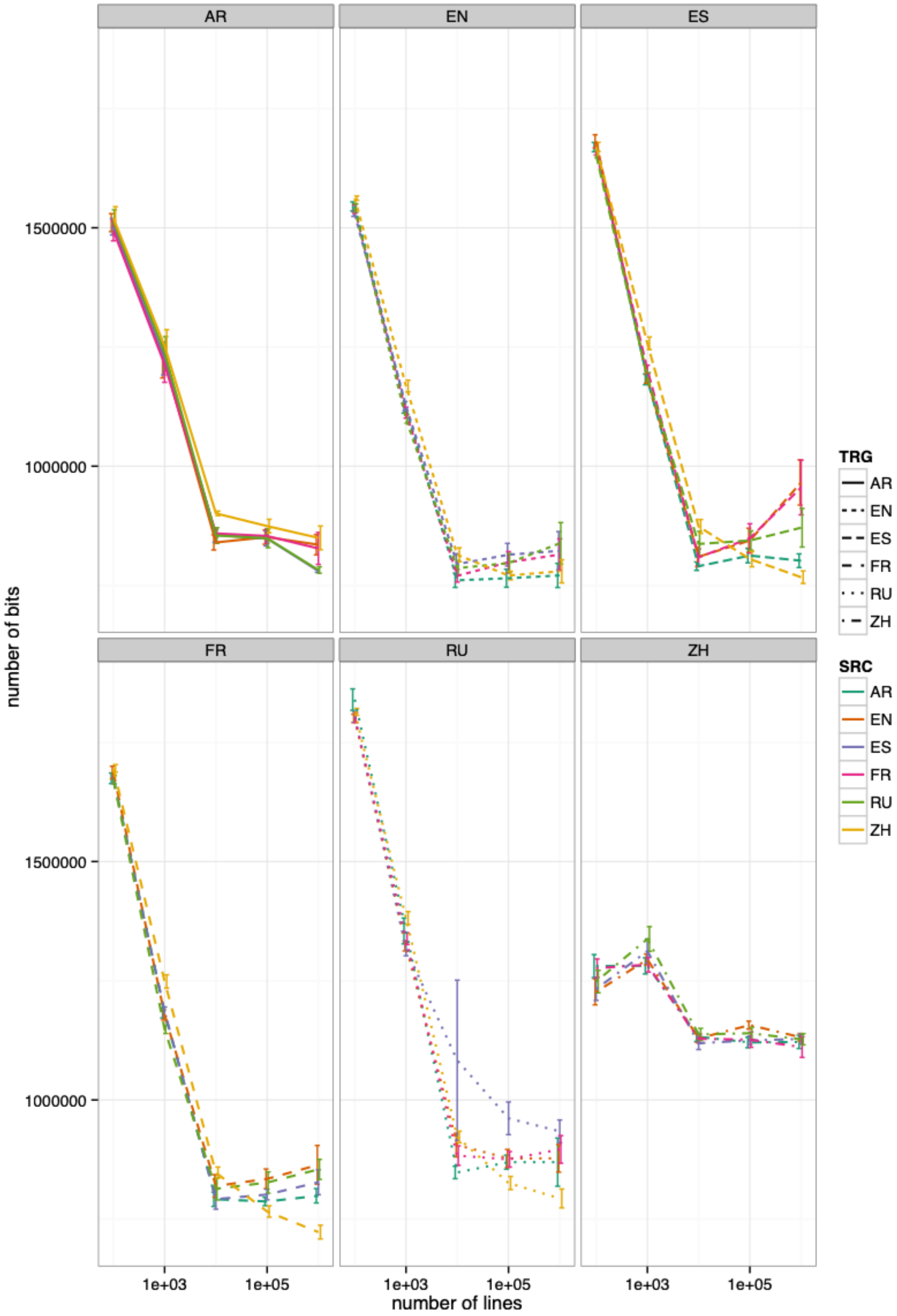


Figure 4: CHAR: character models (target language as facet)

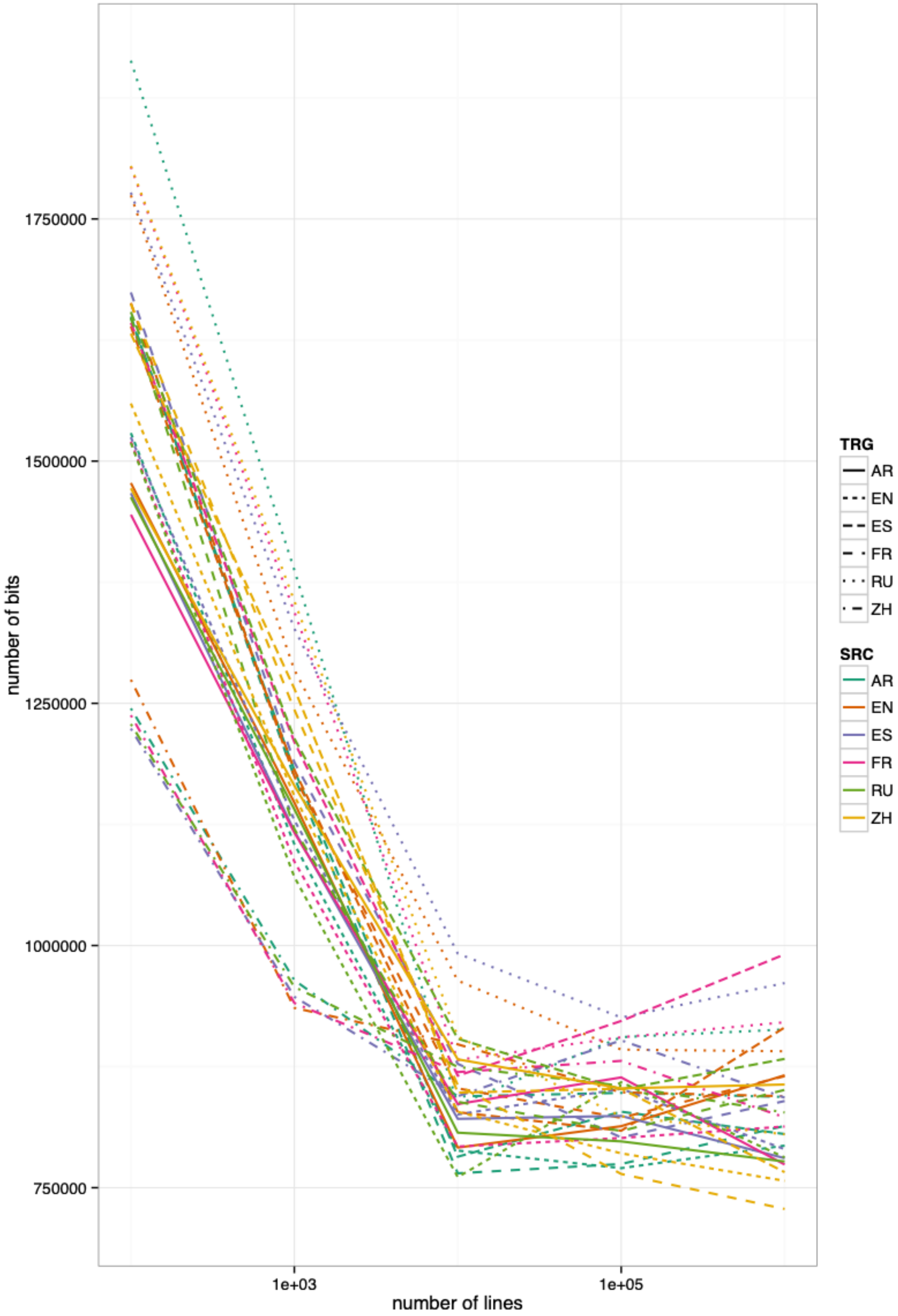


Figure 5: CHAR with Pinyin for $\mathrm{ZH}_{trg}$

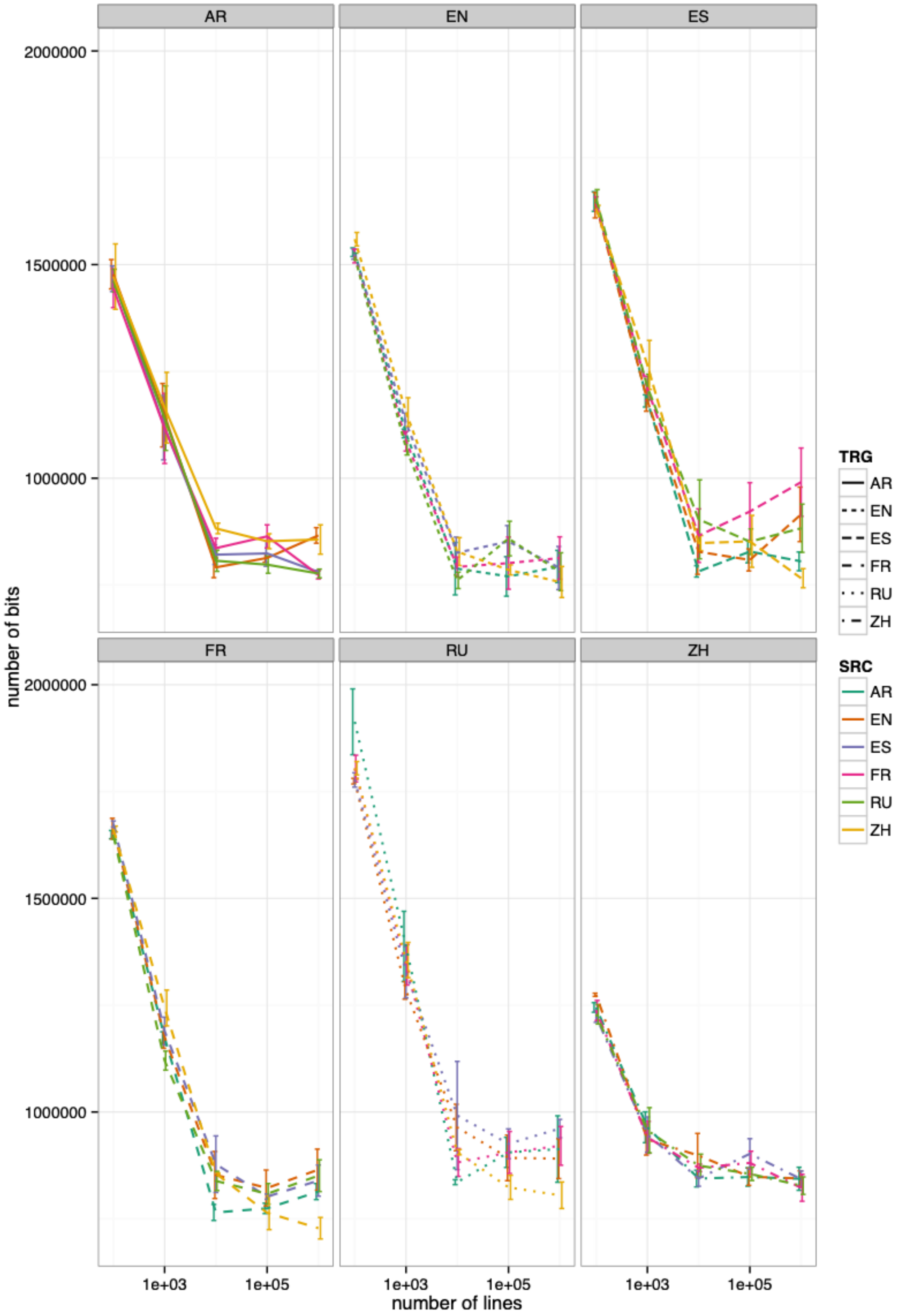


Figure 5: CHAR with Pinyin for $\mathrm{ZH}_{trg}$ (target language as facet)

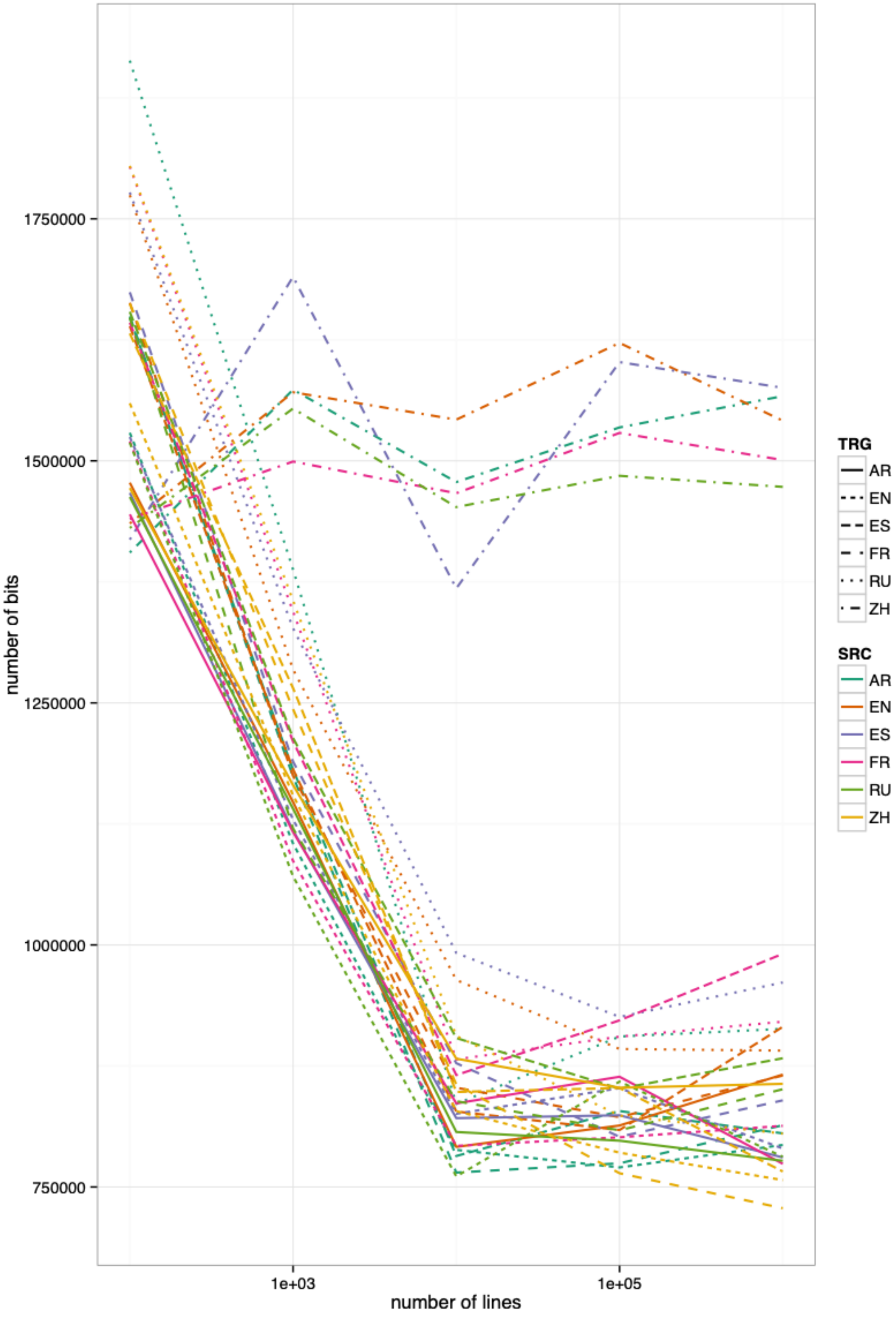


Figure 6: CHAR with Wubi for $\text{ZH}_{trg}$

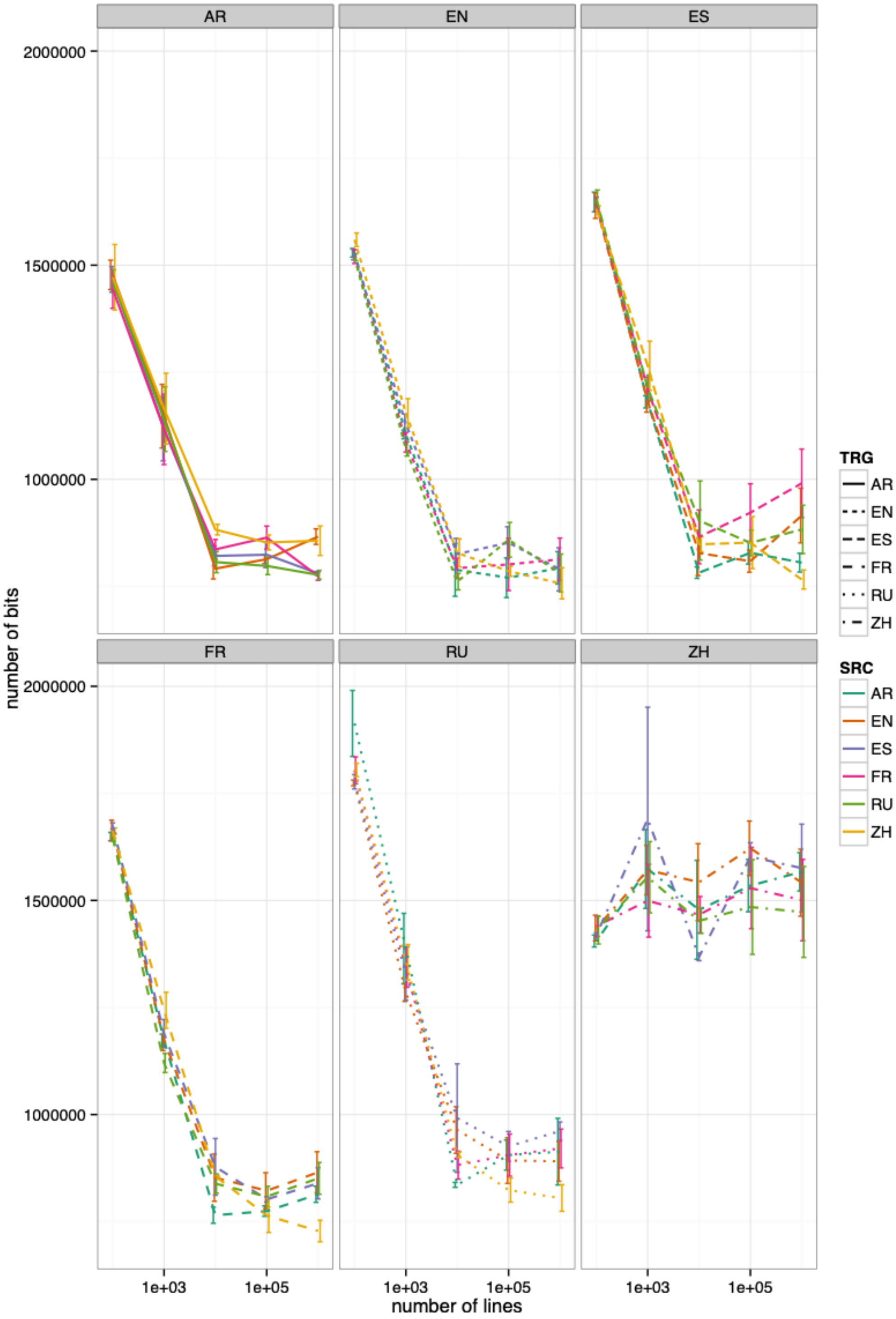


Figure 6: CHAR with Wubi for $\text{ZH}_{trg}$ (target language as facet)

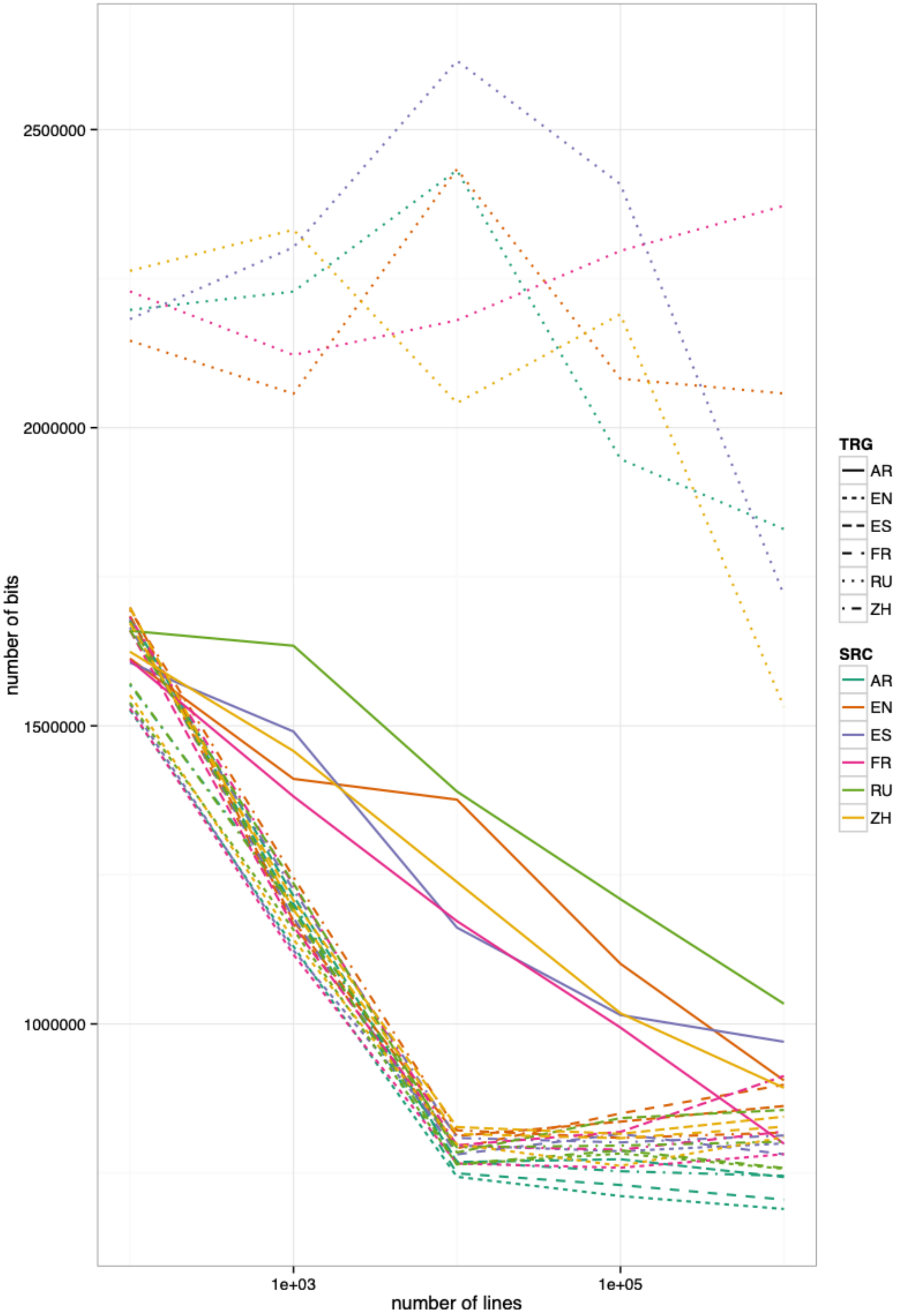


Figure 7: BYTE models with UTF-8 encoding

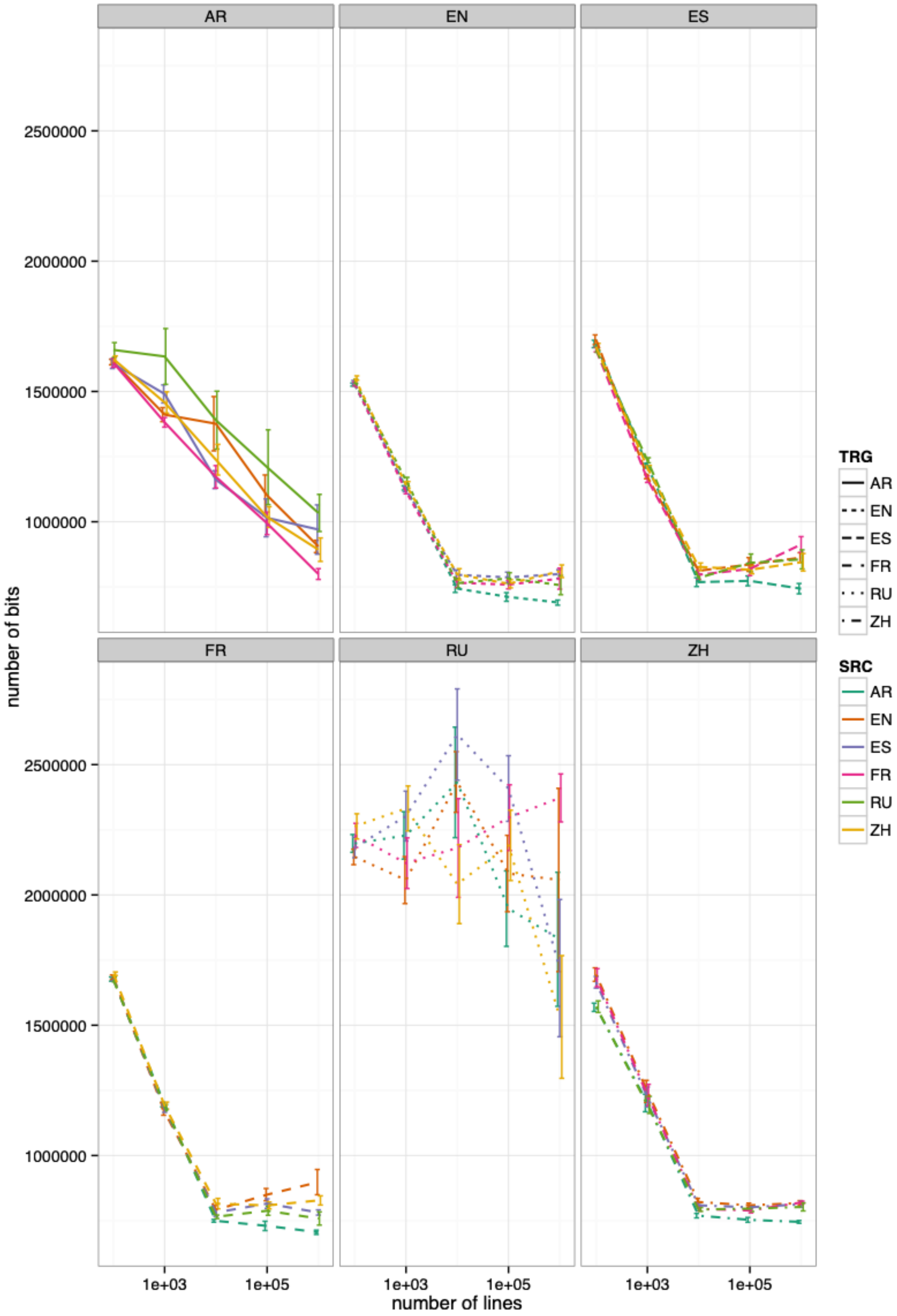


Figure 7: BYTE models with UTF-8 encoding (target language as facet)

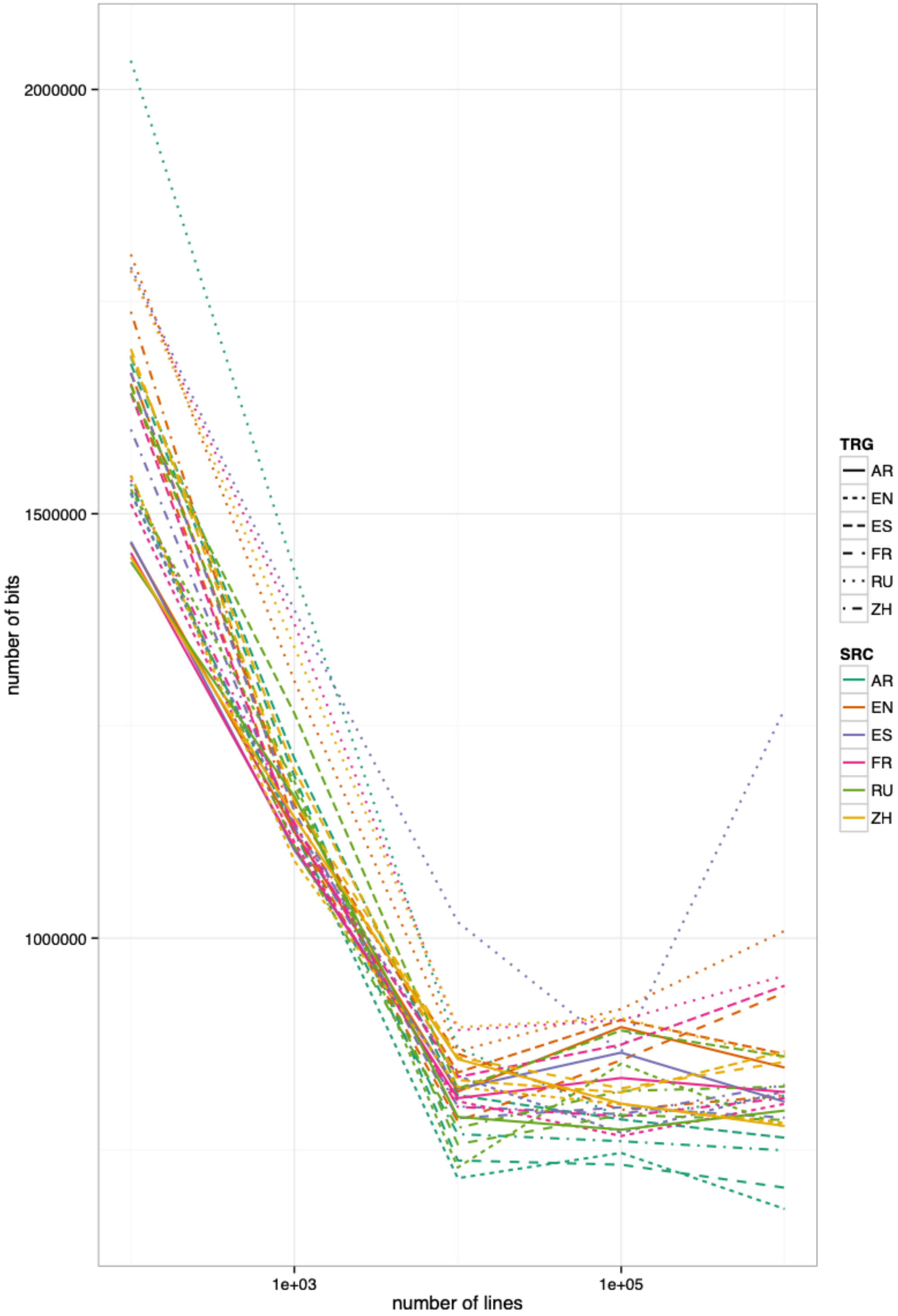


Figure 8: BYTE with $\mathrm{AR}_{trg}$ & $\mathrm{RU}_{trg}$ optimized with code pages 1256 & 1251 ($\mathrm{ARRU}_{trg}$)

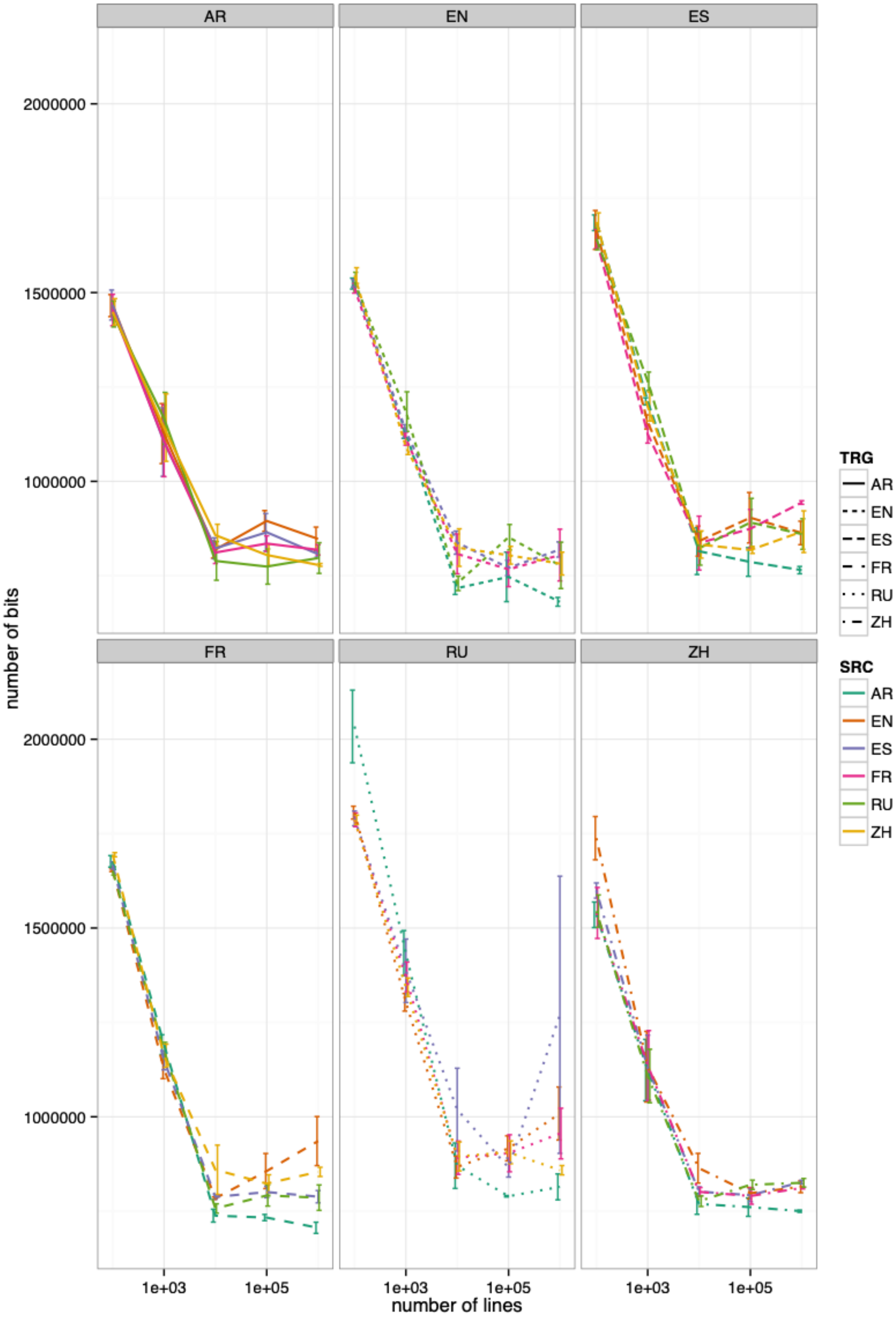


Figure 8: BYTE with $\text{AR}_{trg}$ & $\text{RU}_{trg}$ optimized with code pages 1256 & 1251 (target language as facet)

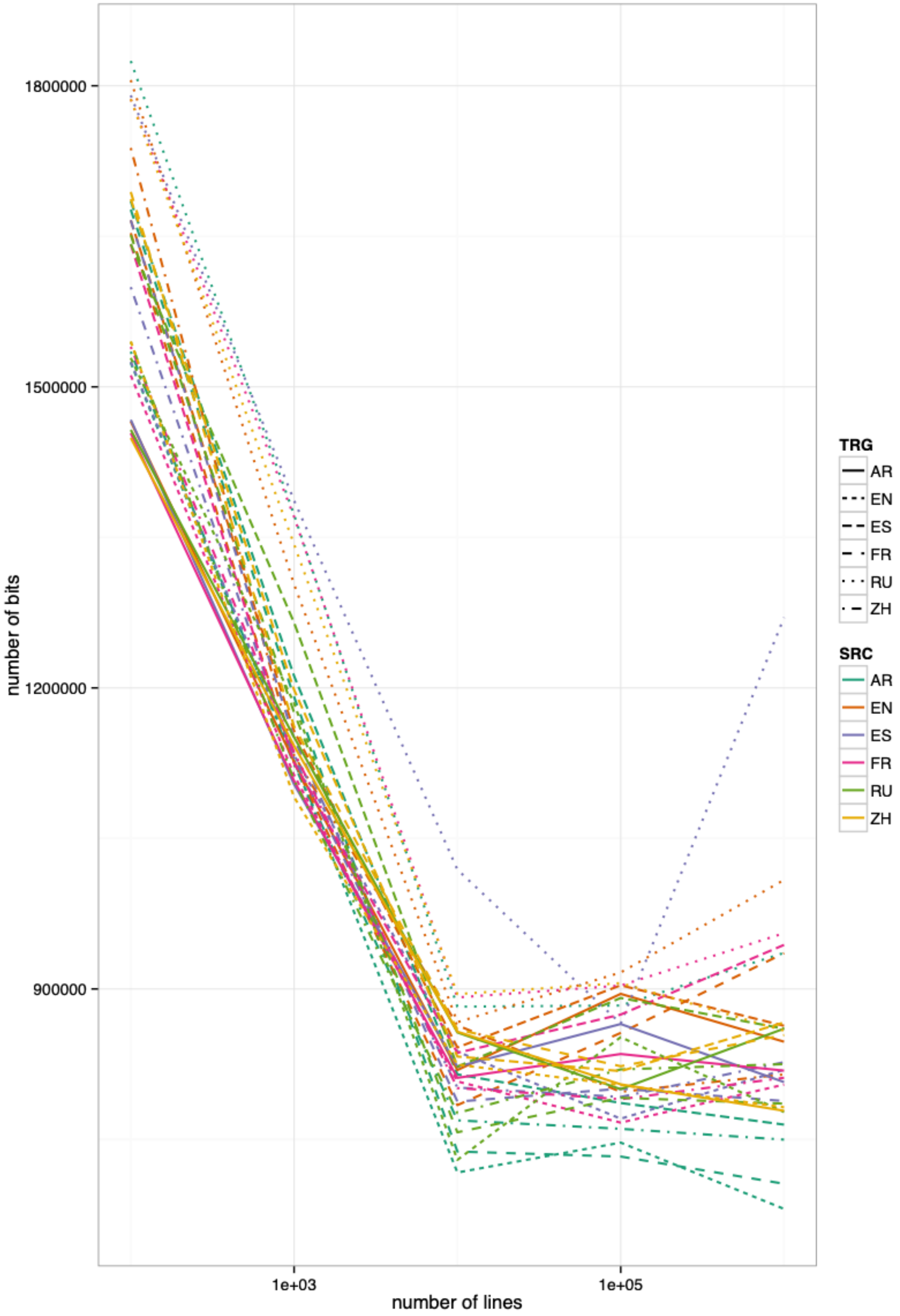


Figure 9: BYTE with directions AR-RU & RU-AR optimized on both source and target sides ($\text{ARRU}_{src,trg}$)

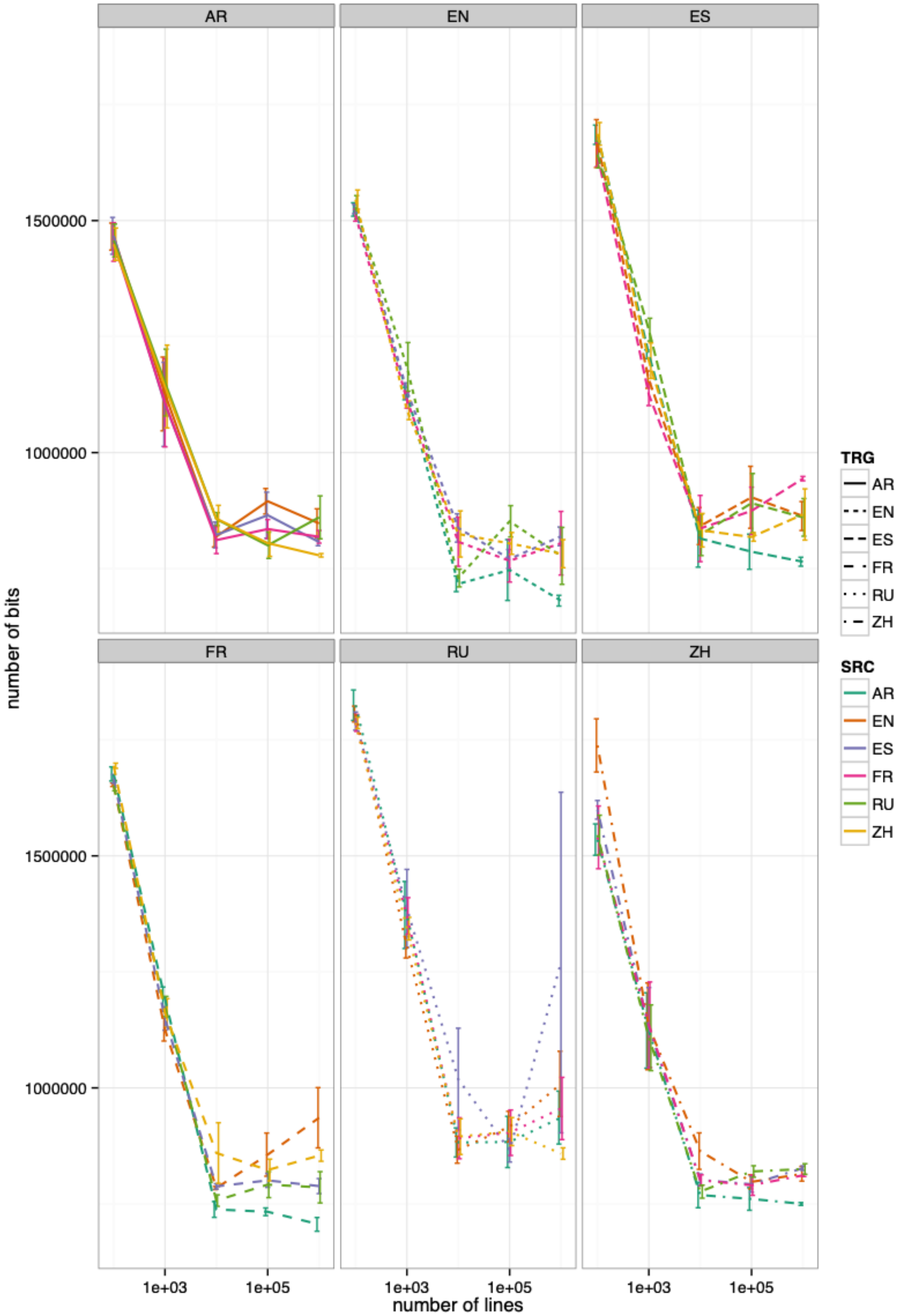


Figure 9: BYTE with directions AR-RU & RU-AR optimized on both source and target sides (target language as facet)

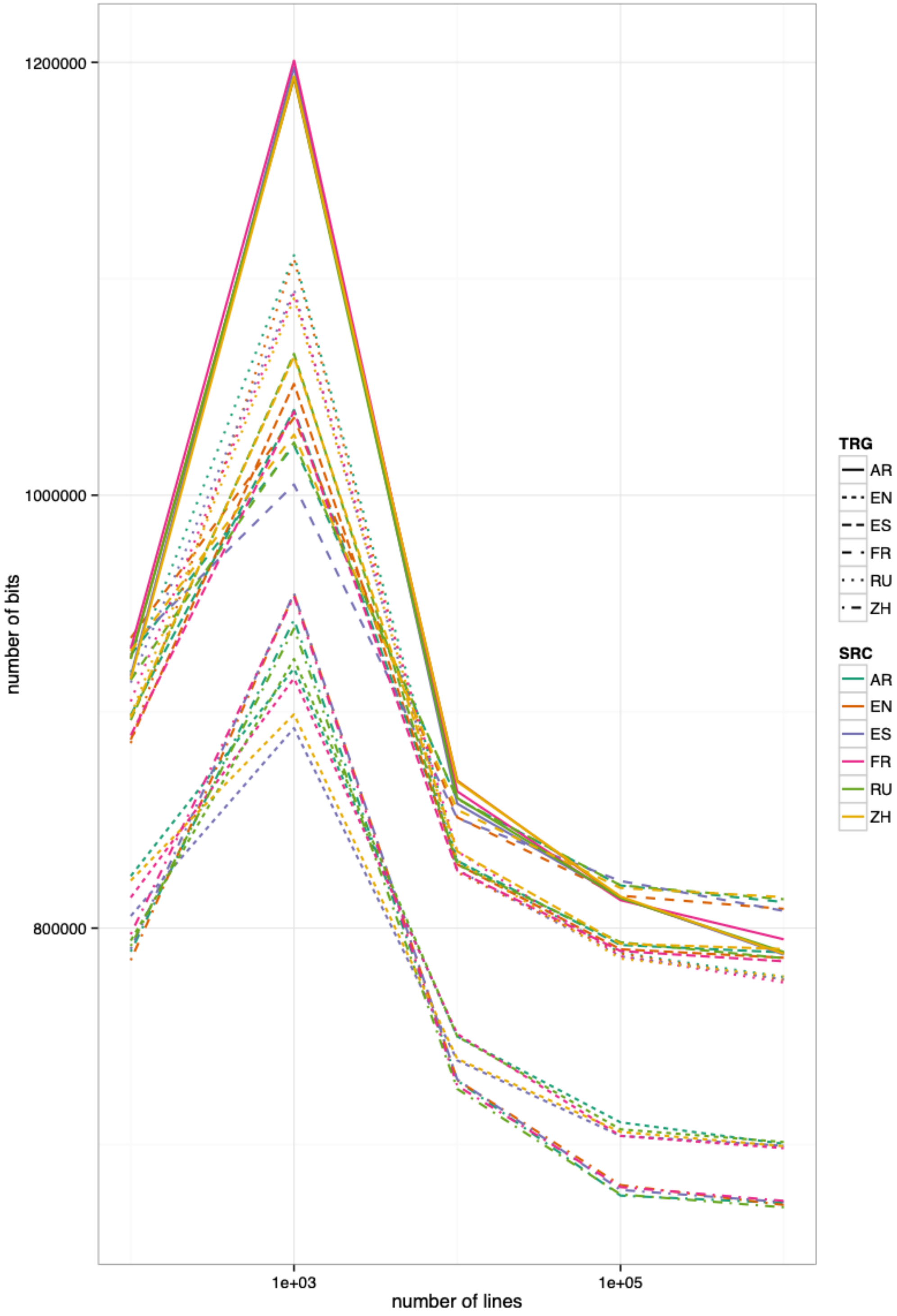


Figure 10: WORD models

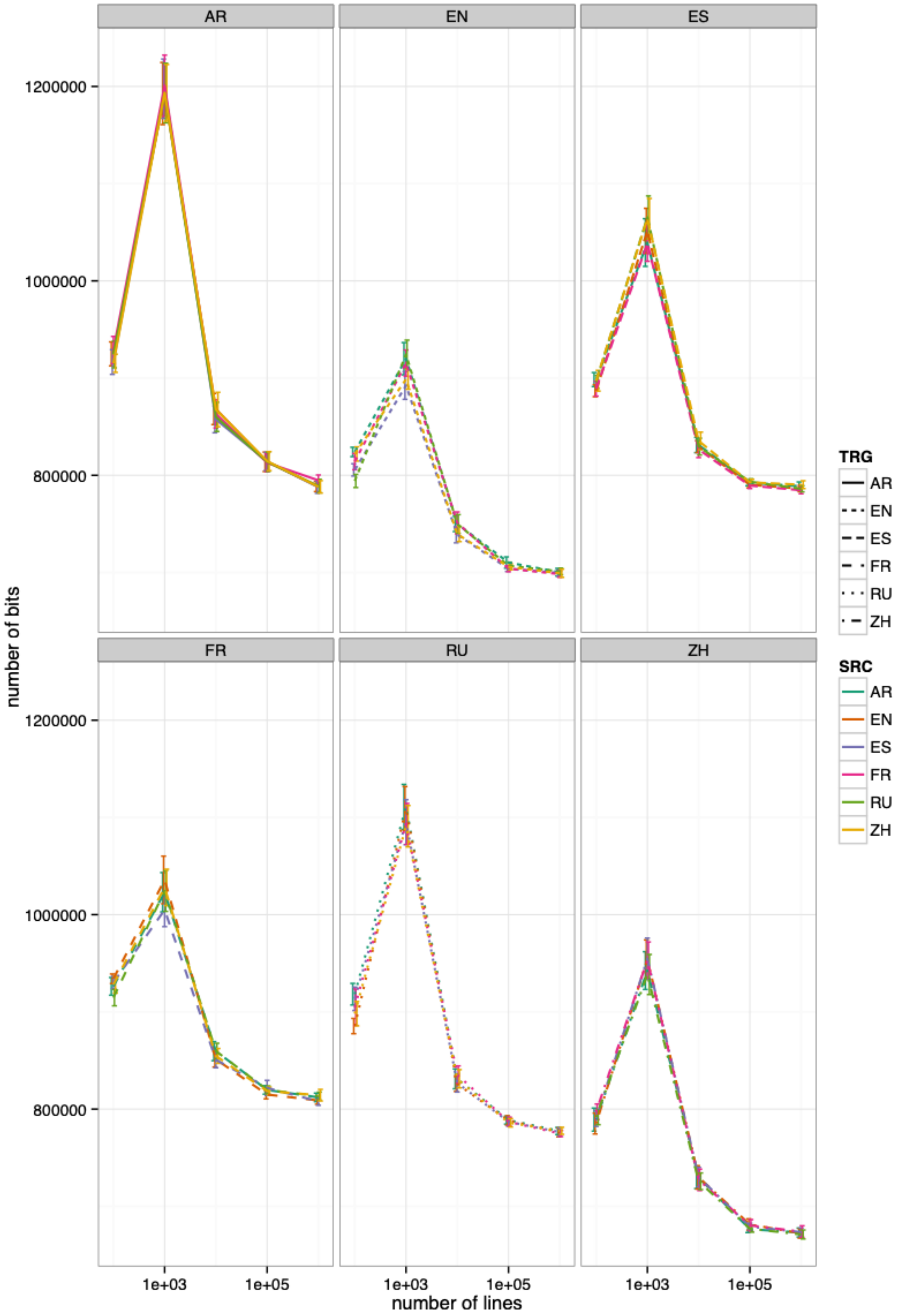


Figure 10: WORD models (target language as facet)

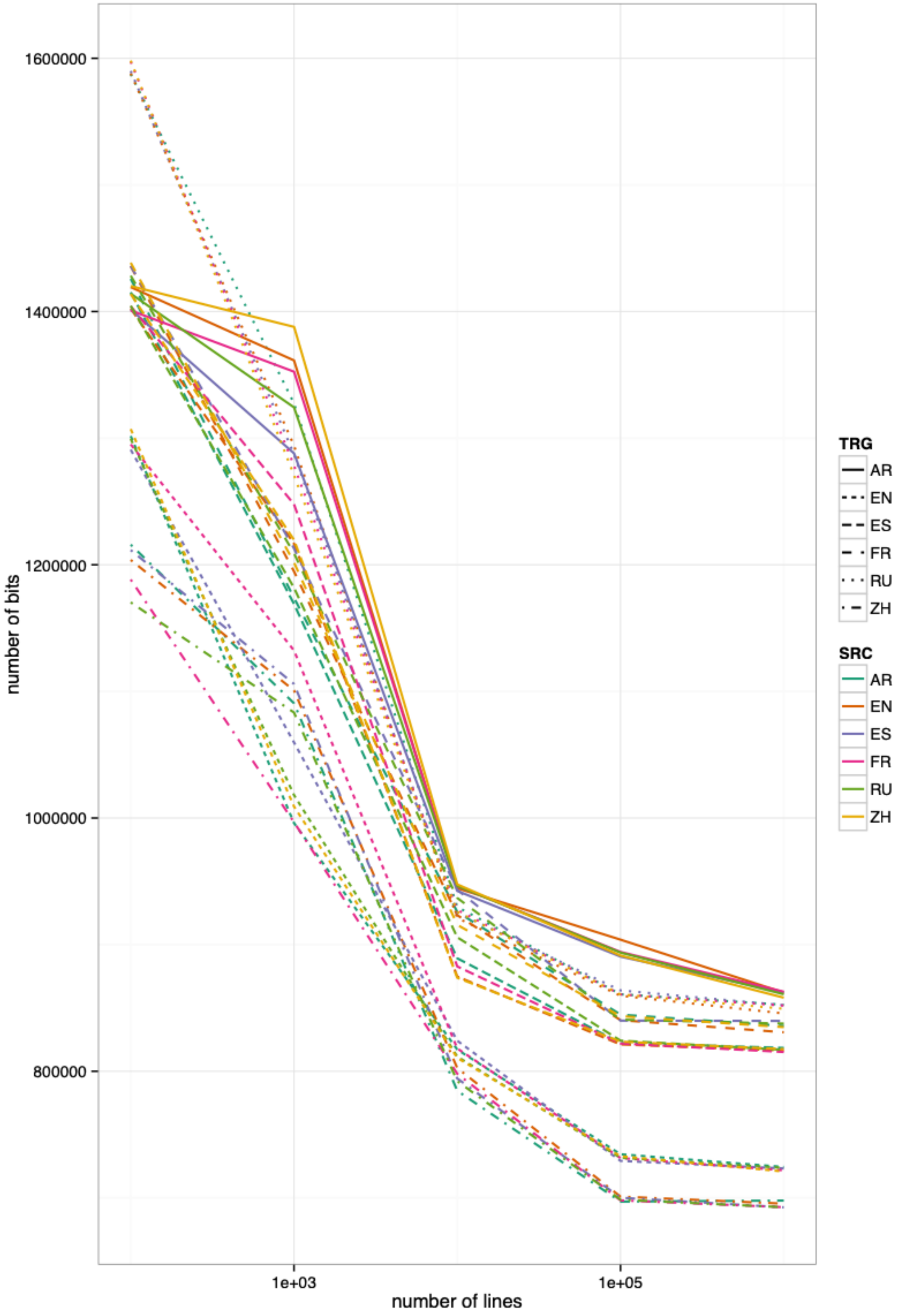


Figure 11: BPE models

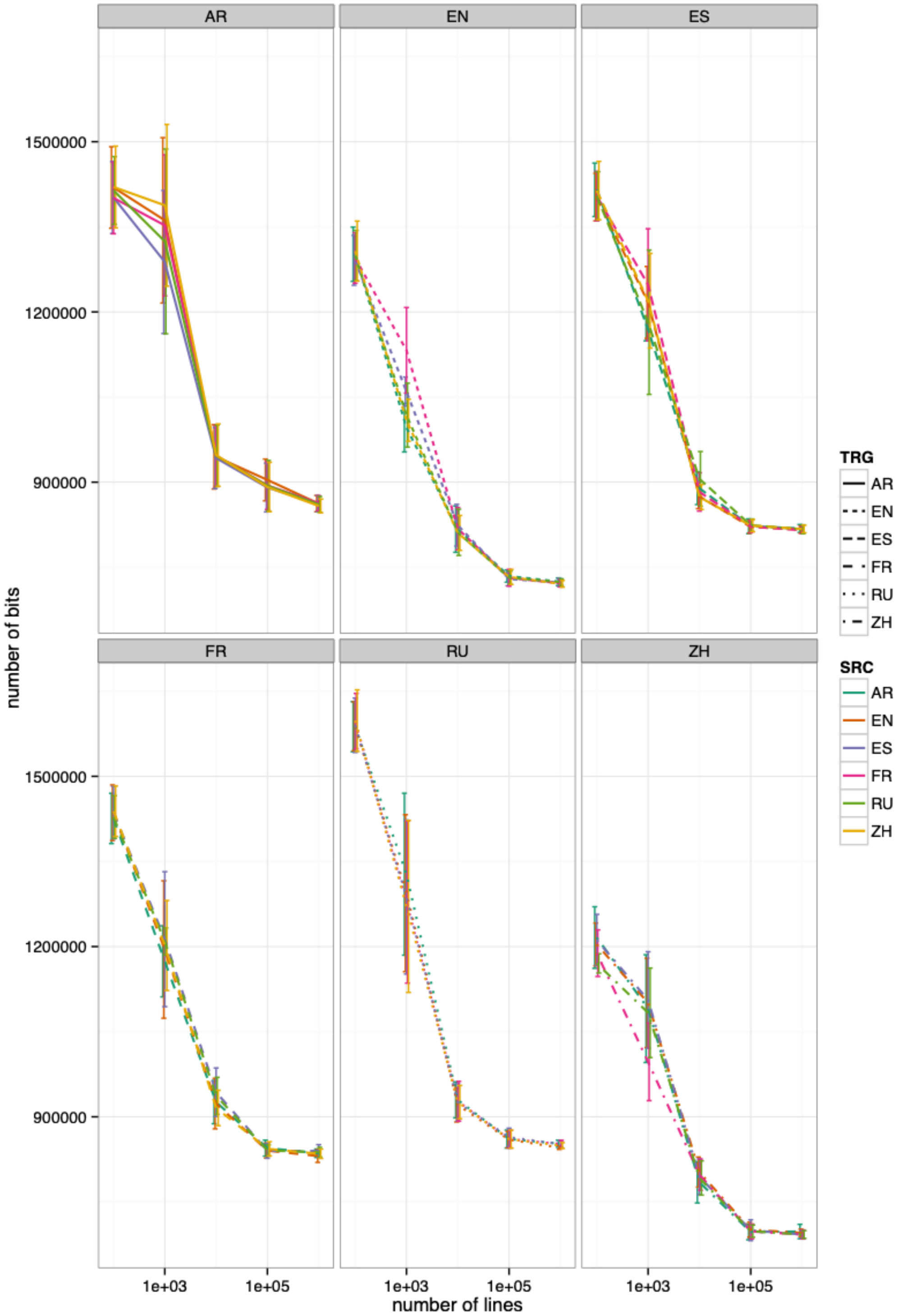


Figure 11: BPE models (target language as facet)

Fairness in Representation for Multilingual NLP: Insights from Controlled Experiments on Conditional Language Modeling

Version 1.0

ICLR 2022 camera-ready copy (20220315)