REPRESENTATION AND BIAS IN MULTILINGUAL NLP: INSIGHTS FROM CONTROLLED EXPERIMENTS ON CONDITIONAL LANGUAGE MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Inspired by the phenomenon of performance disparity between languages in machine translation, we investigate whether and to what extent languages are equally hard to "conditional-language-model". Our goal is to improve our understanding and expectation of the relationship between language, data representation, size, and performance. We study one-to-one, bilingual conditional language modeling through a series of systematically controlled experiments with the Transformer and the 6 languages from the United Nations Parallel Corpus. We examine character, byte, and word models in 30 language directions and 5 data sizes, and observe indications suggesting a script bias on the character level, a length bias on the byte level, and a word bias that gives rise to a hierarchy in performance across languages. We also identify two types of sample-wise non-monotonicity — while word-based representations are prone to exhibit Double Descent, length can induce unstable performance across the size range studied in a novel meta phenomenon which we term *erraticity*. By eliminating statistically significant performance disparity on the character and byte levels by normalizing length and vocabulary in the data, we show that, in the context of computing with the Transformer, there is no complexity intrinsic to languages other than that related to their statistical attributes and that performance disparity is not a necessary condition but a byproduct of word segmentation. Our application of statistical comparisons as a fairness measure also serves as a novel rigorous method for the intrinsic evaluation of languages, resolving a decades-long debate on language complexity. While all these quantitative biases leading to disparity are mitigable through a shallower network, we find room for a human bias to be reflected upon. We hope our work helps open up new directions in the area of language and computing that would be fairer and more flexible and foster a new transdisciplinary perspective for DL-inspired scientific progress.

1 INTRODUCTION

With a transdisciplinary approach to explore a space at the intersection of Deep Learning (DL) / Neural Networks (NNs), language sciences, and language engineering, we report our undertaking in **use-inspired basic research** — with an application-related phenomenon as inspiration, we seek **fundamental scientific understanding** through empirical experimentation. This is *not* an application or machine translation (MT) paper, but one that strives to evaluate and seek new insights on language in the context of DL with a consideration to contribute to our evaluation, segmentation, and model interpretation practice in multilingual Natural Language Processing (NLP).

Our *inspiration*: performance disparity in MT The use case that inspired our investigation is the disparity of MT results reported in Junczys-Dowmunt et al. (2016). Of the 6 official languages of the United Nations (UN) — Arabic (AR), English (EN), Spanish (ES), French (FR), Russian (RU), and Chinese (ZH), results with target languages AR, RU, and ZH seem to be worse than those with EN/ES/FR, regardless of the algorithm, may it be from phrased-based Statistical MT (SMT/Moses

(Koehn et al., 2007)) or Neural MT (NMT).[1] The languages have the same amount of line-aligned, high-quality parallel data available for training, evaluation, and testing. This prompts the question: are some languages indeed harder to translate from or to?

Problem statement: are all languages equally hard to Conditional-Language-Model (CLM)? A similar question concerning (monolingual) language modeling (LMing) was posed in Cotterell et al. (2018) and Mielke et al. (2019) along with the introduction of a method to evaluate LMs with multiway parallel corpora (multitexts) in information-theoretic terms. 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 effort as estimating conditional probabilities with the Transformer, with a bilingual setup where perplexity of one target language (l_{trg}) is estimated given the parallel data in one source language (l_{src}), where $l_{\text{src}} \neq l_{\text{trg}}$. We focus on the very basics and examine the first step in our pipeline — input representation, holding everything else constant. Instead of measuring absolute cross-entropy scores, we evaluate the relative differences between languages from across 5 magnitudes of data sizes in 3 different representation types/levels. **We consider *bias* to be present when performance disparity in our Transformer models is statistically significant.**

1.1 Summary of findings and contributions

In investigating performance disparity as a function of size and data with respect to language and representation on the Transformer in the context of CLMing, we find:

1. in a bilingual (one-to-one) CLMing setup, 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 (see Table 1).
2. We identify 2 types of **sample-wise non-monotonicity** on each of the primary representation levels we studied:
   (a) Double Descent (Belkin et al., 2019; Nakkiran et al., 2020): on the word level, for all languages, performance at 10^2 lines is typically better than at 10^3 before it improves again at 10^4 and beyond. This phenomenon can also be observed in character models with ZH as a target language as well as on the word level with non-neural n-gram LMs;
   (b) *erraticity*: performance is irregular and exhibits great variance across runs. We find sequence length to be predictive of this phenomenon. We show that this can be rectified by data transformation or hyperparameter tuning. In our study, erraticity affects AR and RU on the byte level where the sequences are too long with UTF-8 encoding and ZH when decomposed into strokes on the character level.
3. In eliminating performance disparity through lossless data transformation on the character and byte levels, we resolve language complexity (§ 4 and App. J). We show that, in the context of computing with the Transformer, unless word-based methods are used, there is no linguistic/morphological complexity applicable or necessary. There is **no complexity that is intrinsic to a language aside from its statistical properties**. Hardness in modeling is relative to and bounded by its representation level (**representation relativity**). On the character and byte levels, hardness is correlated with statistical properties concerning sequence length and vocabulary of a language, irrespective of its linguistic typological, phylogenetic, historical, or geographical profile, and can be eliminated. On the word level, hardness is correlated with vocabulary, and a complexity hierarchy arises through the manual preprocessing step of word tokenization. This complexity/disparity effected by word segmentation 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. We find clarification of this expectation of disparity necessary because more diligent error analyses need to be afforded instead of simply accepting massively disparate results or inappropriately attributing under-performance to linguistic reasons.
4. Representational units of **finer granularity** can help close the gap in performance disparity.
5. Bigger/overparameterized models can **magnify/exacerbate the effects of differences in data statistics**. Quantitative biases that lead to disparity are mitigable through numerical methods.

[1]We provide a re-visualization of these grouped in 6 facets by target language in Figure 4 in Appendix A.

Outline of the paper In § 2, we define our method and experimental setup. We present our results and analyses on the primary representations in § 3 and those from secondary set of controls in § 4 in a progressive manner to ease understanding. Meta analyses on fairness evaluation, non-monotonic behavior, and discussion on biases are in § 5. Additional related work is in § 6. We refer our readers to the Appendices for more detailed descriptions/discussions and reports on supplementary experiments.

2 Method and definitions

Controlled experiments as basic research for scientific understanding Using the United Nations Parallel Corpus (Ziemski et al., 2016), the data from which the MT results in Junczys-Dowmunt et al. (2016) stem, we perform a series of controlled experiments on 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 — 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: Byte Pair Encoding (BPE) (Sennrich et al., 2016), an adapted compression algorithm from Gage (1994). 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 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 domain 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). 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 *relations* between scores from different languages and the generalizations derived therefrom.**

Information-theoretic, fair evaluation with multitexts Most sequence-to-sequence models are optimized using a cross-entropy loss (see Appendix B for definition). Cotterell et al. (2018) propose to use "renormalized" perplexity (PP) to evaluate LMs fairly using the total number of bits divided by some constant. In our case, we choose instead a simpler method of using an "unnormalized" PP, directly using the total number of bits needed to encode the development (dev) set, which has a constant size of 3,077 lines per language.

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. 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[2] (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 evaluation.

Experimental setup The systematic, identical treatment we give to our data is described as follows with further preprocessing and hyperparameter details in Appendices B and C, respectively. The distinctive point of our experiment is that the training regime is the same for all (intuition in App. O.1).

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

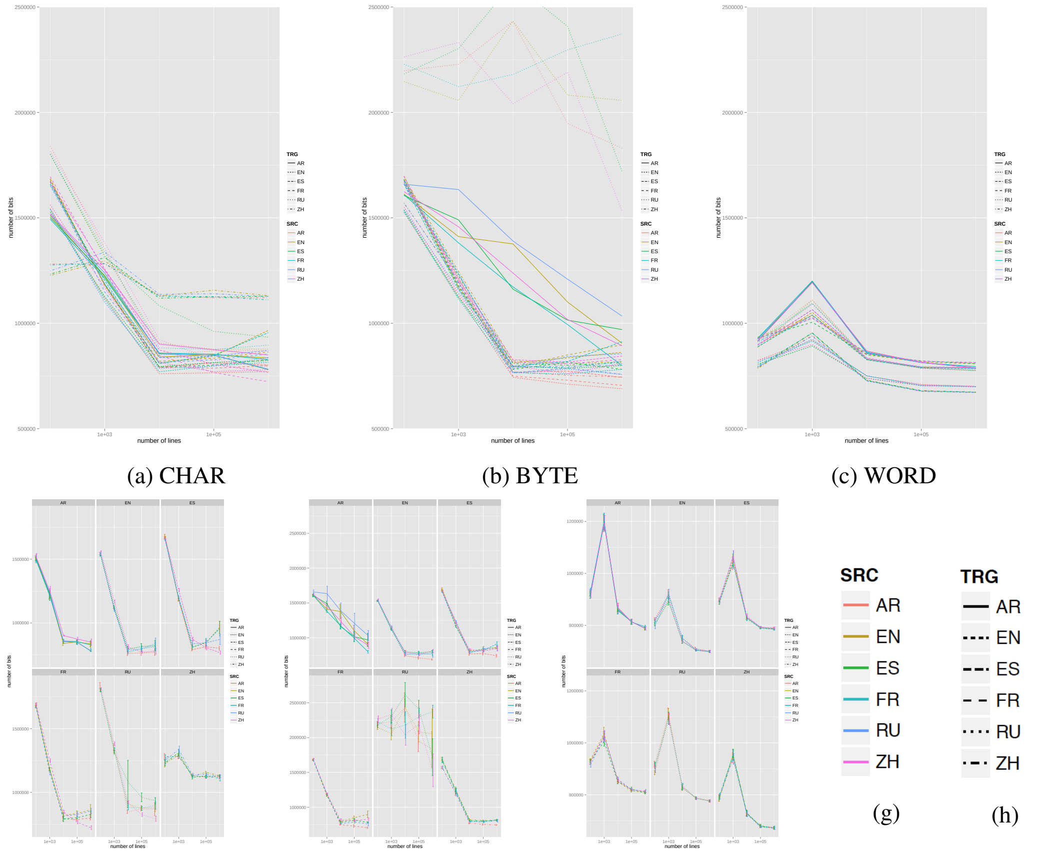
After filtering length to 300 characters maximum per line in parallel for the 6 languages, we made 3 subsets of the data with 1 million lines each — one having lines in the order of the original corpus (dataset A) and two other randomly sampled (without replacement) from the full corpus (datasets B & 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_{src} to one l_{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, vocabulary size ($|V|$), type-token-ratio, OOV rate) as metrics — a complete list thereof is provided in Appendix F. For each of the 3 primary representations — character, byte, and word, we performed 5 runs total in 5 sizes (10^2-10^6 lines) (runs A0, B0, C0, A1, & A2) and 7 more runs in 4 sizes (10^2-10^5 lines) (A3-7, B1, & C1), also controlling for seeds. For the alternate/secondary representations, we ran 3 runs each in 5 sizes (10^2-10^6 lines) (A0, B0, & C0).

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 in Appendix B). 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 10^6 lines) need 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** — 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 (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 (cf. figures sorted by source language in Appendix H) — 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$_{trg}$) shows a different learning pattern throughout. And on the byte level, AR$_{trg}$ and RU$_{trg}$ display non-monotonic and 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



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

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

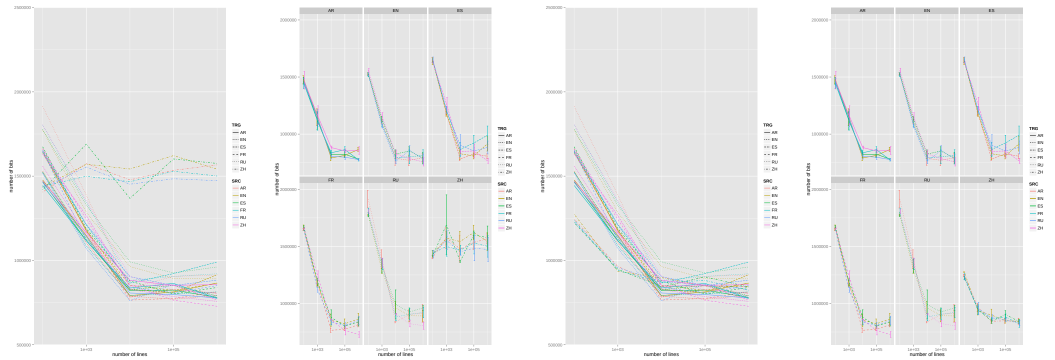
Figure 1: Number of bits (the lower the better) as a function of data size plotted for all 30 directions. Subfigures 1d, 1e, and 1f depict the corresponding information as in 1a, 1b, and 1c (showing mean across 12 runs), respectively, but sorted in 6 facets by target language and with error bars. Legend 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 G.)

ambiguous — 1) it is often aimed to align with that in other languages so to accommodate manual feature engineering and academic theories, 2) there is great variation among different conventions, 3) 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 (cf. Li et al. (2019b) as well as Duanmu (2017) for a summary on the contested nature of wordhood in ZH). A more native analysis of ZH, despite being considered a high-resource language, has not yet been recognized in NLP.

4 Understanding the Phenomena with Alternate Representations

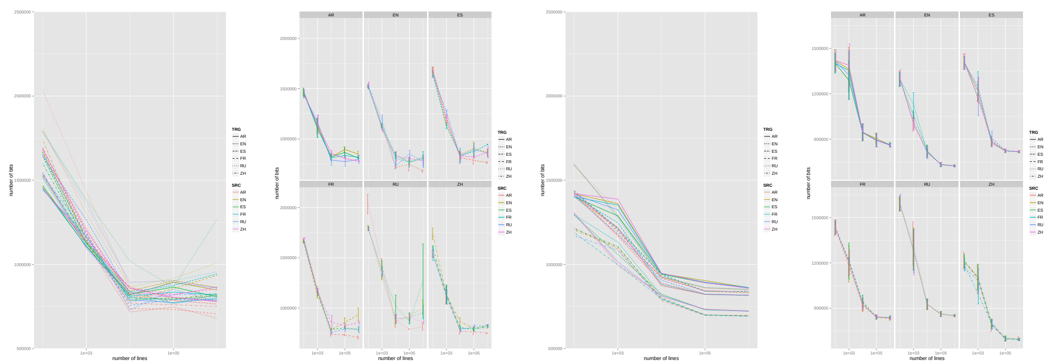
To understand why some languages show different results than others, we carried out a secondary set of control experiments with representations targeting the problematic statistical properties of the corresponding target languages. (An extended version of this section is provided in Appendix P.)

Character level We reduced the high $|V|$ in ZH with representations in ASCII characters — Pinyin and Wubi. The former is a romanization of ZH characters based on their pronunciations and the latter an input algorithm that decomposes character-internal information into stroke shape and ordering and matches these to 5 classes of radicals (Lunde, 2008). We replaced the ZH data in these formats *only on the target side* and reran the experiments involving ZH_{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



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

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



(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.

through Pinyin (transliteration), as in Subfigure 2c. But models with ZH logographic scripts display a behaviorial tendency unlike those with other (phonetic) alphabetic scripts (Subfigure 2a). Work published thus far using Wubi with the Transformer seems to have needed some form of architectural modification (Gao et al., 2020) or a different architecture altogether (Nikolov et al., 2018; Zhang et al., 2019), suggesting a possible script bias (to be further discussed in § 5 under "Basis for biases").

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 AR$_{trg}$ and RU$_{trg}$ — 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 "ARRU$_t$"), 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**[3] **does not exist** in computing, however diverse they may be, as our 6 are, from the conventional linguistic typological, phylogenetic, historical, or geographical perspectives. Please refer to Appendix J for our discussion on language complexity.

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

[3]aside from its statistical properties related to length and vocabulary. "Language" here refers to language represented through all representations.

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

		CHAR		Pinyin		Wubi		BYTE		ARRU_t		$\text{ARRU}_{s,t}$		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

segmentation) from the training data. It is more fine-grained than word segmentation and is known for its capability to model subword units for morphologically complex languages (e.g. AR and RU). 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, showing how **finer-grained modeling has a positive effect on closing the disparity gap**.

5 Meta-results, analysis, and discussion

Performance disparity Table 1 lists the number of language pairs with significant differences under the representations studied. Considering how it is **possible** for our character and byte models to effect no performance disparity for the same languages on the same data, this indicates that disparity is not a necessary condition. In fact, the customary expectation that languages ought to perform differently stems from our word segmentation practice. Furthermore, the order of AR/RU > ES/FR > EN/ZH (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 is practiced today[4] and that which has occurred in the NLP discourse thus far, has only been relevant on and is bounded to the "word" level. The definition of word, however, has been recognized as problematic for a very long time in the language sciences (see Haspelmath (2011) and references therein from the past century). Since the conventional notion of word, which has been centered on English and languages with alphabetic scripts, has a negative impact on languages both morphologically rich (see Minkov et al. (2007), Seddah et al. (2010), inter alia), AR and RU in our case, as well as morphologically "frugal" (Koehn, 2005), as in ZH, finer-grained modeling with characters and bytes (or n-gram variants/pieces thereof) is indeed a more sensible option and enables a greater variety of languages to be handled with more simplicity, fairness, independence, and flexibility.

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 "$\text{ARRU}_{s,t}$"). We defer further investigation on the nature of source language neutralization to future work.

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 ZH_{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

[4]But there are no reasons why linguistics or linguistic typology cannot encompass a statistical science of language beyond/without "words", or with continuous representations of characters and bytes. In fact, that could complement the needs of language engineering and the NNs/DL/ML communities better.

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. If we could use this ratio and logic of higher $|V|$ to automatically detect "non-atomic" units, ones that can be further decomposed, this observation could potentially be beneficial for advancing other sciences, e.g. biology. From a cognitive modeling perspective, the similarity in behavior of ZH characters and words of other languages can affirm the interpretation of wordhood for those ZH speakers who identify ZH characters as words (see also last paragraph in § 3 and Appendix J). While almost all work attribute DD to algorithmic reasons, concurrent work by Chen et al. (2020) corroborates 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 DD and its more recent development can also be found in their work.

We performed additional experiments testing our setting on the datasets used by the Nakkiran et al. (2020) and testing our data on a non-neural LM. Results support our findings and are provided in Appendix K. Number of model parameters can be found in Appendix L.

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 (see Figure 18 in Appendix M). 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. If the connection between erraticity and length bias can indeed be drawn, it could strengthen the case for global conditioning (Sountsov & Sarawagi, 2016). (See Appendix M for more discussion and results.)

Script bias, erraticity, word bias — are these necessary conditions? To assess whether the observed phenomena are particular to this one setting, we performed one run with dataset A in 4 sizes with the primary representations on 1-layer Transformers (see Appendix N). We observed no significant disparity across the board. It shows that **larger/overparameterized models can magnify/exacerbate the differences in the data statistics**. That hyperparameter tuning — in this case, through the reduction of the number of layers — can mitigate effects from data statistics is, to the best of our knowledge, a novel insight, suggesting also that a general expectation of monotonic development as data size increases can indeed be held. Our other findings remain consistent (representational relativity, source language neutralization, and DD on word level).

Bases for biases Recall in § 1, we "consider *bias* to be present when performance disparity in our Transformer models is statistically significant". As shown in our data statistics and analysis (Appendices D and P respectively), script bias, length bias wrt erraticity in CLMing, and word bias are all evident in the vocabulary and length information in the data statistics. Hence these disparities in performance are really a result of the Transformer being able to model these **differences in data** at such a magnitude that the differences are statistically significant. The meta phenomenon of erraticity, however, warrants an additional consideration indicative of the **empirical limits of our compute** (cf. Xu et al. (2020)), even when the non-monotonicity is not observed during the training of each model.

In eliminating performance disparity in character and byte models by normalizing vocabulary and length statistics in the data, we demonstrated that performance disparity as expected from the morphological complexity hierarchy is due to word tokenization, not intrinsic or necessary in language. This is the word bias. Qualitative issues in the concept of word will persist and make crosslinguistic comparison involving "words" unfair even if one were to be able to find a quantitative solution to mitigate the OOV issue, the bottleneck in word-based processing. We humans have a choice in how we see/process languages. That some might still prefer to continue with a crosslinguistic comparison with "words" and exert the superiority of "word" tokenization speaks for a view that is centered on "privileged" languages — in that case, **word bias is a human bias**.

And, in eliminating performance disparity across the board with our one-layer models, we show that all quantitative differences in data statistics between languages can also be modeled in a "zoomedout"/"desensitized" mode, suggesting that while languages can be perceived as being fundamentally different in different ways in different granularities, they can also be viewed as fundamentally similar.

6 ADDITIONAL RELATED WORK

Similar to our work in testing for hardness are Cotterell et al. (2018), Mielke et al. (2019), and Bugliarello et al. (2020). The first two studied (monolingual) LMs — the former tested on the Europarl languages (Koehn, 2005) with n-gram and character models and concluded that morphological complexity was the culprit to hardness, the latter studied 62 languages of the Bible corpus (Mayer & Cysouw, 2014) in addition and refuted the relevance of linguistic features in hardness based on character and BPE models on both corpora in word-tokenized form. Bugliarello et al. (2020) compared translation results of the Europarl languages with BPEs at one data size and concluded that it is easier to translate out of EN than into it, statistical significance was, however, not assessed. In contrast, we ablated away the confound of generation and studied CLMing with controls with a broader range of languages with more diverse statistical profiles in 3 granularities and up to 5 orders of magnitude in data size. 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.

Bender (2009) advocated the relevance of linguistic typology for the design of language-independent NLP systems based on crosslinguistic differences in word-based structural notions, such as parts of speech. Ponti et al. (2019) found typological information to be beneficial in the few-shot setting on the character level for 77 languages with Latin scripts. But no multilingual work has thus far explicitly examined the relation between linguistic typology and the statistical properties of the data, involving languages with diverse statistical profiles in different granularities.

As obtaining training data is often the most difficult part of an NLP or Machine Learning (ML) project, Johnson et al. (2018) introduced an extrapolation methodology to directly model the relation between data size and performance. Our work can be viewed as one preliminary step towards this goal. 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, or the connection between effects of data statistics and modification in architectural depth.

7 CONCLUSION

Summary We performed a novel, rigorous relational assessment of performance disparity across different languages, representations, and data sizes in CLMing with the Transformer. Different disparity patterns were observed on different representation types (character, byte, and word), which can be traced back to the data statistics. The disparity pattern reflected on the word level corresponds to the morphological complexity hierarchy, reminding us that the definition of morphology is predicated on the notion of word and indicating how morphological complexity can be modeled by the Transformer simply through word segmentation. As we were able to eliminate disparity on the same data on the character and byte levels by normalizing length and vocabulary, we showed that morphological complexity is not a necessary concept but one that results from word segmentation and is bounded to the word level, orthogonal to the performance of character or byte models. 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 ZH_{trg} to behave similarly in their being prone to exhibit a peak (as sample-wise DD) around 10^3 lines in our setting. While bigger/overparameterized models can magnify the effect of data statistics, exacerbating the disparity, we found a decrease in model depth can eliminate these quantitative biases, leaving only the qualitative aspect of "word" and the necessity of word segmentation in question.

Outlook Machine learning 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 DL/NNs, language sciences, and language engineering through a data-centric perspective.

We believe a **statistical** science for NLP as a data science can well complement algorithmic analyses with an empirical view contributing to a more generalizable pool of knowledge for NNs/DL/ML. A more comprehensive study not only can lead us to new scientific frontiers, but also better design and evaluation, benefitting the development of a more general, diverse and inclusive Artificial Intelligence.

REFERENCES

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.

E. M. Bender. *Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax*. 2013.

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.

Christian Bentz, Tatyana Ruzsics, Alexander Koplenig, and Tanja Samardžić. A comparison between morphological complexity measures: Typological data vs. language corpora. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pp. 142–153, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://www.aclweb.org/anthology/W16-4117.

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.

Mauro Cettolo, Christian Girardi, and Marcello Federico. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pp. 261–268, Trento, Italy, May 2012.

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

Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech Language*, 13(4):359 – 394, 1999. ISSN 0885-2308. doi: https://doi.org/10.1006/csla.1999.0128. URL http://www.sciencedirect.com/science/article/pii/S0885230899901286.

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.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pp. 103–111, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-4012. URL https://www.aclweb.org/anthology/W14-4012.

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.

Philip Gage. A new algorithm for data compression. *C Users J.*, 12(2):23–38, February 1994. ISSN 0898-9788. URL http://dl.acm.org/citation.cfm?id=177910.177914.

Yingqiang Gao, Nikola I. Nikolov, Yuhuang Hu, and Richard H.R. Hahnloser. Character-level translation with self-attention. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1591–1604, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.145. URL https://www.aclweb.org/anthology/2020.acl-main.145.

Martin Gellerstam. Translationese in Swedish novels translated from English. In Lars Wollin and Hans Lindquist (eds.), *Translation Studies in Scandinavia*, pp. 88–95. CWK Gleerup, 1986.

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.

Kenneth Heafield. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pp. 187–197, Edinburgh, Scotland, July 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W11-2123.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 690–696, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P13-2121.

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.

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.

Mark Johnson, Peter Anderson, Mark Dras, and Mark Steedman. Predicting accuracy on large datasets from smaller pilot data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 450–455. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/P18-2072.

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.

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 *Conference Proceedings: the tenth Machine Translation Summit*, pp. 79–86, Phuket, Thailand, 2005. AAMT, AAMT. URL http://mt-archive.info/MTS-2005-Koehn.pdf.

Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pp. 28–39. Association for Computational Linguistics, 2017. doi: 10.18653/v1/W17-3204. URL http://aclweb.org/anthology/W17-3204.

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. More data can hurt for linear regression: Sample-wise double descent, 2019.

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.

Nikola Nikolov, Yuhuang Hu, Mi Xue Tan, and Richard H.R. Hahnloser. Character-level Chinese-English translation through ASCII encoding. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 10–16, Belgium, Brussels, October 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-6302.

Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003. doi: 10.1162/089120103321337421. URL https://www.aclweb.org/anthology/J03-1002.

M Opper, W Kinzel, J Kleinz, and R Nehl. On the ability of the optimal perceptron to generalise. *Journal of Physics A: Mathematical and General*, 23(11):L581–L586, jun 1990. doi: 10.1088/0305-4470/23/11/012. URL https://doi.org/10.1088%2F0305-4470%2F23%2F11%2F012.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

Edoardo Maria Ponti, Ivan Vulić, Ryan Cotterell, Roi Reichart, and Anna Korhonen. Towards zero-shot language modeling. 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. 2900–2910, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1288. URL https://www.aclweb.org/anthology/D19-1288.

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 Koebler, 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.

Pavel Sountsov and Sunita Sarawagi. Length bias in encoder decoder models and a case for global conditioning. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1516–1525, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1158.

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.

Felix Stahlberg and Bill Byrne. On NMT search errors and model errors: Cat got your tongue? 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. 3356–3362, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1331. URL https://www.aclweb.org/anthology/D19-1331.

Reut Tsarfaty, Djamé Seddah, Yoav Goldberg, Sandra Kuebler, Yannick Versley, Marie Candito, Jennifer Foster, Ines Rehbein, and Lamia Tounsi. Statistical parsing of morphologically rich languages (SPMRL) what, how and whither. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, pp. 1–12, Los Angeles, CA, USA, June 2010. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W10-1401.

Reut Tsarfaty, Djamé Seddah, Sandra Kübler, and Joakim Nivre. Parsing morphologically rich languages: Introduction to the special issue. *Computational Linguistics*, 39(1):15–22, 2013. doi: 10.1162/COLI_a_00133. URL https://www.aclweb.org/anthology/J13-1003.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz 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.

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.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1eBeyHFDH.

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.

Wei Zhang, Feifei Lin, Xiaodong Wang, Zhenshuang Liang, and Zhen Huang. Subcharacter Chinese-English neural machine translation with Wubi encoding, 2019.

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 Re-visualization of Figure 1 in Junczys-Dowmunt et al. (2016) in 6 facets by target language

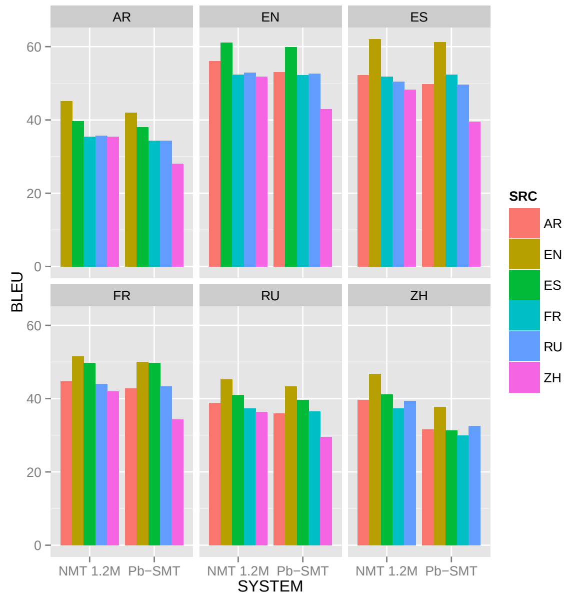


Figure 4: Results of the Moses baseline systems (right group in each facet) and neural models (left) with 1.2 million iterations (1 iteration corresponds to 1 mini-batch) for the 30 directions of the 6-way UN corpus, tokenized (ZH segmented), lowercased, and length filtered to 100 BPE tokens.

B 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[5], 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 75^{th} 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[6]. 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).

[5]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 AR_{trg} or RU_{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.

[6]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).

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[7] for segmentation in ZH.

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`.

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 D for reference.

The systematic training regime that we give to our language directions are identical for all. For each primary representation type (character, byte, and word), we performed:

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

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 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 C.

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 AR_{trg} or RU_{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.

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 (based on whichever granularity of unit is used for training) in the data. Cotterell et al. (2018) propose to use "renormalized" PP to evaluate LMs fairly through the division of an arbitrary constant. In our case, we choose instead a simpler method of using an "unnormalized" PP, i.e. the total number of bits needed to encode the development (dev) set, which has a constant size of 3,077 lines per language (after length filtering of the same dev set used in Junczys-Dowmunt et al. (2016)) for all various training sizes. As the implementation we used (SOCKEYE (Hieber et al., 2018)) only reports PP, we transform it back to entropy as defined above by noting that $H(\boldsymbol{t}, \boldsymbol{s}) = \log_2 PP(\boldsymbol{t}|\boldsymbol{s}) \times N$.

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

C 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.)

D 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				
Number of lines	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
EN	75	81	122	167	244	76	82	127	151	166
ES	84	90	114	173	254	85	91	118	151	166
FR	86	94	130	175	256	87	96	130	153	172
RU	95	108	160	187	260	97	110	148	157	166
ZH	602	1,236	2,267	3,400	4,885	137	143	153	164	193
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
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
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
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
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
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
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
EN	23.71	20.62	11.34	4.12	2.06	26.47	23.53	6.86	0.98	0.00
ES	25.23	20.72	16.22	8.11	7.21	23.64	19.09	11.82	4.55	0.91
FR	25.66	19.47	11.50	4.42	2.65	27.12	20.34	7.63	2.54	0.00
RU	34.93	26.71	4.11	2.05	0.00	32.64	24.31	3.47	1.39	0.00
ZH	69.79	39.73	10.58	1.87	0.56	10.46	7.19	3.27	1.31	0.00
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
EN	0.05	0.02	0.00	0.00	0.00	0.06	0.02	0.00	0.00	0.00
ES	0.07	0.01	0.01	0.01	0.01	0.07	0.02	0.01	0.01	0.00
FR	0.19	0.10	0.07	0.00	0.00	0.21	0.11	0.07	0.00	0.00
RU	0.30	0.24	0.16	0.00	0.00	0.16	0.13	0.08	0.00	0.00
ZH	17.83	3.64	0.38	0.04	0.01	0.05	0.02	0.00	0.00	0.00
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
EN	0.64	0.05	0.01	0.00	0.00	0.65	0.05	0.01	0.00	0.00
ES	0.68	0.05	0.01	0.00	0.00	0.67	0.05	0.01	0.00	0.00
FR	0.69	0.05	0.01	0.00	0.00	0.68	0.05	0.01	0.00	0.00
RU	0.79	0.06	0.01	0.00	0.00	0.45	0.04	0.01	0.00	0.00
ZH	18.14	2.90	0.61	0.09	0.01	1.60	0.12	0.02	0.00	0.00
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 8/33/91/136/223	123.83 ± 47.09 5/94/126/155/263	108.35 ± 53.79 1/69/109/147/299	106.25 ± 54.88 1/64/107/147/299	102.06 ± 57.25 1/54/102/145/300	166.55 ± 99.88 16/62/166/247/404	227.16 ± 86.09 9/172/230/285/479	198.50 ± 99.02 1/126/200/269/554	194.88 ± 101.03 1/117/196/270/561	186.17 ± 104.86 1/98/187/265/565
EN	117.30 ± 72.51 8/49/112/170/265	159.44 ± 58.88 8/117/163/203/285	134.40 ± 66.79 1/86/137/185/300	131.33 ± 68.25 1/79/134/183/300	123.49 ± 69.53 1/65/125/177/300	117.31 ± 72.50 8/49/112/170/265	159.45 ± 58.88 8/117/163/203/285	134.58 ± 66.80 1/87/137/185/300	131.59 ± 68.34 1/80/134/184/300	123.71 ± 69.64 1/65/125/177/311
ES	123.74 ± 78.91 6/42/119/186/296	171.10 ± 64.03 6/127/175/217/297	148.48 ± 73.38 1/95/150/205/300	145.50 ± 75.32 1/88/148/204/300	138.60 ± 77.56 1/73/140/200/300	126.29 ± 80.21 7/44/121/191/302	175.29 ± 65.44 7/130/179/222/306	151.38 ± 74.70 1/97/153/209/307	148.21 ± 76.67 1/89/150/208/312	141.28 ± 78.96 1/74/143/203/318
FR	124.56 ± 79.38 5/45/116/190/287	179.05 ± 66.44 5/133/183/228/300	149.10 ± 74.19 1/93/152/206/300	145.29 ± 75.75 1/87/148/204/300	138.05 ± 77.90 1/72/140/199/300	128.75 ± 82.16 5/47/121/196/297	185.23 ± 68.52 5/138/190/236/312	154.21 ± 76.66 1/96/157/213/322	150.56 ± 78.42 1/90/153/212/322	143.50 ± 80.99 1/75/145/207/331
RU	119.80 ± 72.09 10/44/119/183/260	168.16 ± 62.11 5/126/174/213/298	143.61 ± 71.55 1/91/146/197/300	141.52 ± 73.95 1/85/144/199/300	134.71 ± 76.20 1/70/136/194/300	217.51 ± 134.51 20/77/215/332/489	309.28 ± 114.98 10/231/318/392/554	263.66 ± 132.77 1/167/268/363/562	259.90 ± 137.10 1/155/264/365/566	247.10 ± 141.14 1/127/249/357/567
ZH	33.18 ± 18.38 3/19/30/45/88	42.57 ± 16.91 2/32/42/53/127	37.20 ± 18.46 1/24/37/49/165	36.60 ± 19.05 1/23/36/49/241	34.67 ± 19.29 1/19/34/48/281	85.59 ± 51.18 7/35/81/127/192	116.67 ± 43.82 6/88/117/148/243	102.00 ± 50.61 1/66/103/139/297	99.90 ± 51.49 1/61/101/138/337	94.27 ± 52.80 1/51/95/134/594
ZH_pinyin	108.11 ± 65.51 7/43/98/161/255	148.08 ± 55.57 7/109/149/188/312	127.96 ± 63.96 1/82/129/175/369	125.31 ± 65.25 1/76/126/174/407	118.47 ± 66.88 1/63/119/169/645					
ZH_wubi	92.45 ± 55.26 8/39/86/137/218	127.19 ± 48.00 7/94/129/158/266	110.45 ± 54.71 1/71/111/150/327	108.86 ± 56.04 1/67/110/150/364	102.91 ± 57.41 1/55/103/146/627					
AR_cp1256						90.79 ± 54.45 8/33/91/136/223	123.83 ± 47.09 5/94/126/155/263	108.35 ± 53.79 1/69/109/147/299	106.25 ± 54.88 1/64/107/147/299	102.06 ± 57.25 1/54/102/145/300
RU_cp1251						119.80 ± 72.09 10/44/119/183/260	168.16 ± 62.11 5/126/174/213/298	143.61 ± 71.55 1/91/146/197/300	141.52 ± 73.95 1/85/144/199/300	134.71 ± 76.20 1/70/136/194/300

Representation	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
Number of TYPES										
AR	866	5,865	26,105	97,994	304,975	700	4,171	18,202	29,706	29,961
EN	679	3,289	12,011	37,261	122,454	645	2,856	10,225	26,184	29,068
ES	707	3,836	15,113	49,496	142,806	661	3,232	12,083	27,908	29,445
FR	742	3,878	13,995	42,867	118,201	723	3,193	11,731	27,393	29,333
RU	883	5,430	23,400	81,339	236,085	842	4,319	17,995	29,239	29,807
ZH	627	3,178	13,134	46,938	134,489	800	2,849	10,656	27,284	29,148
Number of TOKENS										
AR	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	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	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	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	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	1,751	23,568	207,714	2,038,639	19,361,101	2,223	25,459	215,179	2,083,632	19,885,789
OOV type rate (%)										
AR	95.80	79.80	48.02	24.19	8.14	10.17	1.39	0.09	0.00	0.00
EN	92.07	69.08	33.56	15.27	7.38	4.73	1.01	0.18	0.05	0.01
ES	93.32	71.90	34.72	13.24	6.12	5.44	1.06	0.27	0.10	0.08
FR	92.50	70.25	32.16	12.31	5.62	5.11	0.99	0.26	0.05	0.02
RU	94.52	75.84	39.14	14.54	5.69	6.38	1.27	0.06	0.02	0.00
ZH	93.19	73.29	38.59	15.63	5.65	76.09	40.95	8.52	0.89	0.21
OOV token rate (%)										
AR	60.86	40.28	17.94	8.62	2.10	2.57	1.99	0.11	0.00	0.00
EN	39.04	18.57	6.16	2.70	1.01	0.35	0.08	0.02	0.01	0.00
ES	36.21	18.28	6.28	2.34	0.86	0.28	0.13	0.05	0.03	0.03
FR	35.44	17.47	6.04	2.13	0.76	0.64	0.52	0.36	0.01	0.00
RU	53.22	30.94	12.59	4.17	1.28	0.88	1.07	0.84	0.01	0.00
ZH	45.48	22.08	8.34	3.20	0.83	31.67	11.76	1.66	0.14	0.03
TTR (%)										
AR	48.76	25.00	12.64	4.81	1.57	24.08	14.78	8.14	1.34	0.14
EN	32.79	12.00	5.08	1.59	0.56	21.64	9.59	4.20	1.10	0.13
ES	31.68	13.02	5.75	1.91	0.58	20.79	10.00	4.44	1.06	0.12
FR	32.29	12.11	5.12	1.60	0.46	22.28	9.18	4.18	1.00	0.11
RU	47.63	21.94	10.80	3.78	1.16	27.07	14.84	7.78	1.27	0.14
ZH	35.81	13.48	6.32	2.30	0.69	35.99	11.19	4.95	1.31	0.15
Mean line length±std 0/25/50/75/100-th										
AR	17.76 ± 10.78 1/7/17/27/48	23.46 ± 9.12 1/18/24/29/56	20.65 ± 10.24 1/13/20/28/104	20.35 ± 10.46 1/13/20/28/107	19.41 ± 10.83 1/11/19/27/137	29.07 ± 17.48 3/14/27/43/75	28.23 ± 11.33 1/21/28/35/71	22.35 ± 11.30 1/14/22/30/113	22.18 ± 11.66 1/14/22/30/122	21.34 ± 12.13 1/12/21/30/173
EN	20.71 ± 12.16 1/9/19/30/46	27.40 ± 10.35 1/20/28/35/60	23.66 ± 11.68 1/15/24/32/93	23.39 ± 12.02 1/15/24/32/106	21.94 ± 12.29 1/12/22/31/136	29.80 ± 17.18 3/16/27/42/70	29.78 ± 11.32 1/22/30/37/68	24.36 ± 12.09 1/16/24/33/101	23.70 ± 12.20 1/15/24/32/106	22.35 ± 12.53 1/12/22/32/138
ES	22.32 ± 13.97 1/7/22/32/52	29.46 ± 11.35 1/22/29/37/63	26.30 ± 13.03 1/17/26/36/104	25.89 ± 13.35 1/16/26/36/107	24.65 ± 13.78 1/13/25/35/139	31.80 ± 19.39 4/16/31/43/79	32.32 ± 12.60 1/24/32/41/72	27.20 ± 13.57 1/18/27/37/106	26.44 ± 13.67 1/16/27/37/107	25.26 ± 14.14 1/14/25/36/140
FR	22.98 ± 14.16 1/9/22/34/50	32.01 ± 11.99 1/24/32/41/70	27.32 ± 13.53 1/17/28/37/107	26.85 ± 13.89 1/16/27/37/110	25.60 ± 14.45 1/14/26/37/138	32.45 ± 19.52 2/16/30/46/76	34.80 ± 13.09 2/27/35/44/77	28.09 ± 13.96 1/18/28/38/107	27.27 ± 14.13 1/17/28/38/110	26.08 ± 14.71 1/14/26/37/138
RU	18.54 ± 10.65 1/8/18/27/39	24.75 ± 9.53 1/19/25/31/58	21.66 ± 10.81 1/14/22/29/104	21.51 ± 11.12 1/13/22/30/107	20.42 ± 11.44 1/11/20/29/136	31.10 ± 17.42 4/18/29/45/69	29.11 ± 11.51 1/21/29/37/68	23.13 ± 11.71 1/15/23/31/104	23.08 ± 12.17 1/14/23/32/110	22.07 ± 12.57 1/12/22/31/166
ZH	17.51 ± 10.68 2/7/17/25/46	23.57 ± 9.52 1/18/23/29/76	20.77 ± 10.48 1/13/20/28/108	20.39 ± 10.60 1/13/20/28/111	19.36 ± 10.87 1/11/19/27/167	22.23 ± 12.33 2/13/20/31/58	25.46 ± 10.29 1/19/25/32/79	21.52 ± 10.94 1/14/21/29/108	20.84 ± 10.92 1/13/21/28/111	19.89 ± 11.23 1/11/20/28/167

Statistics for dataset B

Representation	CHAR					BYTE				
Number of lines	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
EN	76	99	131	189	343	79	103	135	163	181
ES	86	104	134	189	366	89	108	133	161	185
FR	89	112	147	202	352	91	120	142	162	185
RU	109	142	173	204	320	112	141	159	161	180
ZH	761	1,575	2,402	3,465	4,705	123	153	165	175	188
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
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
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
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
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
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
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
EN	25.77	11.34	8.25	2.06	0.00	26.47	11.76	5.88	0.00	0.00
ES	26.13	12.61	11.71	3.60	0.00	24.55	10.91	8.18	0.91	0.00
FR	23.01	11.50	7.96	2.65	0.00	24.58	7.63	5.08	0.85	0.00
RU	26.71	6.16	2.74	0.00	0.00	23.61	4.86	2.08	0.00	0.00
ZH	61.84	26.97	6.68	1.01	0.15	20.26	2.61	0.65	0.00	0.00
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
EN	0.05	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00
ES	0.04	0.01	0.01	0.00	0.00	0.04	0.01	0.01	0.00	0.00
FR	0.07	0.01	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.00
RU	0.22	0.00	0.00	0.00	0.00	0.11	0.00	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
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
EN	0.64	0.08	0.01	0.00	0.00	0.67	0.09	0.01	0.00	0.00
ES	0.64	0.08	0.01	0.00	0.00	0.65	0.08	0.01	0.00	0.00
FR	0.68	0.08	0.01	0.00	0.00	0.67	0.09	0.01	0.00	0.00
RU	0.83	0.11	0.01	0.00	0.00	0.47	0.06	0.01	0.00	0.00
ZH	22.86	4.66	0.71	0.10	0.01	1.35	0.17	0.02	0.00	0.00
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 9/53/94/147/234	100.60 ± 59.60 2/49/101/147/274	101.97 ± 59.10 1/51/101/147/300	102.49 ± 58.36 1/53/103/147/300	102.48 ± 58.39 1/53/102/147/300	176.56 ± 101.97 14/98/170/266/432	181.87 ± 108.68 2/88/182/266/501	184.21 ± 107.49 1/90/184/267/551	185.18 ± 106.22 1/95/185/266/552	185.17 ± 106.29 1/95/185/266/561
EN	118.16 ± 68.05 7/64/114/176/256	119.80 ± 71.15 2/58/122/174/288	120.18 ± 70.00 1/60/121/176/299	120.88 ± 69.43 1/62/122/175/300	120.97 ± 69.54 1/62/122/175/300	118.20 ± 68.05 7/64/114/176/256	119.84 ± 71.17 2/58/122/175/288	120.22 ± 70.02 1/60/121/176/299	120.92 ± 69.44 1/62/122/175/300	121.01 ± 69.55 1/62/122/175/303
ES	135.43 ± 77.54 11/80/133/200/286	135.11 ± 79.66 2/65/137/200/300	136.35 ± 79.00 1/67/138/200/300	137.09 ± 78.22 1/70/139/199/300	137.18 ± 78.45 1/70/139/200/300	137.90 ± 78.93 11/81/134/202/288	137.61 ± 81.03 2/65/140/205/305	138.93 ± 80.43 1/69/140/204/309	139.67 ± 79.61 1/71/141/203/311	139.75 ± 79.84 1/71/141/203/318
FR	131.73 ± 75.95 8/74/128/196/289	134.70 ± 80.48 2/64/135/202/300	135.86 ± 79.51 1/65/137/199/300	136.62 ± 78.71 1/69/138/199/300	136.81 ± 78.90 1/69/138/200/300	136.17 ± 78.50 8/77/132/203/296	139.25 ± 83.09 2/65/140/207/313	140.40 ± 82.18 1/68/142/206/315	141.22 ± 81.35 1/71/143/206/322	141.40 ± 81.55 1/71/143/206/325
RU	131.06 ± 76.89 10/71/123/195/271	130.29 ± 77.80 2/62/130/193/299	132.58 ± 77.80 1/65/133/195/300	133.57 ± 77.09 1/67/134/194/300	133.70 ± 77.19 1/67/134/195/300	240.59 ± 144.25 12/125/229/363/505	239.15 ± 144.35 2/113/238/352/558	243.32 ± 144.14 1/117/245/358/565	245.13 ± 142.89 1/122/246/357/566	245.40 ± 143.08 1/122/247/358/569
ZH	33.29 ± 17.41 4/19/32/47/76	33.77 ± 19.77 2/18/33/48/148	34.07 ± 19.50 1/18/34/48/152	34.20 ± 19.21 1/19/34/48/201	34.21 ± 19.26 1/19/34/48/289	91.22 ± 50.88 8/53/90/134/209	92.69 ± 55.28 2/45/92/133/290	93.52 ± 54.29 1/48/93/135/283	94.03 ± 53.67 1/49/94/135/369	94.06 ± 53.73 1/49/94/135/667
ZH_pinyin	113.91 ± 64.46 8/66/109/174/265	115.58 ± 69.18 2/55/115/167/325	116.62 ± 68.13 1/58/117/169/357	117.32 ± 67.41 1/60/118/169/394	117.37 ± 67.50 1/60/118/169/680					
ZH_wubi	100.35 ± 55.70 8/59/99/150/226	101.05 ± 60.13 2/48/100/146/325	101.91 ± 58.94 1/52/102/147/318	102.43 ± 58.21 1/53/102/147/381	102.46 ± 58.25 1/54/103/146/688					
AR_cp1256						97.98 ± 55.47 9/53/94/147/234	100.60 ± 59.60 2/49/101/147/274	101.97 ± 59.10 1/51/101/147/300	102.49 ± 58.36 1/53/103/147/300	102.48 ± 58.39 1/53/102/147/300
RU_cp1251						131.06 ± 76.89 10/71/123/195/271	130.29 ± 77.80 2/62/130/193/299	132.58 ± 77.80 1/65/133/195/300	133.57 ± 77.09 1/67/134/194/300	133.70 ± 77.19 1/67/134/195/300

Representation	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
Number of TYPES										
AR	1,129	6,719	28,222	97,467	311,355	810	4,408	18,800	29,680	29,959
EN	876	4,148	13,660	42,844	140,912	784	3,408	10,756	26,820	29,178
ES	970	4,775	16,735	51,366	154,482	814	3,705	12,529	28,068	29,471
FR	955	4,636	15,639	45,348	132,881	812	3,670	12,258	27,744	29,370
RU	1,142	6,520	25,959	83,937	251,355	993	4,829	18,837	29,261	29,811
ZH	897	4,162	14,714	46,788	140,188	1,068	3,717	11,069	27,409	29,107
Number of TOKENS										
AR	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	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	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	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	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	1,866	18,697	189,773	1,906,704	19,073,133	2,563	21,630	199,028	1,946,543	19,574,572
OOV type rate (%)										
AR	93.61	72.07	35.51	13.07	4.82	7.86	0.30	0.02	0.00	0.00
EN	89.58	61.04	26.28	10.25	4.18	4.00	0.44	0.15	0.04	0.00
ES	90.59	64.27	27.73	9.61	3.95	4.08	0.44	0.17	0.06	0.00
FR	90.23	62.61	25.79	8.71	3.36	3.87	0.47	0.12	0.05	0.00
RU	93.20	69.75	32.20	11.10	3.91	4.39	0.23	0.03	0.01	0.00
ZH	89.73	64.10	29.54	10.44	3.84	67.89	27.00	5.27	0.67	0.04
OOV token rate (%)										
AR	53.94	25.89	9.26	3.14	1.15	0.82	0.08	0.00	0.00	0.00
EN	34.11	12.18	3.50	1.22	0.50	0.30	0.02	0.01	0.01	0.00
ES	31.52	12.57	3.72	1.15	0.46	0.19	0.03	0.04	0.02	0.00
FR	30.96	11.91	3.38	0.99	0.39	0.22	0.05	0.01	0.01	0.00
RU	49.05	22.75	7.55	2.36	0.81	0.71	0.02	0.00	0.00	0.00
ZH	37.81	13.74	4.42	1.42	0.53	22.02	4.99	0.79	0.10	0.00
TTR (%)										
AR	61.86	35.82	14.82	5.09	1.63	23.91	17.47	8.93	1.42	0.14
EN	42.79	19.70	6.47	2.02	0.66	24.03	13.83	4.88	1.24	0.13
ES	40.59	20.01	6.92	2.11	0.64	21.99	13.15	4.95	1.13	0.12
FR	40.43	18.97	6.33	1.82	0.53	22.16	12.86	4.76	1.10	0.12
RU	58.27	33.19	12.97	4.17	1.25	26.41	18.48	8.63	1.35	0.14
ZH	48.07	22.26	7.75	2.45	0.74	41.67	17.18	5.56	1.41	0.15
Mean line length±std 0/25/50/75/100-th										
AR	18.25 ± 10.12 2/10/17/27/43	18.76 ± 11.10 1/9/18/27/73	19.05 ± 10.95 1/10/19/27/84	19.16 ± 10.83 1/10/19/27/143	19.14 ± 10.82 1/10/19/27/129	33.87 ± 18.38 2/20/32/50/73	25.23 ± 15.26 1/13/24/36/96	21.06 ± 12.27 1/11/21/30/95	20.90 ± 12.02 1/11/21/29/146	20.99 ± 12.08 1/11/21/29/165
EN	20.47 ± 11.02 1/12/19/31/42	21.06 ± 12.60 1/11/21/31/74	21.10 ± 12.25 1/11/21/30/86	21.22 ± 12.14 1/11/21/30/117	21.23 ± 12.15 1/11/21/30/127	32.62 ± 17.93 2/19/31/47/81	24.63 ± 14.90 1/13/24/35/85	22.04 ± 12.78 1/11/22/31/89	21.60 ± 12.34 1/12/22/31/117	21.67 ± 12.40 1/12/22/31/129
ES	23.90 ± 13.29 1/14/24/34/49	23.86 ± 14.12 1/12/24/35/80	24.18 ± 13.94 1/12/24/35/87	24.32 ± 13.83 1/13/24/35/119	24.33 ± 13.86 1/13/24/35/130	37.02 ± 20.68 2/22/36/54/81	28.18 ± 16.96 1/14/27/41/93	25.31 ± 14.60 1/13/25/37/91	24.85 ± 14.13 1/13/25/36/119	24.93 ± 14.21 1/13/25/36/133
FR	23.62 ± 13.40 1/13/23/34/53	24.44 ± 14.69 1/12/24/36/78	24.70 ± 14.41 1/12/25/36/92	24.85 ± 14.27 1/13/25/36/118	24.87 ± 14.30 1/13/25/36/132	36.64 ± 21.18 2/21/35/54/83	28.54 ± 17.31 1/15/27/42/94	25.73 ± 14.96 1/13/26/37/97	25.28 ± 14.49 1/13/25/36/118	25.37 ± 14.57 1/13/26/36/134
RU	19.60 ± 10.58 1/12/20/28/43	19.64 ± 11.58 1/10/20/29/72	20.02 ± 11.63 1/10/20/29/85	20.15 ± 11.50 1/11/20/29/117	20.15 ± 11.51 1/11/20/29/128	37.60 ± 21.30 2/21/33/53/91	26.14 ± 15.95 1/13/25/38/89	21.83 ± 12.79 1/11/22/31/87	21.70 ± 12.55 1/11/22/31/120	21.79 ± 12.63 1/11/22/31/162
ZH	18.66 ± 10.28 2/11/18/26/48	18.70 ± 11.18 1/10/18/27/68	18.98 ± 11.05 1/10/19/27/99	19.07 ± 10.88 1/10/19/27/116	19.07 ± 10.90 1/10/19/27/167	25.63 ± 13.65 2/16/24/34/63	21.63 ± 13.19 1/11/21/31/86	19.90 ± 11.61 1/10/20/28/101	19.47 ± 11.13 1/11/19/27/116	19.57 ± 11.24 1/11/19/28/167

Statistics for dataset C

Representation / Number of lines	CHAR 100	CHAR 1,000	CHAR 10,000	CHAR 100,000	CHAR 1,000,000	BYTE 100	BYTE 1,000	BYTE 10,000	BYTE 100,000	BYTE 1,000,000
Number of TYPES										
AR	97	136	162	261	427	92	123	144	160	176
EN	78	98	126	186	310	81	103	131	159	178
ES	84	108	133	186	306	85	112	133	160	175
FR	87	113	139	195	330	91	117	134	159	179
RU	98	141	167	211	287	100	140	154	166	177
ZH	783	1,550	2,443	3,477	4,698	128	152	162	177	189
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
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
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
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
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
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
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
EN	21.65	12.37	8.25	2.06	0.00	22.55	11.76	4.90	0.98	0.00
ES	25.23	14.41	10.81	5.41	0.90	23.64	10.91	6.36	0.91	0.00
FR	23.01	9.73	7.08	1.77	0.00	22.88	9.32	5.08	0.85	0.00
RU	33.56	7.53	2.05	0.00	0.00	31.25	5.56	1.39	0.00	0.00
ZH	61.23	26.87	6.73	1.42	0.35	16.99	3.92	0.65	0.00	0.00
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
EN	0.03	0.00	0.00	0.00	0.00	0.03	0.00	0.00	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
FR	0.14	0.00	0.00	0.00	0.00	0.12	0.00	0.00	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
ZH	11.15	1.25	0.18	0.04	0.01	0.10	0.00	0.00	0.00	0.00
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
EN	0.63	0.08	0.01	0.00	0.00	0.66	0.08	0.01	0.00	0.00
ES	0.59	0.08	0.01	0.00	0.00	0.59	0.08	0.01	0.00	0.00
FR	0.63	0.08	0.01	0.00	0.00	0.64	0.08	0.01	0.00	0.00
RU	0.71	0.10	0.01	0.00	0.00	0.39	0.06	0.01	0.00	0.00
ZH	23.23	4.50	0.72	0.10	0.01	1.39	0.16	0.02	0.00	0.00
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 5/57/96/152/213	104.14 ± 57.83 4/54/106/149/264	102.11 ± 58.06 3/53/102/147/293	102.31 ± 58.33 1/53/102/147/298	102.45 ± 58.46 1/53/102/147/300	184.48 ± 105.32 9/101/173/280/383	188.51 ± 105.63 6/98/191/270/487	184.51 ± 105.78 3/94/184/266/533	184.83 ± 106.15 1/95/184/266/551	185.11 ± 106.40 1/95/185/266/559
EN	123.00 ± 72.30 9/63/116/187/268	123.65 ± 71.12 3/60/128/178/300	120.88 ± 69.62 1/61/122/175/300	120.89 ± 69.58 1/62/122/175/300	120.93 ± 69.60 1/61/122/175/300	123.05 ± 72.35 9/63/116/187/268	123.68 ± 71.15 3/60/128/178/300	120.92 ± 69.64 1/61/122/175/300	120.93 ± 69.60 1/62/122/175/300	120.97 ± 69.62 1/61/122/175/304
ES	141.25 ± 82.83 11/59/134/215/292	139.56 ± 79.59 3/67/145/201/300	136.98 ± 78.35 1/70/138/200/300	137.01 ± 78.45 1/70/138/199/300	137.13 ± 78.47 1/70/139/199/300	143.87 ± 84.37 11/60/136/218/298	142.24 ± 81.05 3/69/149/206/306	139.53 ± 79.72 1/71/140/204/308	139.58 ± 79.84 1/71/141/203/312	139.70 ± 79.87 1/71/141/203/316
FR	137.97 ± 80.12 6/62/133/215/293	138.53 ± 79.49 3/68/141/203/300	136.54 ± 78.80 1/68/137/200/300	136.61 ± 78.92 1/69/138/199/300	136.75 ± 78.96 1/69/138/199/300	142.43 ± 82.97 6/64/137/219/303	143.14 ± 82.18 3/70/145/209/315	141.15 ± 81.47 1/71/142/206/319	141.20 ± 81.56 1/71/142/206/323	141.34 ± 81.61 1/71/143/206/330
RU	137.63 ± 80.80 11/71/130/212/296	136.67 ± 78.77 3/63/139/200/300	133.93 ± 77.25 1/67/135/196/300	133.70 ± 77.26 1/67/134/195/300	133.67 ± 77.25 1/67/134/195/300	253.45 ± 150.90 21/126/242/392/550	251.23 ± 146.21 4/118/258/366/561	245.79 ± 143.31 1/121/247/360/562	245.36 ± 143.18 1/122/247/358/566	245.34 ± 143.17 1/122/247/358/569
ZH	33.71 ± 18.03 4/19/33/48/70	34.43 ± 19.40 2/18/35/49/127	34.15 ± 19.06 1/19/34/48/133	34.20 ± 19.27 1/19/34/48/220	34.20 ± 19.28 1/19/34/48/272	92.35 ± 52.42 12/47/93/132/198	95.10 ± 54.29 3/50/96/138/248	93.88 ± 53.42 3/49/94/135/291	94.04 ± 53.75 1/49/94/135/378	94.05 ± 53.79 1/48/94/135/594
ZH_pinyin	116.05 ± 67.54 11/58/117/167/263	119.26 ± 68.54 3/61/122/173/320	117.16 ± 67.25 3/61/117/169/335	117.32 ± 67.49 1/60/118/169/376	117.36 ± 67.55 1/60/118/169/584					
ZH_wubi	100.71 ± 57.05 13/54/101/139/217	103.94 ± 59.48 3/54/105/150/296	102.29 ± 57.92 3/54/102/147/297	102.46 ± 58.29 1/53/102/146/398	102.45 ± 58.31 1/53/103/146/627					
AR_cp1256						102.12 ± 57.40 5/57/96/152/213	104.14 ± 57.83 4/54/106/149/264	102.11 ± 58.06 3/53/102/147/293	102.31 ± 58.33 1/53/102/147/298	102.45 ± 58.46 1/53/102/147/300
RU_cp1251						137.63 ± 80.80 11/71/130/212/296	136.67 ± 78.77 3/63/139/200/300	133.93 ± 77.25 1/67/135/196/300	133.70 ± 77.26 1/67/134/195/300	133.67 ± 77.25 1/67/134/195/300

Representation / Number of lines	WORD 100	WORD 1,000	WORD 10,000	WORD 100,000	WORD 1,000,000	BPE 100	BPE 1,000	BPE 10,000	BPE 100,000	BPE 1,000,000
Number of TYPES										
AR	1,165	6,823	28,296	97,380	310,632	864	4,530	18,735	29,673	29,954
EN	906	4,173	13,658	42,677	141,163	802	3,484	10,800	26,766	29,215
ES	970	4,734	16,850	51,281	154,649	803	3,686	12,470	28,134	29,470
FR	991	4,603	15,586	45,235	132,493	853	3,693	12,170	27,731	29,355
RU	1,165	6,490	26,112	84,137	250,778	1,002	4,858	19,029	29,322	29,814
ZH	882	4,160	14,730	46,667	139,895	1,084	3,716	11,214	27,443	29,110
Number of TOKENS										
AR	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	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	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	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	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	1,824	19,120	190,485	1,907,643	19,073,463	2,559	22,067	199,706	1,947,300	19,573,372
OOV type rate (%)										
AR	93.72	71.13	35.70	13.44	5.02	6.06	0.34	0.00	0.00	0.00
EN	89.47	60.56	26.27	10.43	4.32	3.52	0.49	0.18	0.04	0.00
ES	90.71	63.96	27.18	9.71	3.88	4.44	0.53	0.19	0.08	0.00
FR	89.92	62.27	25.96	8.89	3.48	3.96	0.49	0.15	0.02	0.00
RU	93.11	69.66	32.15	10.91	3.81	4.82	0.29	0.02	0.01	0.00
ZH	90.34	63.48	29.89	10.31	4.09	67.98	26.22	5.03	0.55	0.10
OOV token rate (%)										
AR	53.40	25.24	9.31	3.25	1.18	0.73	0.12	0.00	0.00	0.00
EN	34.26	11.63	3.41	1.26	0.51	0.20	0.03	0.02	0.00	0.00
ES	32.17	12.05	3.56	1.16	0.46	0.29	0.09	0.04	0.02	0.00
FR	31.39	11.66	3.30	1.03	0.41	0.51	0.04	0.02	0.00	0.00
RU	49.48	22.44	7.53	2.33	0.79	0.38	0.04	0.00	0.00	0.00
ZH	38.65	13.96	4.52	1.42	0.56	22.01	4.75	0.81	0.09	0.01
TTR (%)										
AR	61.67	35.39	14.84	5.09	1.62	24.56	17.69	8.89	1.42	0.14
EN	42.82	19.37	6.44	2.01	0.66	23.81	13.87	4.88	1.24	0.13
ES	39.24	19.24	6.94	2.11	0.64	21.22	12.83	4.91	1.13	0.12
FR	40.40	18.45	6.27	1.82	0.53	22.72	12.78	4.71	1.10	0.12
RU	57.56	31.70	12.95	4.17	1.24	26.28	18.15	8.66	1.35	0.14
ZH	48.36	21.76	7.73	2.45	0.73	42.36	16.84	5.62	1.41	0.15
Mean line length±std 0/25/50/75/100-th										
AR	18.89 ± 10.27 2/11/18/27/43	19.28 ± 10.59 1/10/20/27/50	19.07 ± 10.72 1/10/19/27/82	19.13 ± 10.82 1/10/19/27/120	19.14 ± 10.84 1/10/19/27/143	35.18 ± 18.92 4/18/34/50/73	25.60 ± 14.42 1/14/26/36/79	21.07 ± 12.01 1/11/21/30/86	20.87 ± 12.00 1/11/21/29/123	20.98 ± 12.10 1/11/21/29/147
EN	21.16 ± 12.24 2/10/21/31/48	21.55 ± 12.31 1/11/22/31/55	21.20 ± 12.15 1/11/21/30/83	21.23 ± 12.17 1/11/21/30/118	21.23 ± 12.17 1/11/21/30/136	33.69 ± 19.32 3/15/33/49/78	25.11 ± 14.32 1/13/26/36/86	22.13 ± 12.67 1/12/22/32/88	21.60 ± 12.36 1/12/22/31/120	21.67 ± 12.42 1/12/22/31/138
ES	24.72 ± 14.39 2/12/24/37/56	24.60 ± 13.91 1/12/25/36/60	24.27 ± 13.80 1/13/24/35/84	24.31 ± 13.87 1/13/24/35/119	24.32 ± 13.87 1/13/24/35/139	37.84 ± 21.74 4/17/35/56/86	28.73 ± 16.27 1/14/30/41/90	25.41 ± 14.46 1/13/25/36/90	24.84 ± 14.17 1/13/25/36/121	24.93 ± 14.22 1/13/25/36/140
FR	24.53 ± 13.79 2/13/24/37/50	24.95 ± 14.19 1/12/26/36/62	24.85 ± 14.28 1/13/25/36/84	24.85 ± 14.31 1/13/25/36/118	24.86 ± 14.31 1/13/25/36/138	37.55 ± 20.97 3/20/37/56/80	28.91 ± 16.41 1/14/30/42/86	25.86 ± 14.85 1/13/26/37/88	25.28 ± 14.53 1/13/25/36/120	25.36 ± 14.58 1/13/26/36/138
RU	20.24 ± 11.27 2/11/20/29/47	20.47 ± 11.61 1/10/21/29/55	20.16 ± 11.44 1/11/20/29/83	20.16 ± 11.53 1/11/20/29/117	20.16 ± 11.53 1/11/20/29/136	38.13 ± 21.10 3/18/38/56/81	26.77 ± 15.26 1/14/27/39/78	21.98 ± 12.61 1/12/22/31/87	21.73 ± 12.59 1/11/22/31/121	21.79 ± 12.64 1/11/22/31/138
ZH	18.24 ± 9.97 2/10/17/26/39	19.12 ± 10.87 1/10/20/27/55	19.05 ± 10.86 1/10/19/27/87	19.08 ± 10.91 1/10/19/27/113	19.07 ± 10.92 1/10/19/27/248	25.59 ± 13.84 2/14/26/37/52	22.07 ± 12.68 1/12/22/31/90	19.97 ± 11.43 1/11/20/28/92	19.47 ± 11.17 1/11/19/27/115	19.57 ± 11.26 1/11/19/28/249

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 Number of lines in train set	CHAR	BYTE	WORD	BPE_A					BPE_B					BPE_C				
				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 3/60/110/153/277	196.79 ± 105.85 6/107/199/277/503	19.95 ± 10.58 1/12/20/27/58	54.46 ± 29.51 1/30/54/76/152	37.60 ± 20.51 1/22/37/52/125	26.97 ± 14.80 1/16/27/37/95	24.79 ± 13.59 1/14/24/34/86	22.92 ± 12.56 1/14/22/32/75	48.65 ± 26.35 1/27/49/68/156	31.80 ± 17.76 1/18/32/44/110	23.83 ± 13.15 1/14/24/33/80	22.27 ± 12.12 1/13/22/31/71	22.19 ± 12.07 1/13/22/31/71	48.63 ± 26.21 1/28/49/67/145	31.60 ± 17.53 1/18/31/44/110	23.90 ± 13.20 1/14/24/33/87	22.29 ± 12.14 1/13/22/31/73	22.19 ± 12.06 1/13/22/31/72
EN	127.14 ± 68.64 6/68/130/181/299	127.16 ± 68.65 6/68/130/181/299	21.98 ± 11.81 1/13/22/31/61	50.64 ± 27.44 1/27/51/72/136	33.08 ± 18.07 1/19/33/46/110	25.05 ± 13.56 1/15/25/35/78	23.16 ± 12.42 1/14/23/32/71	22.86 ± 12.27 1/13/23/32/71	45.58 ± 24.86 1/26/46/65/129	29.53 ± 16.30 1/17/29/41/94	23.90 ± 12.95 1/14/24/33/74	22.63 ± 12.13 1/13/23/31/65	22.54 ± 12.06 1/13/23/31/65	45.62 ± 24.67 1/26/46/64/122	29.37 ± 16.13 1/17/29/41/92	23.88 ± 12.91 1/14/24/33/73	22.63 ± 12.12 1/13/23/31/65	22.54 ± 12.05 1/13/23/31/66
ES	144.28 ± 77.78 5/77/145/207/300	146.96 ± 79.21 5/78/148/211/307	25.38 ± 13.56 1/15/25/36/63	55.39 ± 30.13 1/30/56/79/141	37.04 ± 20.13 1/21/37/52/111	28.81 ± 15.53 1/17/29/41/85	26.74 ± 14.29 1/15/27/38/76	26.44 ± 14.14 1/15/26/37/73	50.60 ± 27.66 1/27/51/72/133	33.72 ± 18.48 1/19/34/47/103	27.60 ± 14.88 1/15/28/39/75	26.18 ± 13.98 1/15/26/37/71	26.11 ± 13.94 1/15/26/37/70	50.29 ± 27.27 1/28/51/71/133	33.53 ± 18.38 1/19/34/47/95	27.58 ± 14.86 1/15/28/39/75	26.19 ± 13.99 1/15/26/37/71	26.12 ± 13.94 1/15/26/37/71
FR	142.37 ± 77.81 4/74/145/205/300	147.08 ± 80.30 4/77/150/212/310	25.59 ± 13.86 1/14/26/36/66	54.04 ± 29.55 1/29/55/76/139	37.27 ± 20.65 1/20/38/52/119	28.84 ± 15.76 1/16/29/40/89	26.79 ± 14.47 1/15/27/37/74	26.51 ± 14.35 1/15/27/37/70	50.78 ± 27.79 1/27/52/72/135	33.82 ± 18.74 1/18/34/47/108	27.66 ± 15.11 1/15/28/39/80	26.27 ± 14.20 1/14/27/37/69	26.22 ± 14.16 1/14/27/37/69	49.97 ± 27.25 1/27/51/70/131	33.83 ± 18.80 1/18/34/47/107	27.66 ± 15.09 1/15/28/39/84	26.30 ± 14.21 1/14/27/37/69	26.21 ± 14.16 1/14/27/37/69
RU	140.25 ± 76.21 5/75/141/200/300	257.79 ± 141.53 7/136/259/370/569	20.86 ± 11.25 1/12/21/29/69	57.79 ± 32.12 1/31/58/82/185	36.93 ± 20.73 1/21/36/52/185	25.92 ± 14.48 1/15/26/36/128	23.56 ± 12.95 1/14/23/33/101	23.10 ± 12.71 1/13/23/32/95	53.08 ± 29.00 1/29/53/75/185	32.59 ± 18.42 1/18/32/46/166	24.39 ± 13.59 1/14/24/34/112	22.81 ± 12.50 1/13/23/32/85	22.75 ± 12.46 1/13/23/32/93	53.24 ± 29.18 2/30/53/76/185	32.49 ± 18.49 1/18/32/45/161	24.41 ± 13.57 1/14/24/34/105	22.81 ± 12.50 1/13/23/32/90	22.74 ± 12.45 1/13/23/32/92
ZH	35.10 ± 18.48 2/21/35/49/125	97.85 ± 52.10 4/53/99/138/268	19.50 ± 10.42 1/11/20/27/64	31.44 ± 16.96 1/18/31/44/109	26.07 ± 14.19 1/15/26/36/93	22.14 ± 12.02 1/13/22/31/80	20.67 ± 11.16 1/12/21/29/74	20.41 ± 11.05 1/12/20/28/74	30.48 ± 16.47 1/17/30/43/104	24.58 ± 13.42 1/14/24/34/92	21.19 ± 11.49 1/12/21/29/71	20.11 ± 10.79 1/12/20/28/64	20.09 ± 10.79 1/12/20/28/66	30.59 ± 16.45 1/18/30/43/101	24.61 ± 13.49 1/14/24/34/93	21.19 ± 11.49 1/12/21/29/76	20.12 ± 10.81 1/12/20/28/64	20.09 ± 10.79 1/12/20/28/65
ZH_pinyin	122.52 ± 65.59 6/67/125/173/353																	
ZH_wubi	107.49 ± 56.73 4/60/108/151/294																	
AR_cp1256		108.66 ± 58.01 3/60/110/153/277																
RU_cp1251		140.25 ± 76.21 5/75/141/200/300																

E SCORE TABLES

Number of bits to encode the dev data for each of the 30 language directions. Shown is mean±std over:

- 12 runs for CHAR, BYTE, and WORD from 100 to 100,000 lines,
- 5 runs for 1,000,000 lines, and
- 3 runs for all sizes involving alternate representations (BPE, Pinyin, Wubi, cp 1256 and cp 1251).

LG DIR	CHAR 100	CHAR 1,000	CHAR 10,000	CHAR 100,000	CHAR 1,000,000
AR-EN	1,544,736.23 ±33,345.46	1,121,015.03 ±33,586.51	760,732.50 ±51,886.35	765,028.72 ±65,502.62	770,924.52 ±56,622.95
AR-ES	1,669,230.67 ±33,827.05	1,182,122.14 ±38,918.23	790,296.63 ±30,536.00	812,541.59 ±52,705.76	801,797.31 ±32,089.03
AR-FR	1,674,603.44 ±38,229.92	1,182,589.30 ±36,913.53	791,660.51 ±51,782.45	787,110.43 ±30,306.43	799,124.88 ±33,696.76
AR-RU	1,839,654.03 ±77,843.20	1,354,401.84 ±93,525.81	848,031.25 ±45,311.41	869,551.01 ±50,929.85	869,612.93 ±112,969.06
AR-ZH	1,280,817.03 ±83,683.91	1,281,197.26 ±58,574.87	1,132,398.00 ±33,188.89	1,120,681.62 ±38,495.77	1,121,645.99 ±30,840.47
EN-AR	1,510,740.01 ±65,375.89	1,219,914.71 ±120,433.84	840,090.49 ±55,048.71	851,269.34 ±51,874.44	835,292.90 ±47,718.68
EN-ES	1,681,165.45 ±48,333.33	1,183,388.65 ±32,316.44	810,007.62 ±71,162.44	844,186.06 ±88,946.74	965,942.27 ±106,264.42
EN-FR	1,686,664.41 ±45,780.11	1,178,168.83 ±36,672.34	819,628.02 ±82,055.42	834,537.84 ±71,529.51	863,864.62 ±91,124.27
EN-RU	1,800,285.21 ±29,925.05	1,322,698.36 ±35,269.32	905,404.66 ±85,669.60	877,356.92 ±65,407.02	877,611.01 ±64,609.22
EN-ZH	1,226,660.68 ±94,088.76	1,294,737.73 ±39,344.17	1,129,167.14 ±30,676.13	1,157,019.45 ±28,965.79	1,131,332.86 ±17,992.05
ES-AR	1,502,118.26 ±59,531.42	1,225,761.43 ±118,947.77	855,696.37 ±53,179.44	848,447.61 ±54,355.00	780,375.16 ±6,622.94
ES-EN	1,528,879.47 ±17,380.08	1,125,929.72 ±33,012.11	794,613.91 ±75,558.20	814,325.63 ±82,452.93	822,495.30 ±90,357.69
ES-FR	1,671,931.04 ±27,730.35	1,185,237.77 ±33,338.32	792,016.62 ±73,058.08	801,625.47 ±38,060.73	827,406.55 ±58,343.17
ES-RU	1,802,953.02 ±36,905.60	1,316,280.08 ±47,934.44	1,083,050.54 ±582,627.72	961,497.33 ±118,556.02	934,228.95 ±53,202.75
ES-ZH	1,232,611.05 ±82,270.75	1,312,046.07 ±88,361.91	1,119,023.46 ±45,898.67	1,124,896.58 ±31,824.16	1,127,969.47 ±23,414.96
FR-AR	1,493,998.54 ±73,487.06	1,213,257.69 ±129,450.70	858,980.22 ±27,406.28	853,556.67 ±52,803.51	827,446.97 ±74,995.89
FR-EN	1,539,331.82 ±30,645.00	1,112,235.95 ±39,703.27	770,130.63 ±45,619.46	798,443.60 ±77,432.73	814,711.40 ±73,631.19
FR-ES	1,660,503.46 ±28,332.32	1,202,263.37 ±32,925.18	809,042.36 ±60,412.82	848,666.85 ±105,934.45	955,334.11 ±128,210.24
FR-RU	1,801,787.81 ±35,248.27	1,338,302.47 ±42,609.76	883,277.92 ±70,791.84	875,546.37 ±56,524.21	896,296.70 ±64,751.37
FR-ZH	1,276,350.99 ±67,068.64	1,283,583.70 ±52,962.76	1,128,361.74 ±33,735.96	1,126,473.28 ±55,610.45	1,110,946.65 ±48,488.07
RU-AR	1,514,660.49 ±80,209.43	1,238,154.76 ±116,146.83	854,605.61 ±56,760.19	848,816.02 ±68,333.29	782,218.32 ±15,171.38
RU-EN	1,541,284.19 ±30,975.35	1,099,124.56 ±31,720.92	786,171.98 ±54,777.83	796,825.08 ±68,295.54	838,299.51 ±97,298.93
RU-ES	1,653,035.53 ±21,799.04	1,186,609.00 ±33,265.12	836,904.99 ±92,154.88	844,107.64 ±69,917.63	870,993.17 ±90,756.24
RU-FR	1,671,911.24 ±31,514.53	1,148,237.21 ±30,323.20	813,638.02 ±81,037.73	826,770.83 ±77,272.11	854,381.90 ±47,166.38
RU-ZH	1,248,287.61 ±81,835.09	1,337,592.90 ±89,849.94	1,137,380.32 ±44,231.23	1,139,940.12 ±44,106.75	1,127,306.95 ±26,305.47
ZH-AR	1,520,133.55 ±83,307.06	1,251,844.54 ±119,331.91	901,003.65 ±16,826.04	874,419.36 ±49,241.53	849,667.86 ±55,537.72
ZH-EN	1,562,210.25 ±15,272.79	1,168,166.65 ±42,453.69	814,223.17 ±50,875.39	770,705.16 ±30,958.09	779,540.82 ±54,173.29
ZH-ES	1,669,628.69 ±32,990.23	1,257,703.22 ±45,720.33	873,579.22 ±50,836.73	805,216.88 ±55,927.91	767,172.77 ±29,211.85
ZH-FR	1,695,484.34 ±27,107.05	1,248,595.78 ±48,321.79	847,461.91 ±41,077.17	766,047.76 ±41,162.85	722,938.10 ±32,182.37
ZH-RU	1,813,830.27 ±25,062.52	1,381,636.93 ±46,711.35	922,328.27 ±41,901.72	825,878.06 ±47,665.75	793,500.11 ±44,128.61

LG DIR	BYTE 100	BYTE 1,000	BYTE 10,000	BYTE 100,000	BYTE 1,000,000
AR-EN	1,526,255.58 ±20,035.54	1,129,961.64 ±28,268.82	743,567.36 ±51,321.62	711,353.65 ±57,220.79	689,639.79 ±21,842.78
AR-ES	1,682,368.21 ±49,057.87	1,214,712.80 ±31,341.49	768,329.55 ±58,935.67	773,077.60 ±65,376.92	743,187.12 ±44,315.67
AR-FR	1,677,201.07 ±27,985.35	1,195,378.25 ±29,518.16	749,358.45 ±17,860.40	729,881.05 ±61,847.12	705,375.66 ±18,364.57
AR-RU	2,197,541.98 ±118,172.43	2,228,319.17 ±315,190.29	2,431,853.83 ±734,531.48	1,947,404.65 ±501,380.34	1,829,839.51 ±575,069.74
AR-ZH	1,568,743.13 ±54,845.33	1,201,566.37 ±116,203.87	769,183.30 ±29,222.58	752,991.20 ±31,759.12	745,264.85 ±11,916.25
EN-AR	1,612,887.02 ±36,858.63	1,411,047.84 ±93,468.28	1,376,211.98 ±360,810.34	1,100,958.99 ±272,393.50	905,275.82 ±52,591.22
EN-ES	1,698,584.22 ±63,942.73	1,172,736.58 ±26,544.61	812,042.42 ±49,375.34	836,078.21 ±91,257.68	862,362.78 ±41,834.99
EN-FR	1,683,971.89 ±31,327.55	1,167,324.39 ±44,210.75	791,437.92 ±56,977.42	849,888.50 ±80,981.37	897,983.76 ±108,432.81
EN-RU	2,145,929.38 ±103,021.89	2,056,962.26 ±312,126.19	2,433,405.41 ±403,385.28	2,082,168.25 ±507,789.38	2,057,326.47 ±786,351.50
EN-ZH	1,694,305.78 ±89,612.62	1,246,027.07 ±146,988.77	821,592.74 ±48,392.46	808,848.95 ±29,206.91	817,737.15 ±20,401.29
ES-AR	1,605,955.78 ±62,098.79	1,490,411.37 ±120,316.50	1,161,601.70 ±118,816.27	1,014,994.24 ±251,491.55	970,076.35 ±211,495.90
ES-EN	1,536,479.43 ±19,856.01	1,123,311.02 ±28,408.31	796,207.81 ±63,756.37	786,442.26 ±69,696.47	799,490.21 ±36,829.85
ES-FR	1,677,699.91 ±23,982.36	1,177,374.22 ±40,721.13	781,253.45 ±49,925.27	815,425.77 ±63,815.54	781,399.85 ±22,542.74
ES-RU	2,182,142.73 ±135,719.75	2,302,608.53 ±330,863.07	2,615,505.97 ±607,001.26	2,408,379.29 ±436,157.86	1,719,144.46 ±588,839.57
ES-ZH	1,664,833.83 ±77,412.09	1,224,859.94 ±131,833.28	808,765.40 ±29,211.12	801,317.28 ±24,022.28	812,927.38 ±26,410.04
FR-AR	1,610,076.62 ±58,763.74	1,381,524.21 ±63,989.06	1,172,428.64 ±149,066.89	993,944.78 ±149,970.87	799,319.48 ±46,367.05
FR-EN	1,528,759.79 ±25,574.77	1,116,278.27 ±31,373.53	765,871.86 ±63,821.27	758,813.76 ±56,822.33	781,520.62 ±89,958.57
FR-ES	1,660,208.47 ±29,354.33	1,163,346.80 ±43,432.14	796,585.17 ±62,756.79	818,921.73 ±88,147.45	912,707.81 ±67,643.78
FR-RU	2,228,537.88 ±160,434.99	2,121,931.71 ±336,980.35	2,180,180.77 ±656,345.58	2,296,574.43 ±436,772.45	2,372,214.02 ±205,878.26
FR-ZH	1,682,167.79 ±120,999.47	1,232,599.06 ±140,833.43	795,219.46 ±27,723.08	789,061.67 ±31,881.81	819,523.77 ±11,493.76
RU-AR	1,659,403.19 ±96,866.48	1,634,298.58 ±370,634.47	1,389,921.86 ±385,005.35	1,209,265.05 ±497,397.12	1,033,769.40 ±159,220.24
RU-EN	1,538,863.25 ±27,229.71	1,158,053.12 ±42,767.55	764,175.77 ±63,298.64	782,666.90 ±75,348.29	757,005.62 ±82,551.42
RU-ES	1,660,066.45 ±35,196.89	1,235,809.74 ±33,713.11	784,414.71 ±48,871.45	842,104.27 ±117,598.95	855,656.79 ±81,889.11
RU-FR	1,674,970.54 ±31,577.29	1,187,457.17 ±26,923.73	764,864.49 ±23,061.64	787,983.47 ±60,116.15	758,333.45 ±56,654.08
RU-ZH	1,571,141.84 ±77,145.40	1,194,344.98 ±113,430.97	792,907.70 ±24,982.71	795,999.26 ±37,662.58	802,600.33 ±34,068.53
ZH-AR	1,624,209.85 ±42,159.98	1,457,744.80 ±141,303.10	1,238,075.32 ±204,195.12	1,017,971.85 ±137,533.76	892,807.05 ±100,570.10
ZH-EN	1,551,513.48 ±28,726.75	1,141,688.29 ±43,583.83	798,052.31 ±75,530.17	762,011.92 ±54,989.05	809,986.05 ±54,007.86
ZH-ES	1,671,717.60 ±44,963.44	1,207,194.98 ±30,309.28	827,037.59 ±50,529.76	815,660.85 ±52,744.73	844,427.40 ±75,201.08
ZH-FR	1,697,714.49 ±23,205.94	1,191,418.13 ±46,529.35	815,463.30 ±69,139.02	808,388.78 ±45,459.67	827,621.80 ±39,370.89
ZH-RU	2,263,112.04 ±167,356.23	2,332,008.35 ±298,747.54	2,041,371.91 ±524,829.59	2,190,263.98 ±467,425.12	1,531,660.29 ±526,243.84

LG DIR	WORD 100	WORD 1,000	WORD 10,000	WORD 100,000	WORD 1,000,000
AR-EN	824,021.66 ±16,664.29	919,947.50 ±57,452.07	749,970.52 ±27,184.31	710,251.48 ±20,148.97	700,341.00 ±8,853.30
AR-ES	898,467.63 ±24,959.67	1,039,423.53 ±84,956.04	831,131.85 ±27,013.27	792,199.69 ±10,468.31	788,770.52 ±9,821.52
AR-FR	926,213.05 ±31,620.32	1,023,048.10 ±70,043.69	859,552.52 ±33,915.72	819,669.80 ±15,239.71	812,027.11 ±9,823.48
AR-RU	918,172.29 ±38,438.63	1,110,884.62 ±79,972.05	831,112.97 ±35,373.49	788,608.20 ±14,059.96	777,418.25 ±8,180.38
AR-ZH	789,140.01 ±41,101.53	942,441.68 ±67,255.20	729,612.42 ±39,347.58	676,409.89 ±11,966.22	673,012.11 ±10,276.14
EN-AR	924,898.31 ±42,745.80	1,192,684.90 ±110,700.20	868,140.02 ±55,331.86	813,874.25 ±34,472.49	788,414.28 ±11,428.26
EN-ES	887,045.15 ±21,462.75	1,051,451.41 ±80,381.73	829,470.61 ±21,121.79	790,231.87 ±10,828.77	786,306.07 ±7,241.92
EN-FR	934,050.87 ±17,022.63	1,035,714.99 ±84,847.57	851,171.41 ±27,442.34	814,975.89 ±15,513.71	808,886.69 ±8,763.55
EN-RU	885,403.30 ±27,223.84	1,109,220.75 ±77,956.99	825,785.76 ±26,464.92	788,325.99 ±16,717.11	776,041.11 ±10,251.12
EN-ZH	785,138.74 ±37,169.41	954,362.82 ±68,065.90	729,739.64 ±41,640.71	681,291.82 ±20,230.39	672,100.12 ±10,946.81
ES-AR	916,526.69 ±43,961.48	1,197,624.01 ±105,778.66	857,489.43 ±47,566.38	814,208.10 ±32,471.60	787,750.61 ±13,980.23
ES-EN	805,578.60 ±21,817.93	892,368.72 ±49,290.59	738,870.42 ±28,774.80	704,025.54 ±9,889.00	699,448.80 ±7,276.77
ES-FR	930,115.35 ±22,116.82	1,004,960.20 ±59,888.08	850,753.28 ±28,585.58	821,771.76 ±26,797.43	807,950.92 ±9,117.70
ES-RU	913,399.94 ±41,650.78	1,095,113.59 ±80,209.25	826,919.25 ±32,666.68	786,955.75 ±15,361.23	776,362.37 ±9,475.61
ES-ZH	790,313.04 ±28,023.63	954,665.23 ±73,411.42	730,068.53 ±34,568.02	679,049.77 ±12,124.26	673,434.94 ±9,415.99
FR-AR	929,013.51 ±48,067.94	1,200,779.58 ±108,925.07	863,154.85 ±50,040.02	812,918.99 ±32,030.29	794,825.62 ±12,341.00
FR-EN	814,124.76 ±28,993.16	915,392.27 ±46,743.33	751,116.43 ±38,848.58	704,000.97 ±10,860.95	698,360.27 ±7,404.34
FR-ES	889,058.36 ±26,451.47	1,038,679.60 ±64,167.38	826,599.83 ±29,086.15	789,413.25 ±10,453.66	784,658.89 ±8,168.89
FR-RU	904,743.86 ±64,881.30	1,093,309.95 ±72,982.62	835,428.69 ±31,260.71	786,984.31 ±15,909.24	774,883.56 ±7,098.45
FR-ZH	797,170.14 ±27,567.81	953,202.69 ±64,725.30	727,208.97 ±38,176.47	680,361.44 ±20,563.36	673,978.72 ±13,365.69
RU-AR	924,303.45 ±47,592.08	1,193,233.65 ±106,075.39	860,103.40 ±52,037.99	813,816.48 ±34,050.28	788,969.01 ±15,502.48
RU-EN	794,196.11 ±23,425.16	924,349.04 ±51,003.38	749,955.35 ±33,221.45	707,043.65 ±12,423.97	701,154.52 ±7,645.50
RU-ES	896,072.52 ±23,048.98	1,065,327.27 ±76,375.04	829,514.31 ±29,979.01	793,461.61 ±10,224.66	786,102.56 ±7,027.46
RU-FR	914,593.98 ±28,944.46	1,024,369.02 ±72,855.37	859,692.20 ±27,817.90	819,751.33 ±14,198.54	813,337.85 ±9,244.53
RU-ZH	791,737.80 ±27,003.01	938,418.28 ±71,392.01	725,715.65 ±29,584.88	676,971.77 ±12,435.61	671,004.64 ±10,250.36
ZH-AR	915,209.06 ±32,380.96	1,193,426.50 ±100,933.92	867,644.69 ±61,385.13	814,599.04 ±33,901.66	788,133.05 ±14,247.79
ZH-EN	821,944.58 ±25,378.46	898,782.12 ±35,481.92	739,595.70 ±26,988.74	705,837.56 ±10,692.02	699,272.86 ±9,417.54
ZH-ES	897,303.52 ±36,860.29	1,064,057.78 ±71,631.03	835,422.15 ±31,104.35	792,767.51 ±10,798.32	790,344.66 ±9,003.49
ZH-FR	928,038.15 ±30,193.21	1,028,004.46 ±65,007.37	854,656.55 ±26,572.39	818,513.53 ±14,201.03	814,320.54 ±13,277.16
ZH-RU	898,116.31 ±43,471.25	1,090,957.68 ±73,782.63	831,154.56 ±32,627.39	785,884.85 ±14,883.92	777,642.70 ±8,004.77

LG DIR	BPE 100	BPE 1,000	BPE 10,000	BPE 100,000	BPE 1,000,000
AR-EN	1,301,665.51 ±82,857.86	996,047.63 ±74,175.08	817,170.91 ±70,705.51	734,443.61 ±18,626.37	724,413.83 ±11,671.49
AR-ES	1,415,204.23 ±81,544.75	1,169,086.06 ±34,798.65	889,203.72 ±49,722.10	822,286.95 ±22,391.54	818,450.34 ±11,976.63
AR-FR	1,425,834.11 ±76,674.62	1,173,956.40 ±108,448.13	925,854.48 ±66,008.74	844,751.77 ±24,413.23	835,802.11 ±12,815.20
AR-RU	1,587,521.54 ±76,564.12	1,327,483.08 ±247,187.63	929,790.50 ±54,954.78	860,789.46 ±27,324.96	852,520.24 ±10,909.30
AR-ZH	1,215,900.34 ±93,970.60	1,090,617.74 ±164,499.46	784,579.94 ±63,112.68	697,066.73 ±24,073.74	697,866.07 ±21,932.07
EN-AR	1,419,275.88 ±124,341.43	1,361,398.11 ±252,343.74	944,554.15 ±98,089.19	903,941.69 ±63,710.37	862,162.92 ±24,882.58
EN-ES	1,402,311.29 ±72,607.56	1,216,699.14 ±109,807.25	874,802.00 ±37,226.63	821,171.75 ±19,088.59	816,403.05 ±11,016.65
EN-FR	1,435,857.04 ±84,872.44	1,194,782.82 ±209,676.17	923,240.09 ±77,036.43	840,383.94 ±22,776.12	830,788.81 ±19,227.13
EN-RU	1,588,098.71 ±74,670.88	1,294,095.36 ±239,512.03	921,079.54 ±52,414.55	859,648.14 ±24,907.38	845,655.45 ±5,376.96
EN-ZH	1,203,801.62 ±65,261.40	1,100,747.41 ±136,366.71	802,543.68 ±45,927.74	700,991.91 ±22,850.33	695,438.33 ±11,291.51
ES-AR	1,401,583.29 ±109,916.84	1,288,013.56 ±218,310.60	942,437.09 ±93,254.70	890,564.90 ±74,417.84	862,240.60 ±24,404.85
ES-EN	1,290,921.22 ±76,027.66	1,059,725.25 ±44,181.19	824,122.71 ±63,609.75	729,224.37 ±20,354.82	723,241.13 ±12,076.16
ES-FR	1,436,026.95 ±79,889.54	1,212,957.04 ±206,011.41	943,307.74 ±74,036.72	839,985.47 ±22,650.13	839,760.81 ±19,334.38
ES-RU	1,589,951.85 ±82,913.65	1,287,241.22 ±235,002.06	925,282.41 ±57,152.55	863,856.53 ±27,797.33	852,106.52 ±9,433.15
ES-ZH	1,211,610.91 ±78,154.97	1,105,851.14 ±147,269.98	794,327.21 ±54,849.25	700,122.25 ±31,801.70	692,539.65 ±13,102.53
FR-AR	1,400,857.45 ±109,342.89	1,352,502.41 ±215,335.19	946,306.64 ±93,590.10	894,209.96 ±73,243.54	862,815.27 ±22,818.24
FR-EN	1,295,095.24 ±77,110.80	1,132,324.91 ±130,607.50	818,004.33 ±64,478.34	731,459.64 ±25,711.83	721,742.61 ±11,056.08
FR-ES	1,404,135.74 ±75,580.97	1,247,827.63 ±171,052.39	882,779.24 ±58,463.42	821,925.17 ±18,151.49	815,220.86 ±10,261.57
FR-RU	1,597,346.90 ±84,430.89	1,278,350.91 ±247,342.95	927,393.74 ±60,727.51	860,679.74 ±26,941.22	851,834.07 ±11,582.41
FR-ZH	1,188,320.74 ±71,097.38	996,717.44 ±118,260.78	798,048.58 ±47,630.61	698,049.53 ±22,478.41	692,642.52 ±13,709.21
RU-AR	1,414,105.57 ±103,238.28	1,324,315.20 ±281,897.95	946,345.19 ±93,731.54	893,725.53 ±76,923.49	860,822.73 ±24,195.80
RU-EN	1,299,444.10 ±76,572.12	1,018,148.81 ±97,567.11	811,614.42 ±70,685.80	731,687.21 ±19,913.12	723,716.51 ±10,210.85
RU-ES	1,404,521.63 ±73,696.49	1,181,840.49 ±220,298.17	905,806.89 ±83,428.87	824,187.74 ±18,836.41	816,919.61 ±13,536.25
RU-FR	1,428,278.92 ±64,957.77	1,209,074.62 ±41,132.76	936,867.26 ±56,628.76	840,765.99 ±17,096.97	837,647.05 ±14,732.39
RU-ZH	1,170,310.61 ±29,290.24	1,083,295.36 ±136,602.30	792,170.54 ±52,238.92	698,538.35 ±18,709.40	692,610.20 ±11,836.86
ZH-AR	1,420,118.83 ±124,896.73	1,387,883.51 ±246,916.77	947,595.38 ±95,207.49	891,411.12 ±75,378.14	858,110.42 ±20,957.97
ZH-EN	1,307,175.00 ±91,659.81	1,009,106.88 ±64,100.34	810,542.12 ±53,309.75	732,930.41 ±23,107.23	720,827.32 ±10,243.06
ZH-ES	1,413,771.04 ±89,173.36	1,219,644.77 ±144,552.08	873,855.62 ±38,175.84	823,366.50 ±18,577.96	817,536.03 ±11,550.00
ZH-FR	1,438,493.68 ±77,039.61	1,201,853.73 ±137,343.07	915,603.53 ±54,104.68	843,513.45 ±21,707.51	834,857.87 ±14,234.13
ZH-RU	1,598,147.95 ±94,023.22	1,270,765.18 ±262,527.08	925,864.53 ±50,845.80	860,678.52 ±27,516.26	849,359.58 ±8,882.60

	100	1,000	10,000	100,000	1,000,000
AR-ZH_pinyin	1,244,719.00 ±19,404.89	964,206.32 ±62,447.97	844,187.85 ±33,777.37	847,890.69 ±37,172.75	843,679.52 ±46,688.28
EN-ZH_pinyin	1,274,326.38 ±6,446.73	935,634.70 ±63,139.67	897,991.31 ±89,866.11	848,401.47 ±34,886.12	844,440.43 ±31,096.20
ES-ZH_pinyin	1,223,892.67 ±24,286.37	947,071.23 ±71,669.78	844,416.88 ±31,479.07	902,142.51 ±61,295.27	842,915.10 ±32,926.82
FR-ZH_pinyin	1,237,645.51 ±40,668.83	940,696.08 ±64,686.93	869,345.88 ±15,272.39	880,928.85 ±47,130.09	822,583.48 ±54,568.18
RU-ZH_pinyin	1,228,549.67 ±39,355.53	957,208.34 ±91,715.81	874,649.15 ±46,176.22	855,246.78 ±26,953.55	827,620.65 ±35,329.25
AR-ZH_wubi	1,405,265.49 ±24,177.93	1,573,523.54 ±159,519.24	1,478,060.70 ±200,125.36	1,534,523.85 ±105,807.52	1,566,862.14 ±77,396.40
EN-ZH_wubi	1,435,305.55 ±52,171.91	1,570,871.90 ±100,392.45	1,542,564.62 ±155,534.95	1,621,805.89 ±110,098.10	1,541,417.82 ±135,453.17
ES-ZH_wubi	1,418,676.31 ±10,094.82	1,690,040.92 ±452,306.14	1,368,251.22 ±15,457.44	1,601,982.51 ±56,511.54	1,575,431.36 ±177,731.27
FR-ZH_wubi	1,439,945.68 ±41,647.38	1,499,277.03 ±147,249.10	1,466,536.18 ±73,655.57	1,528,845.79 ±164,414.47	1,500,804.94 ±164,599.21
RU-ZH_wubi	1,430,659.37 ±56,290.04	1,554,048.88 ±144,200.65	1,452,249.13 ±50,514.38	1,484,555.96 ±191,429.18	1,473,114.58 ±183,723.10
EN-AR_cp1256	1,465,480.73 ±50,492.46	1,126,423.78 ±137,598.58	819,198.83 ±39,191.91	895,139.90 ±47,555.78	847,425.26 ±54,355.60
ES-AR_cp1256	1,467,064.48 ±68,719.73	1,103,860.65 ±157,000.70	822,523.19 ±48,108.48	865,080.06 ±85,901.58	807,204.25 ±13,669.73
FR-AR_cp1256	1,453,850.33 ±71,978.45	1,107,306.50 ±164,397.01	811,412.78 ±50,296.08	835,193.59 ±35,489.11	818,608.84 ±23,676.94
RU-AR_cp1256	1,443,181.79 ±59,366.24	1,166,915.22 ±118,862.54	789,271.03 ±88,802.25	774,284.66 ±80,721.58	796,788.17 ±70,399.46
ZH-AR_cp1256	1,449,043.78 ±60,412.59	1,142,324.09 ±154,618.66	857,447.94 ±49,944.59	804,967.70 ±49,661.83	778,655.50 ±6,406.73
AR-RU_cp1251	2,033,848.35 ±166,696.08	1,433,360.25 ±103,744.53	870,350.56 ±104,241.98	789,482.91 ±3,272.59	814,319.13 ±59,519.66
EN-RU_cp1251	1,805,622.90 ±29,412.48	1,303,158.16 ±40,220.95	866,878.02 ±50,482.79	916,441.19 ±57,277.46	1,008,644.13 ±121,622.34
ES-RU_cp1251	1,790,425.65 ±33,818.40	1,386,700.30 ±145,629.15	1,018,909.55 ±189,934.38	865,520.61 ±43,694.66	1,270,142.14 ±635,645.01
FR-RU_cp1251	1,786,597.29 ±31,551.71	1,370,522.14 ±68,248.46	891,537.51 ±76,492.28	903,389.38 ±84,856.48	955,577.80 ±116,158.01
ZH-RU_cp1251	1,786,634.07 ±21,326.49	1,343,079.46 ±42,128.54	894,879.79 ±66,548.00	905,825.07 ±52,282.91	858,440.41 ±21,368.00
AR_cp1256-RU_cp1251	1,824,623.51 ±56,859.77	1,372,546.79 ±125,115.31	882,342.43 ±54,341.13	883,493.70 ±95,577.24	935,969.31 ±98,765.65
RU_cp1251-AR_cp1256	1,457,056.24 ±61,387.32	1,150,943.30 ±124,707.99	856,651.67 ±24,159.92	800,129.52 ±48,924.14	860,462.66 ±80,105.83

F 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	ARRU_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.

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

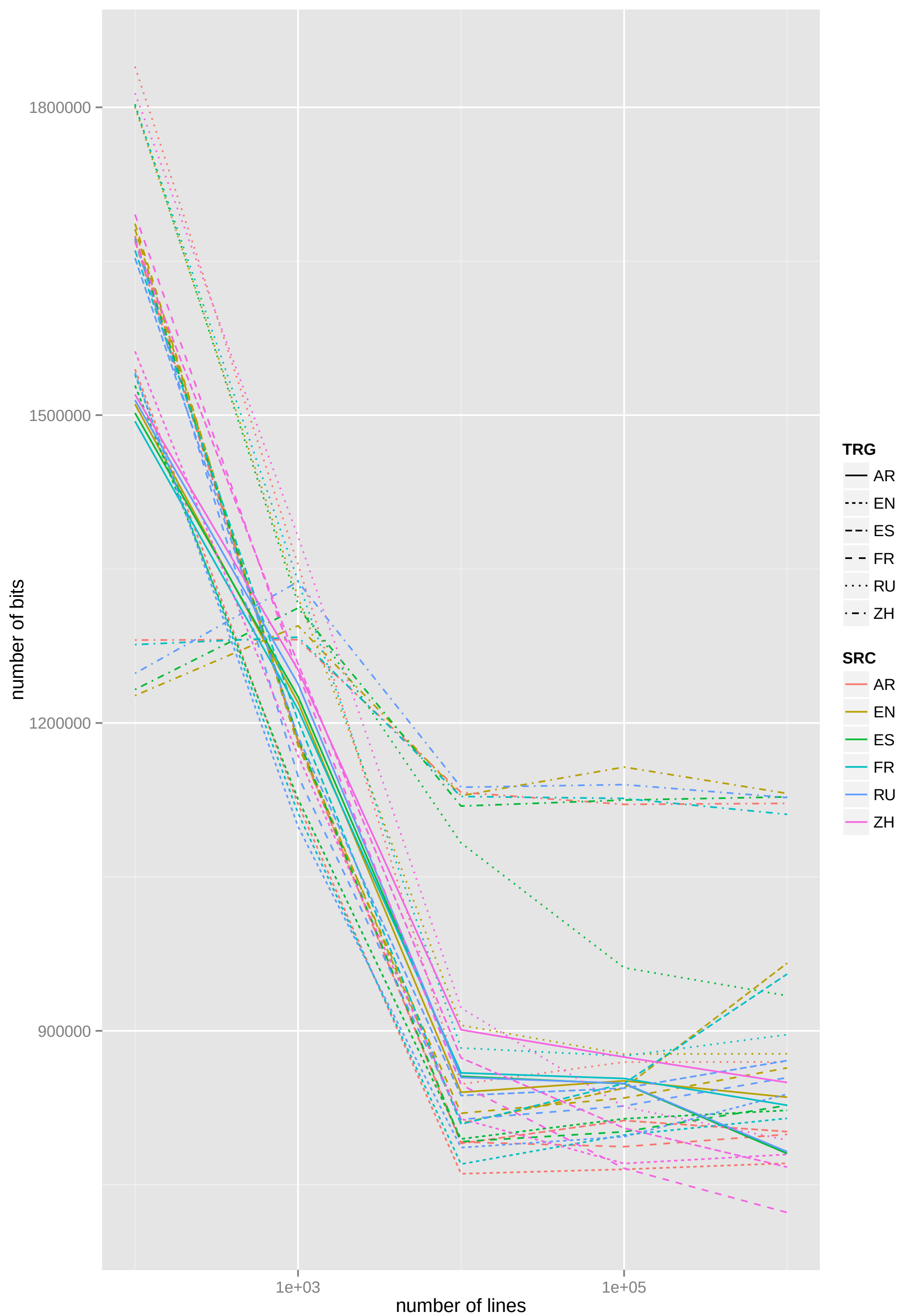


Figure 5: CHAR: character models

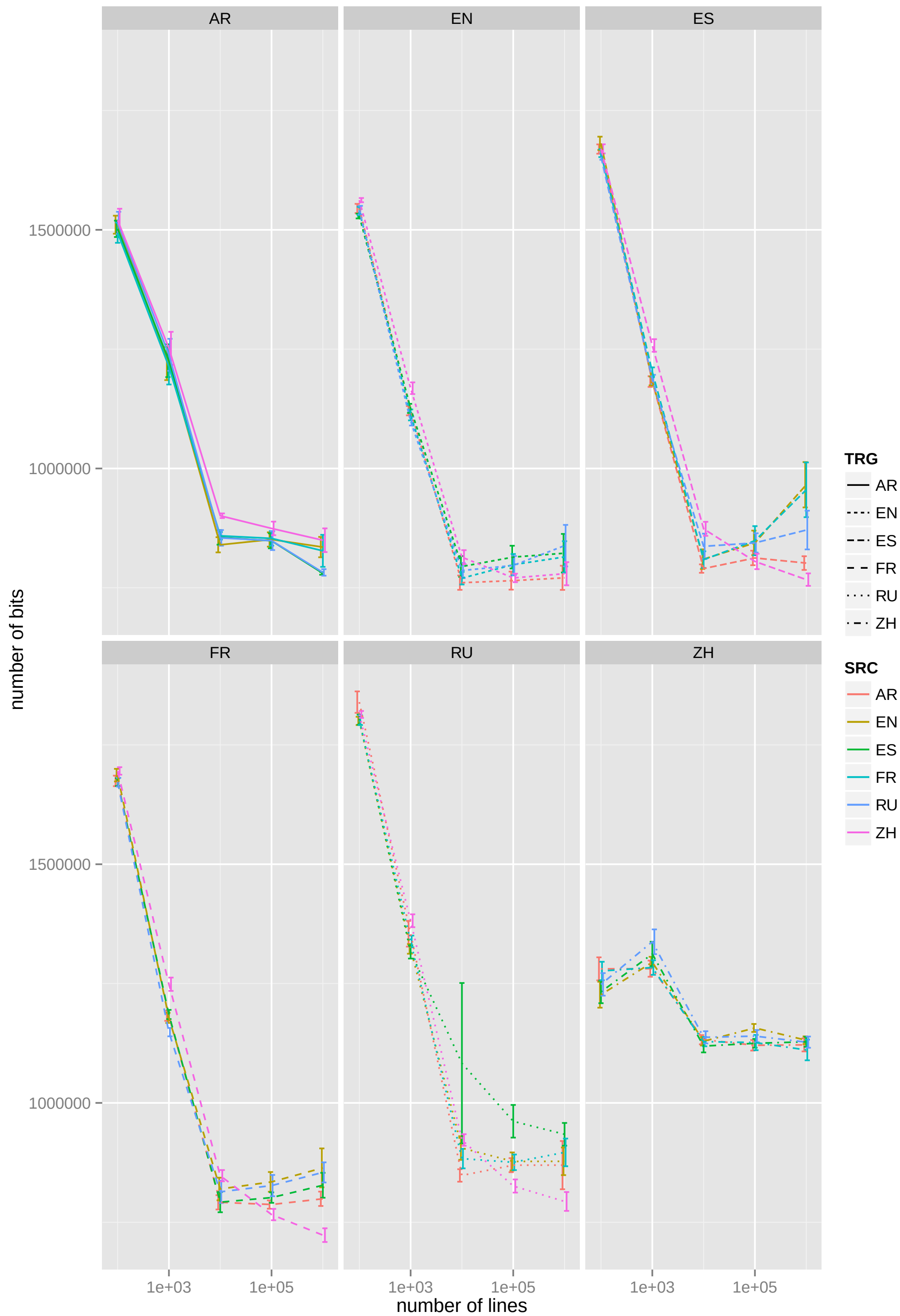


Figure 5: CHAR: character models (target language as facet)

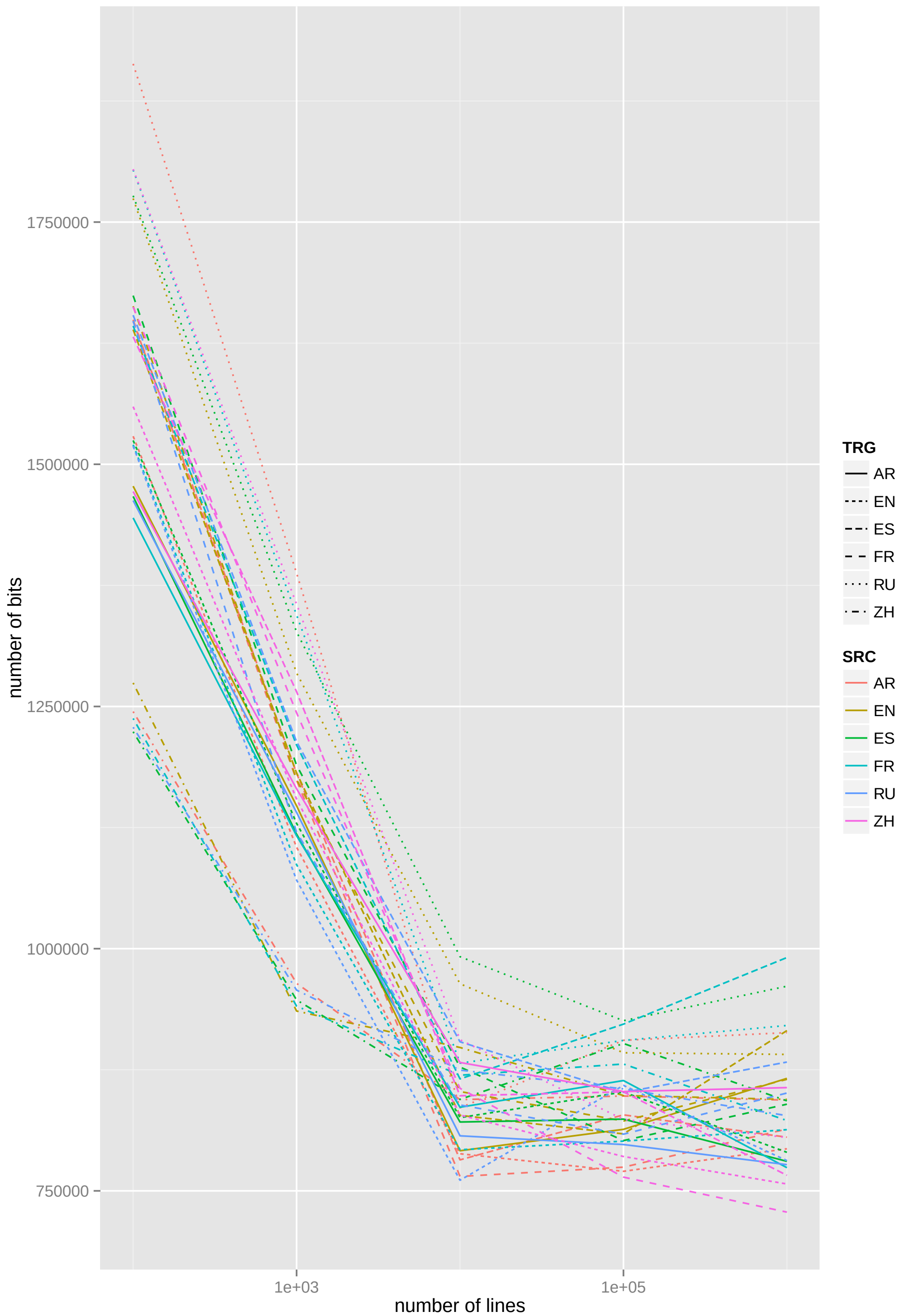


Figure 6: CHAR with Pinyin for ZH_{trg}

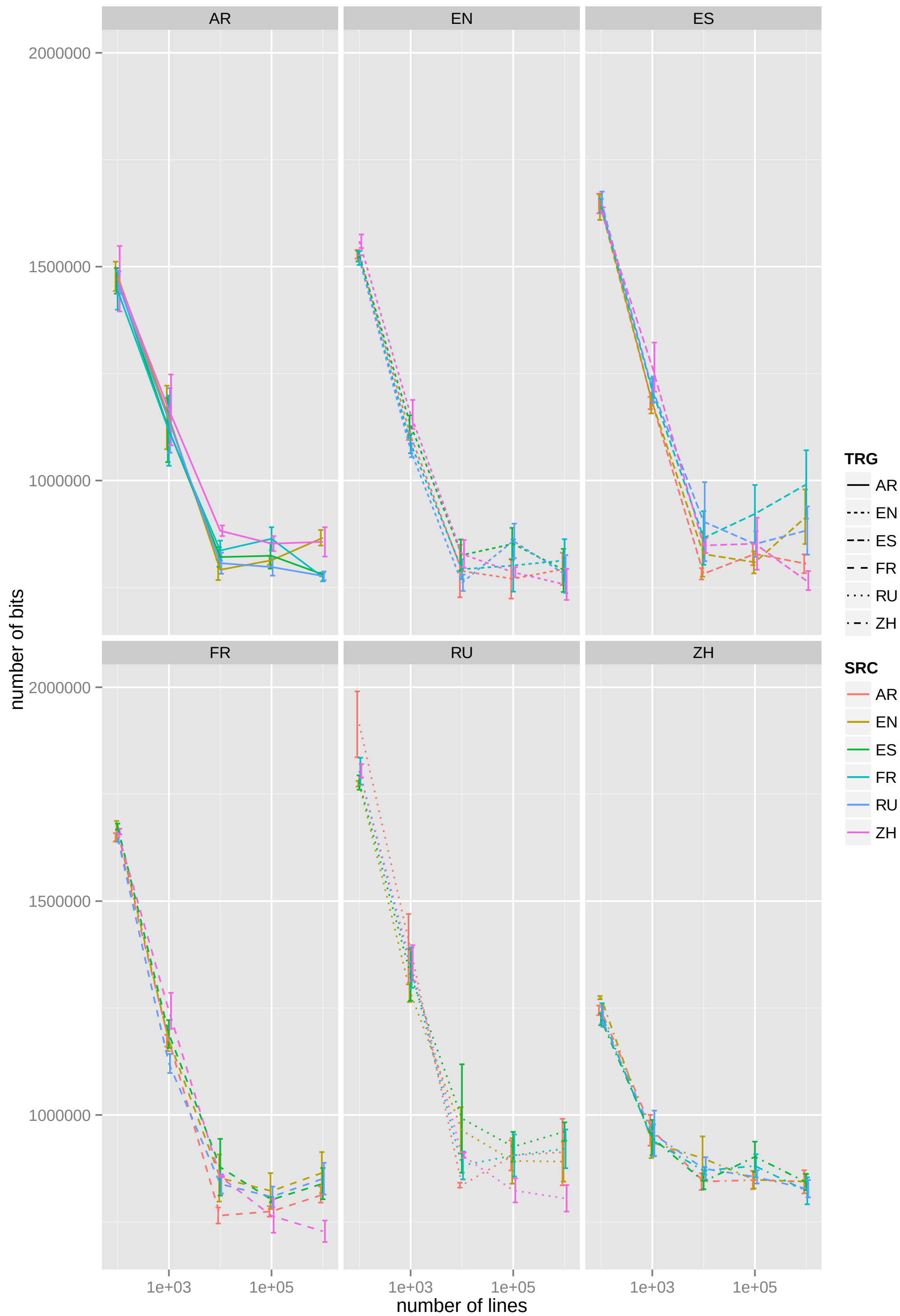


Figure 6: CHAR with Pinyin for ZH_{trg} (target language as facet)

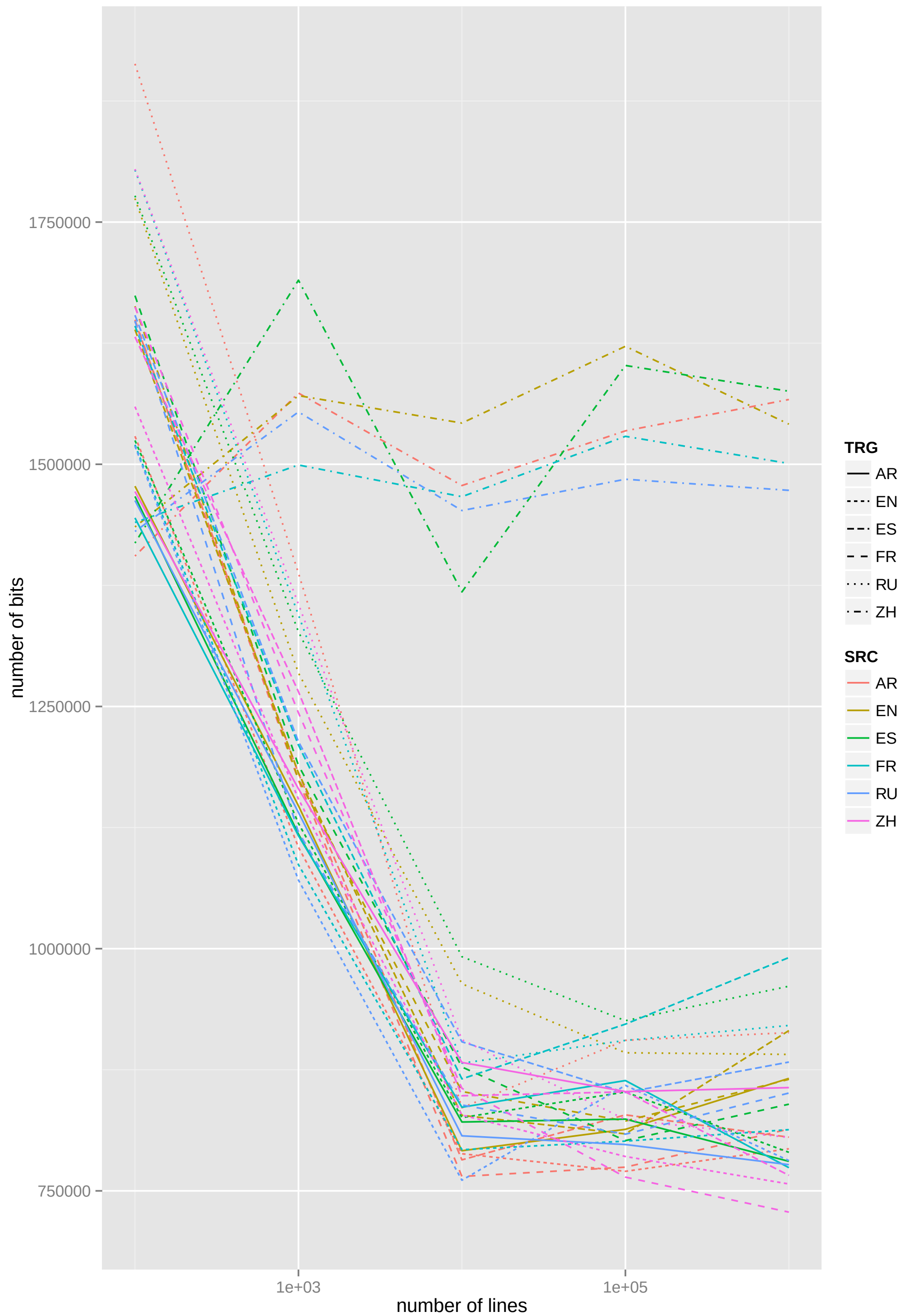


Figure 7: CHAR with Wubi for ZH_{trg}

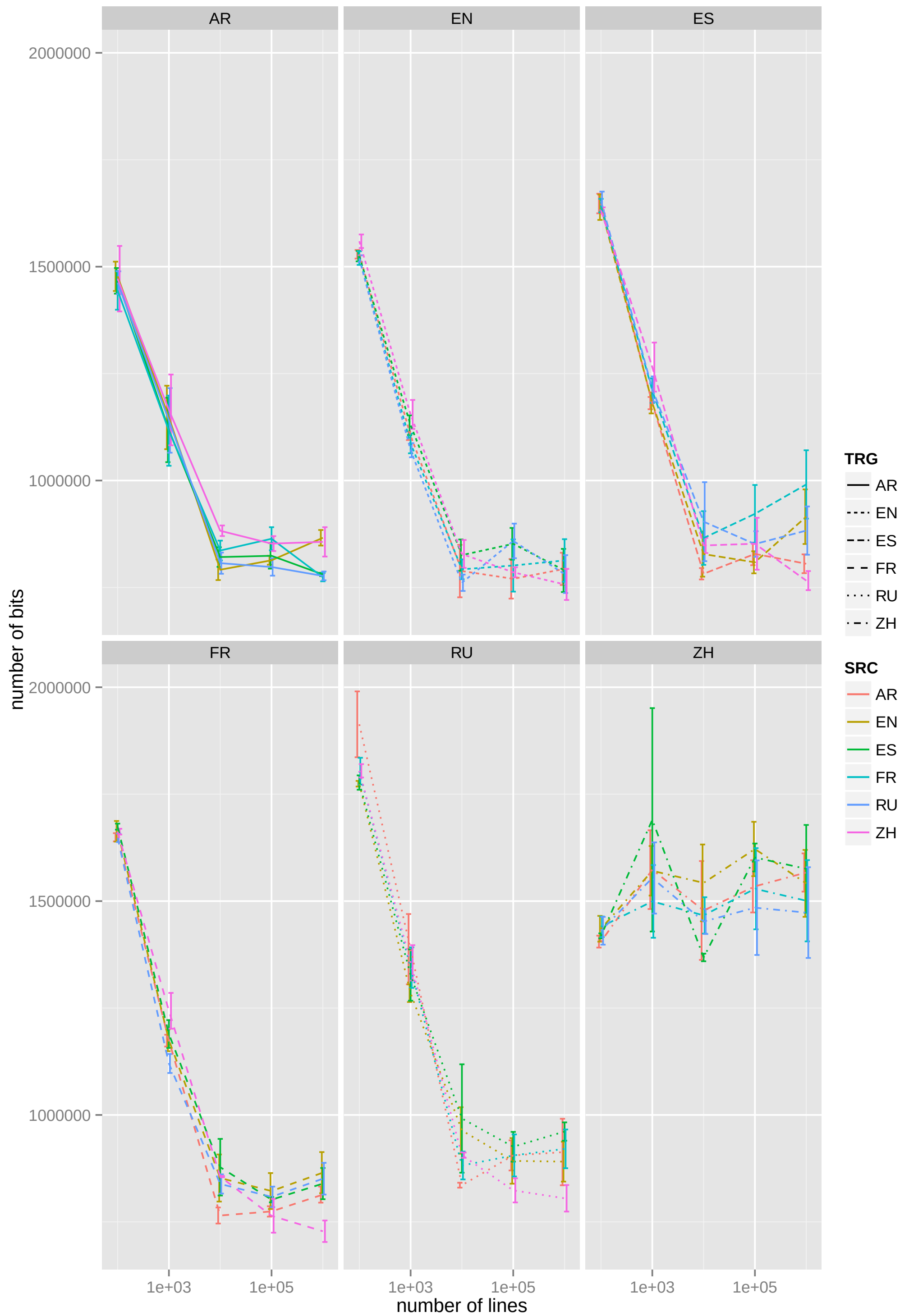


Figure 7: CHAR with Wubi for ZH_{trg} (target language as facet)

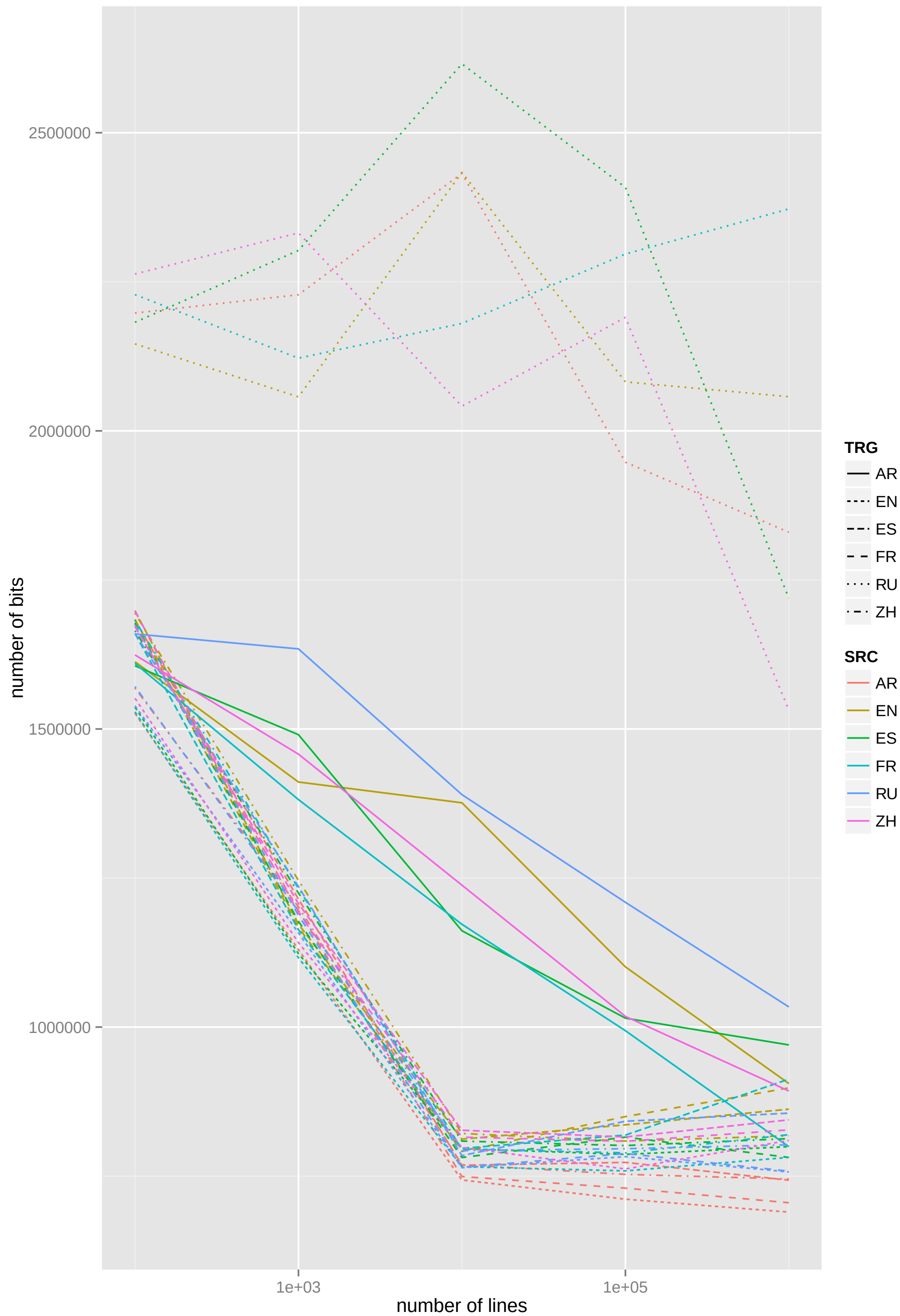


Figure 8: BYTE models with UTF-8 encoding

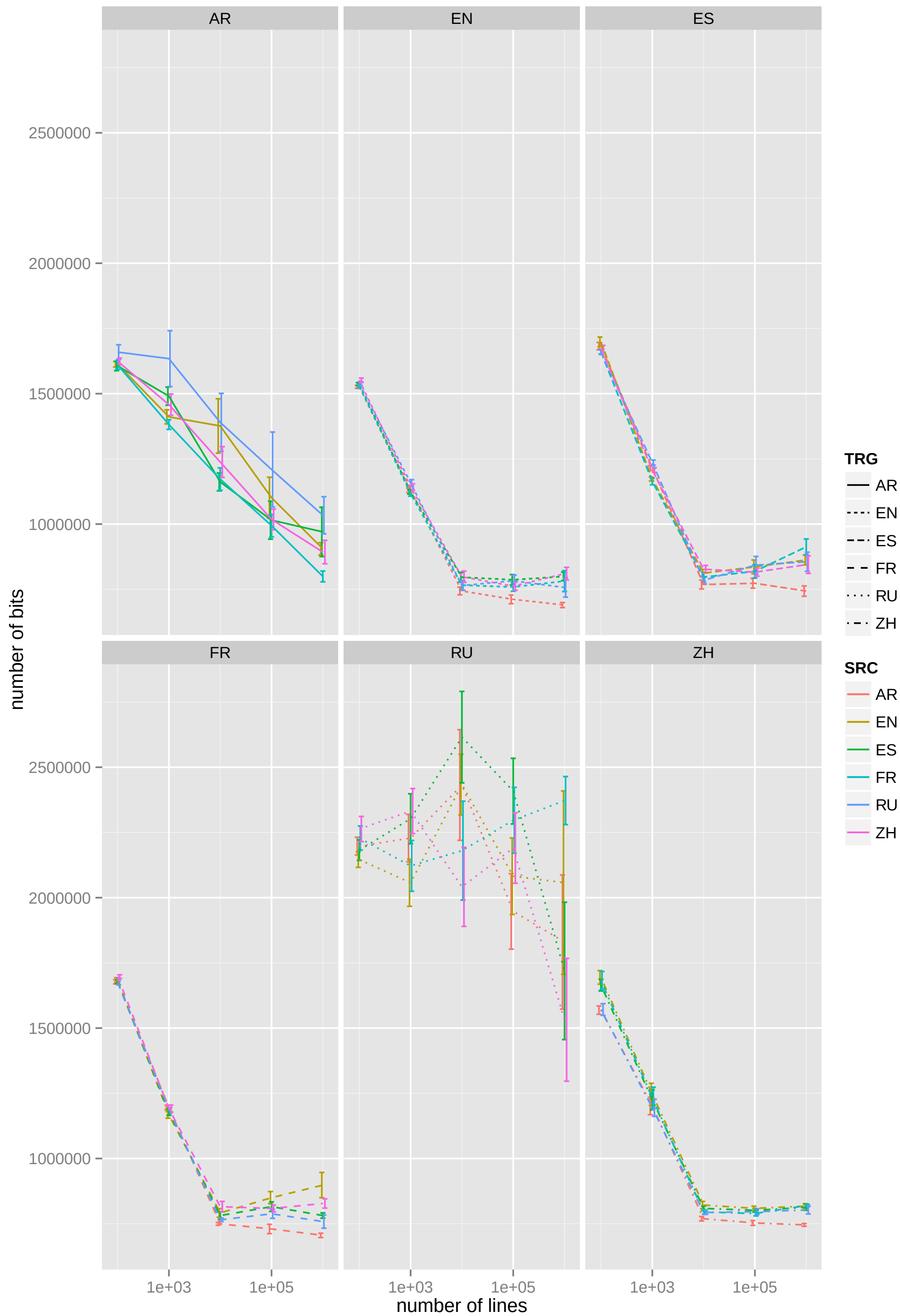


Figure 8: BYTE models with UTF-8 encoding (target language as facet)

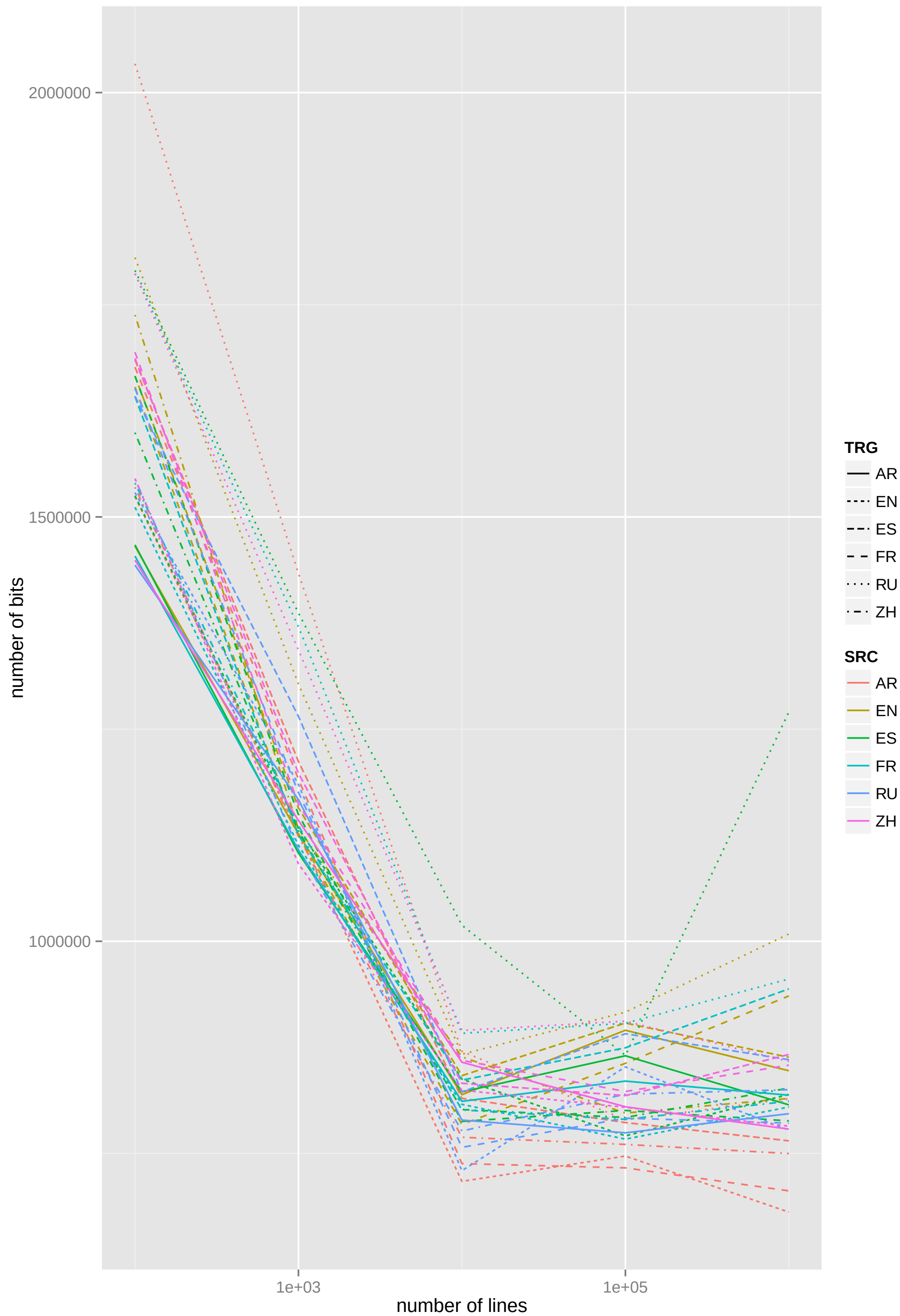


Figure 9: BYTE with AR_{trg} & RU_{trg} optimized with code pages 1256 & 1251 ($ARRU_{trg}$)

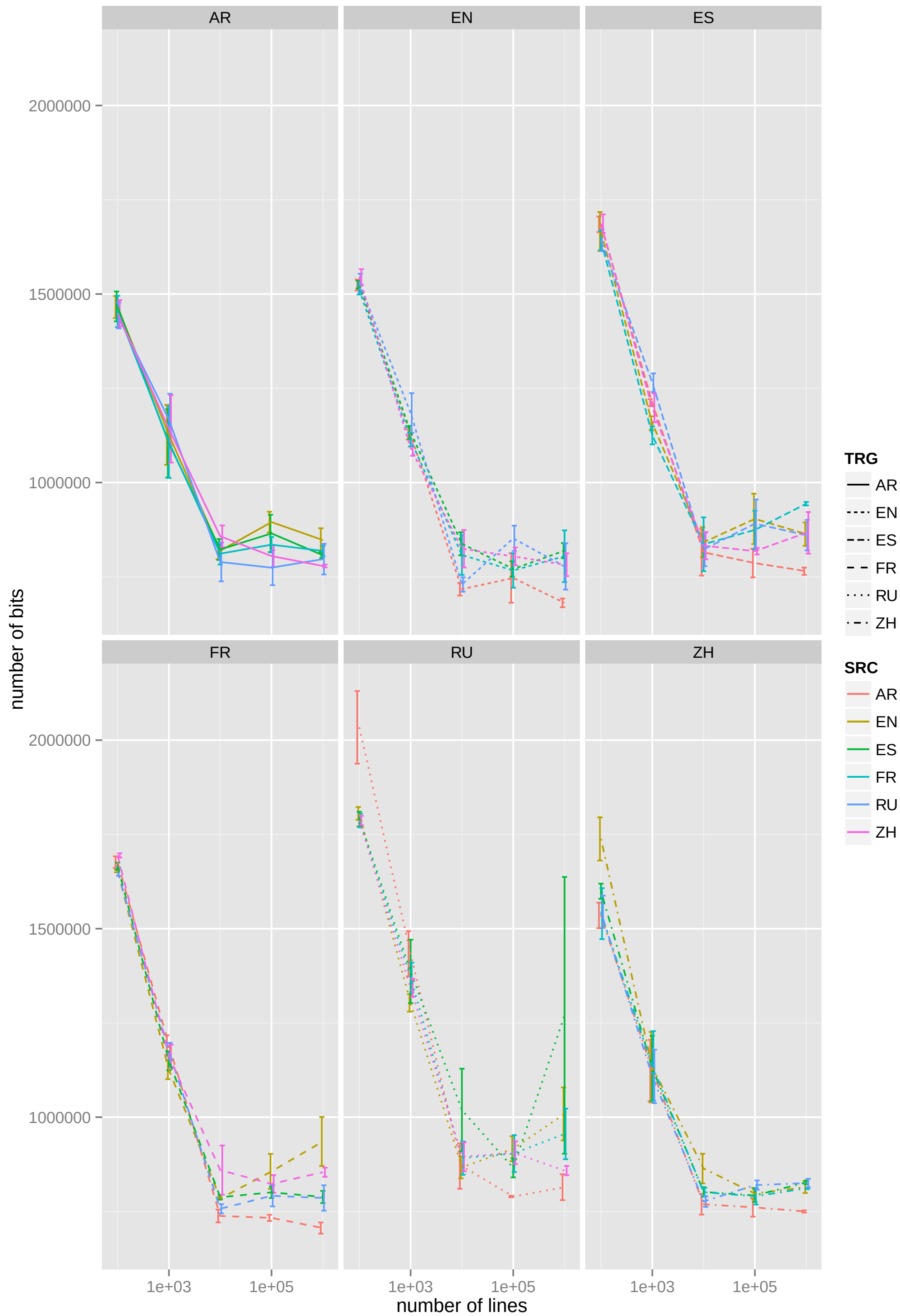


Figure 9: BYTE with AR_{trg} & RU_{trg} optimized with code pages 1256 & 1251 (target language as facet)

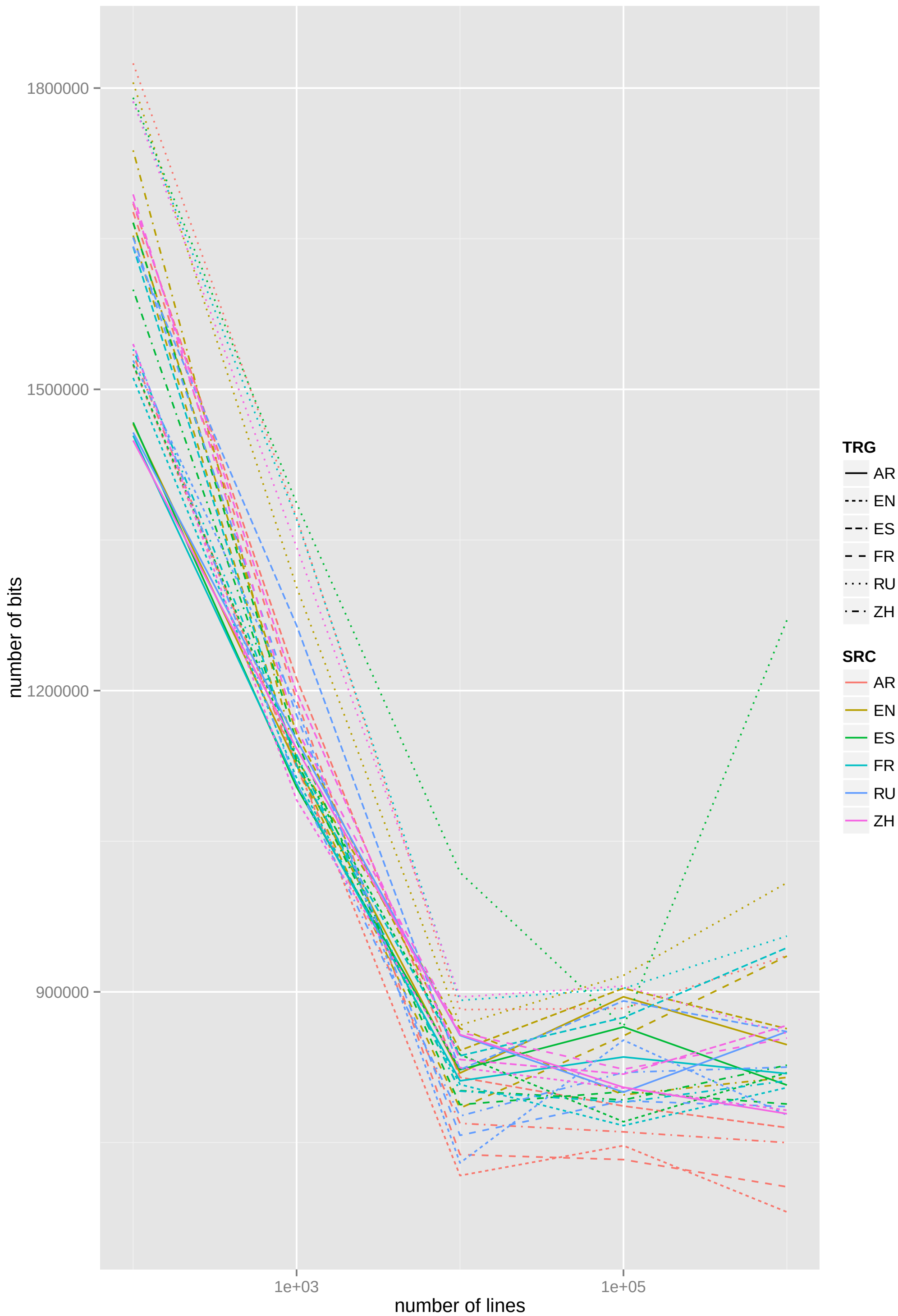


Figure 10: BYTE with directions AR-RU & RU-AR optimized on both source and target sides ($\text{ARRU}_{src,trg}$)

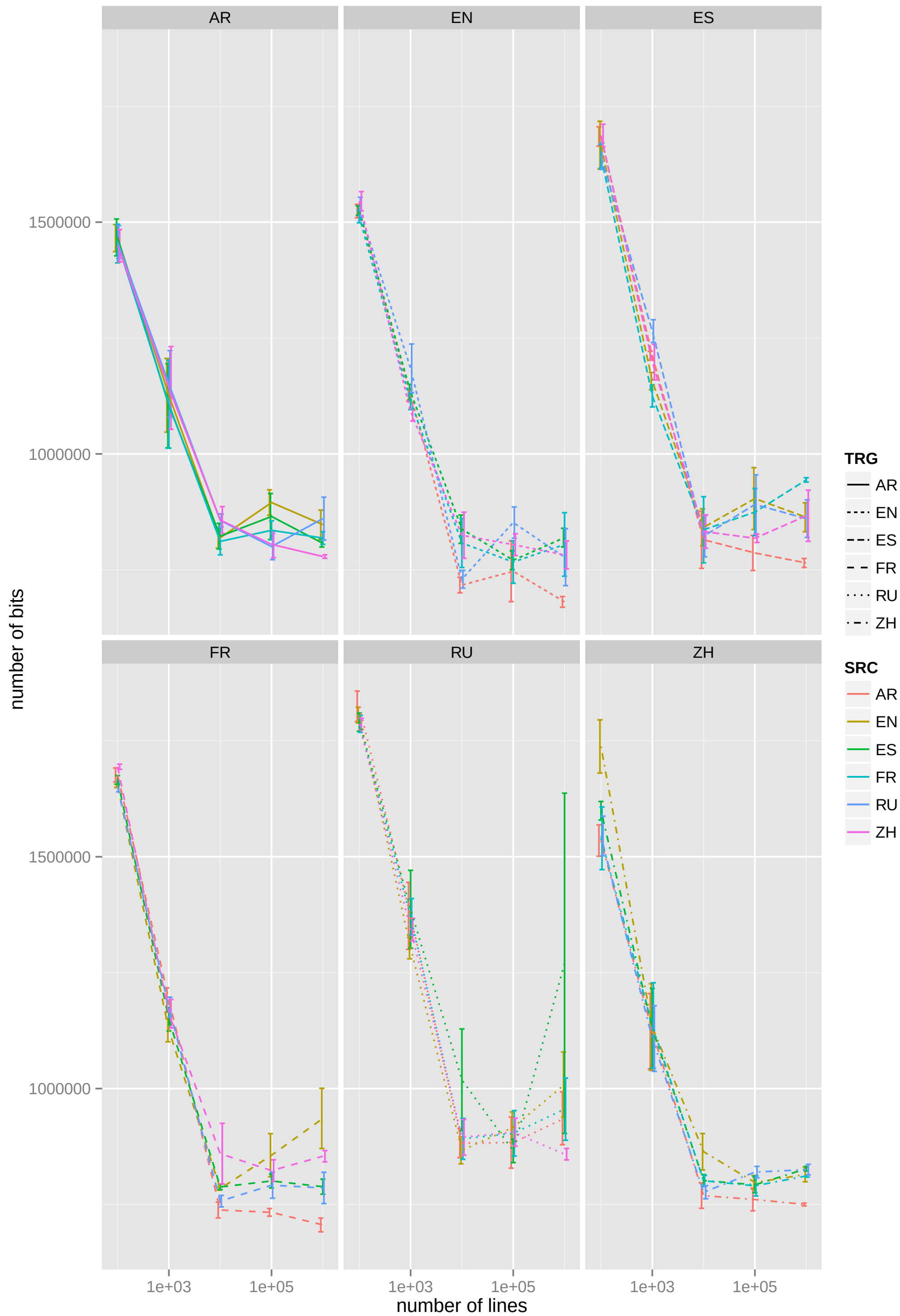


Figure 10: BYTE with directions AR-RU & RU-AR optimized on both source and target sides (target language as facet)

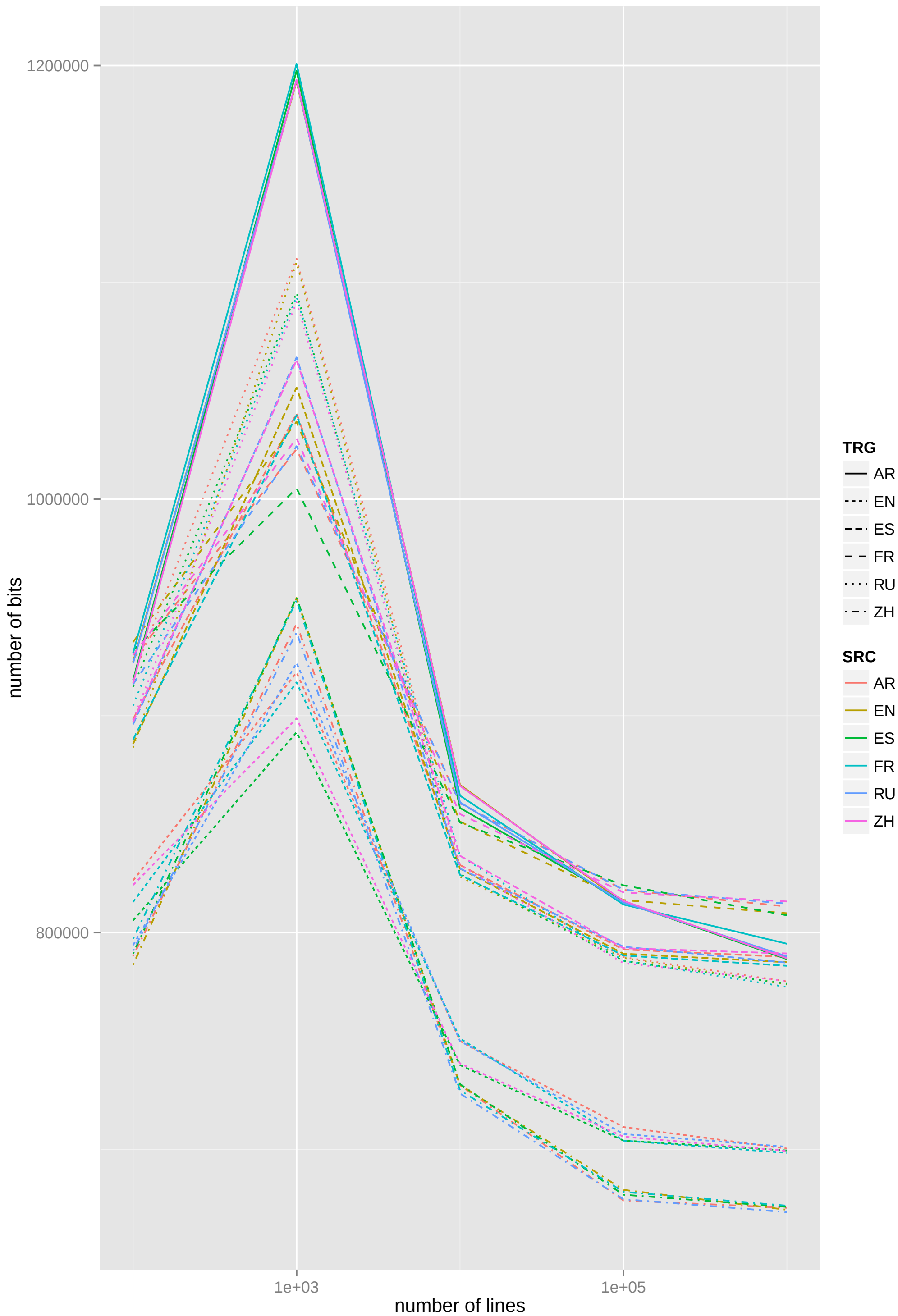


Figure 11: WORD models

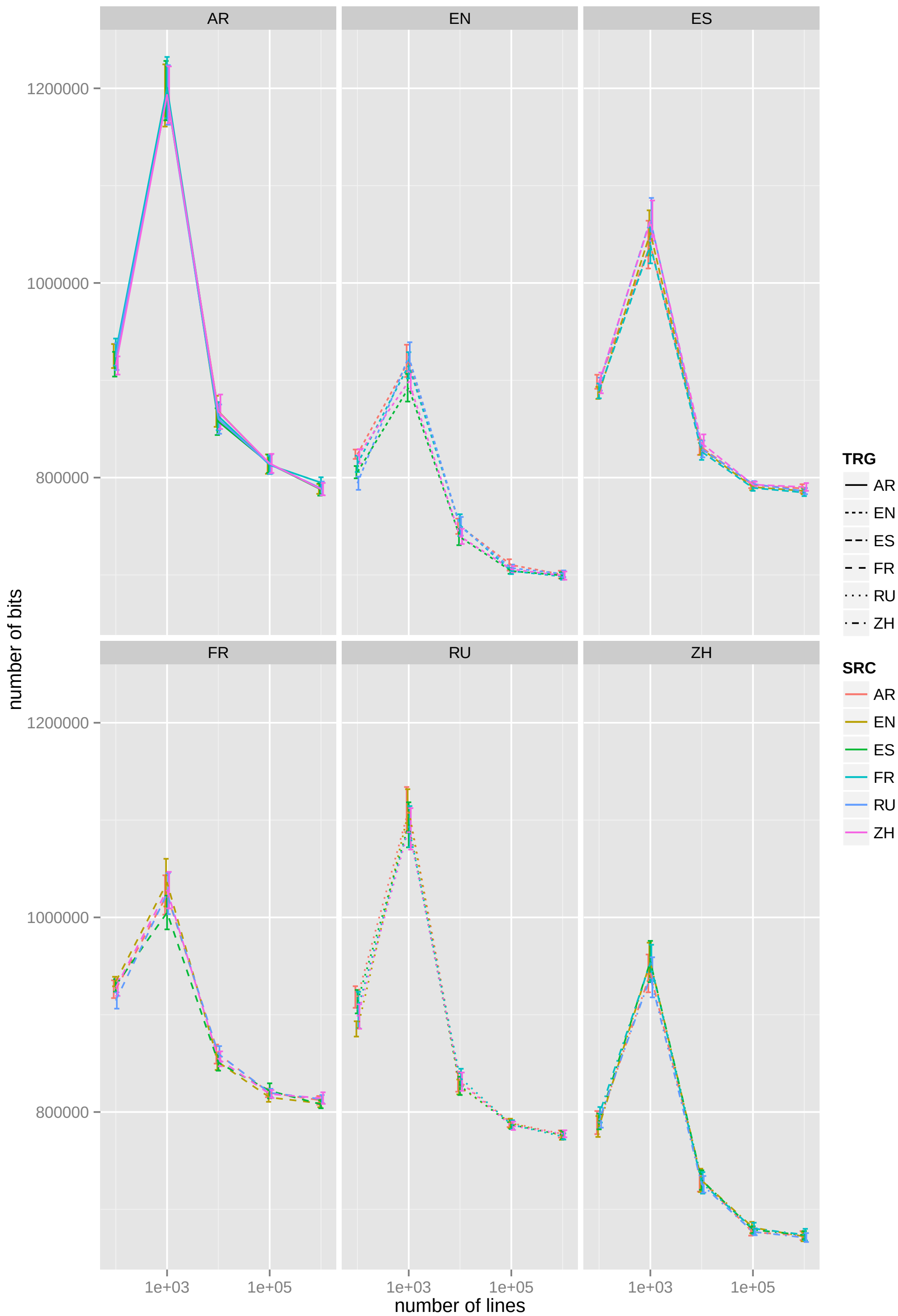


Figure 11: WORD models (target language as facet)

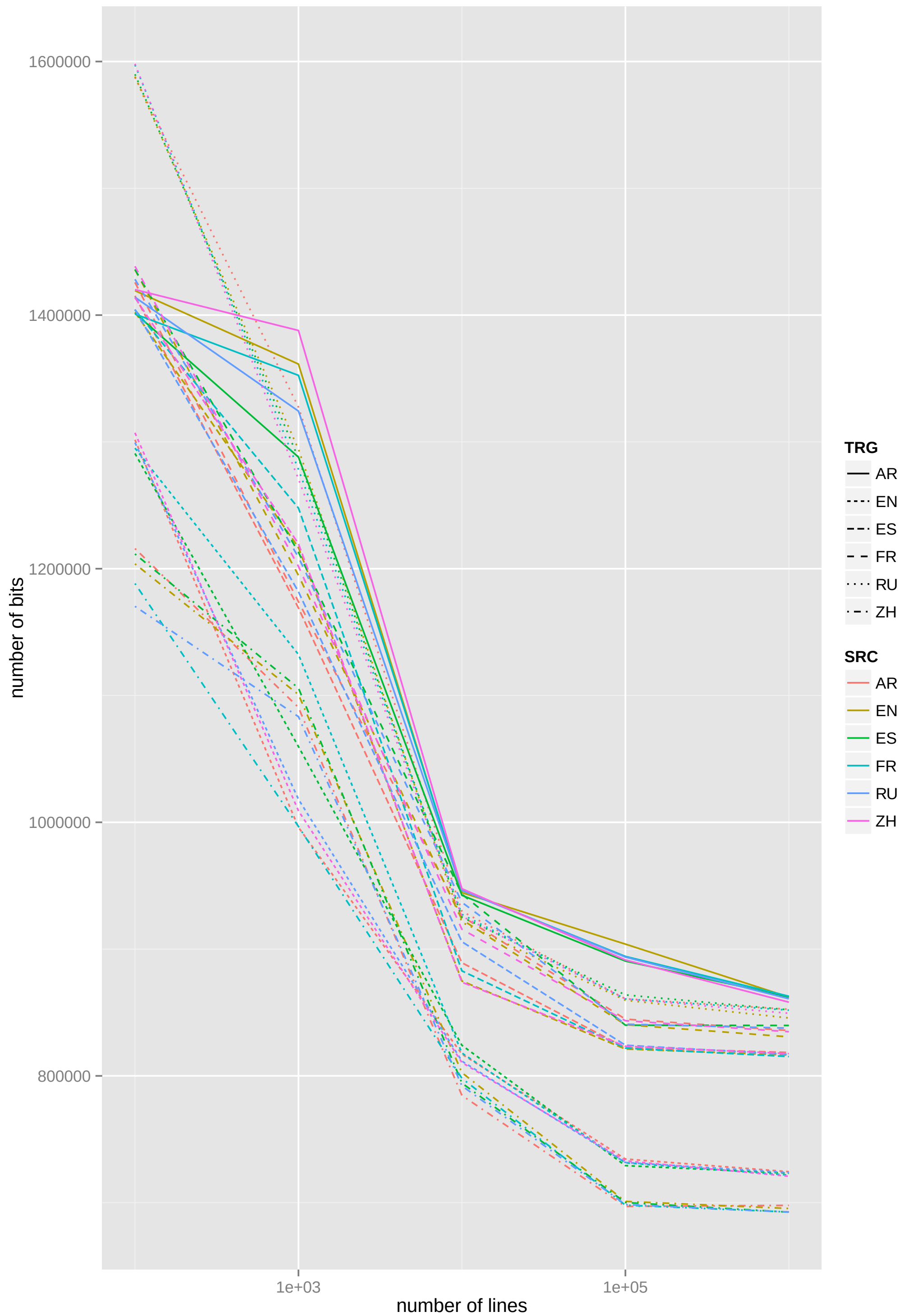


Figure 12: BPE models

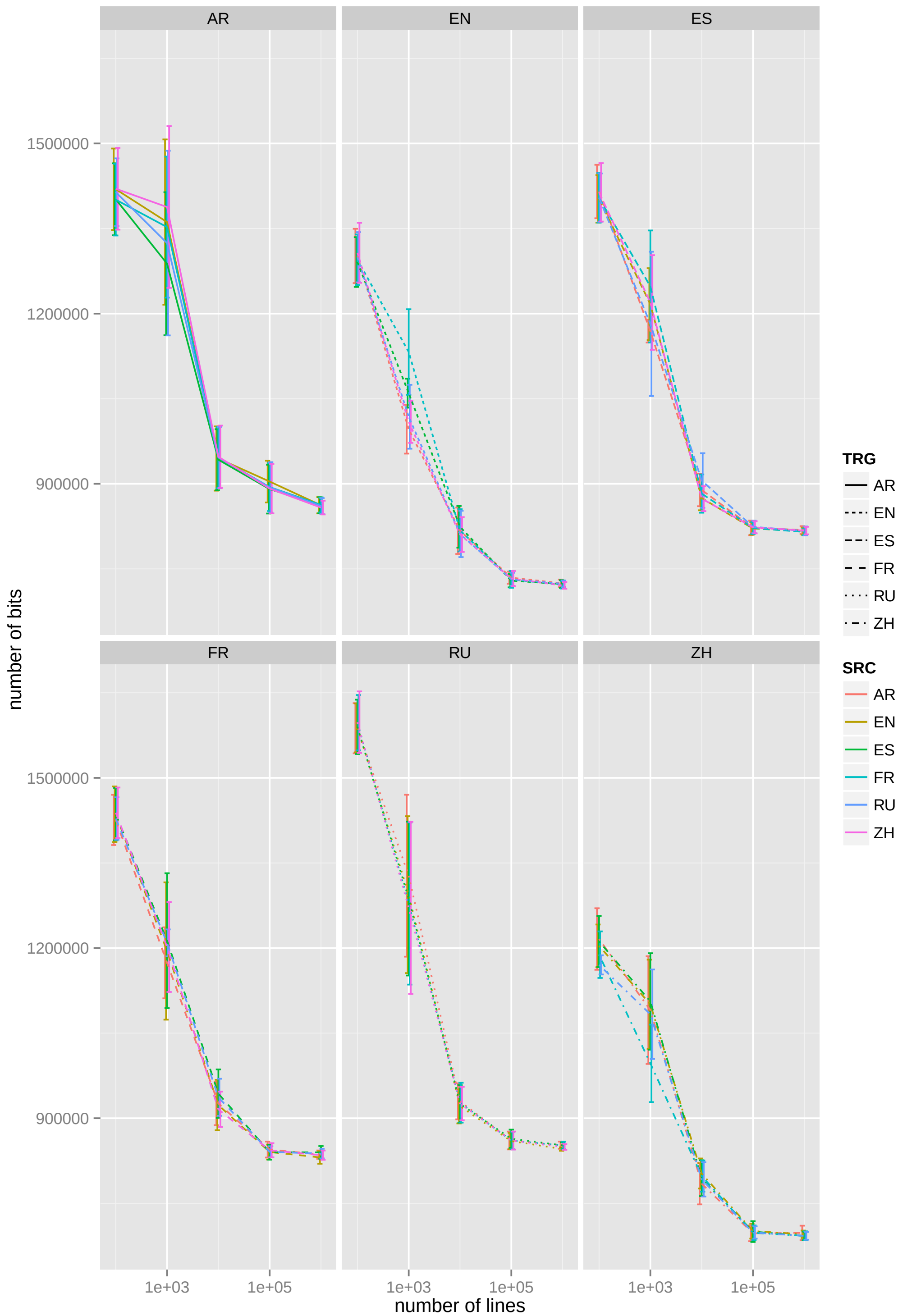
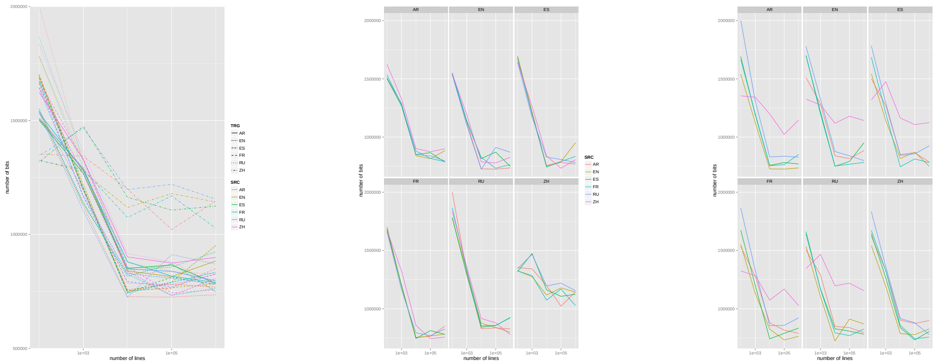


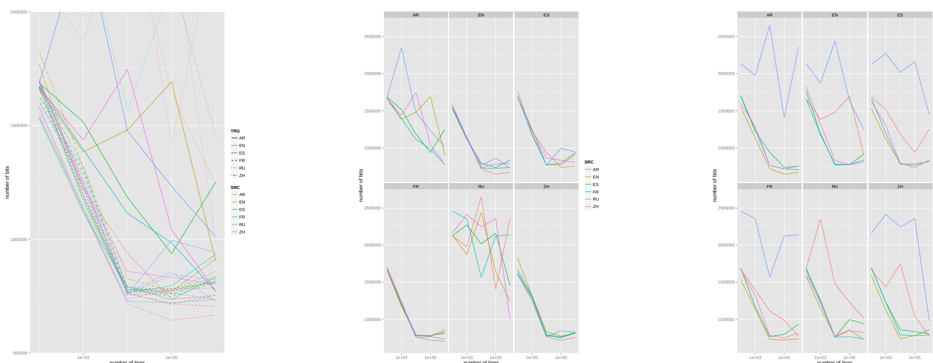
Figure 12: BPE models (target language as facet)

H SAMPLE FIGURES FROM RUN A0, ALSO SORTED BY SOURCE LANGUAGE FOR CONTRAST



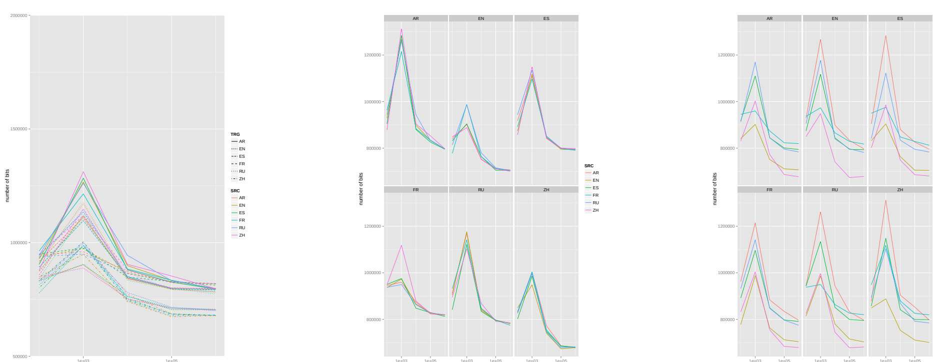
(a) CHAR (b) CHAR by target (c) CHAR by source

Figure 13: CHAR: character models from run A0



(a) BYTE (b) BYTE by target (c) BYTE by source

Figure 14: BYTE: byte models from run A0



(a) WORD (b) WORD by target (c) WORD by source

Figure 15: WORD: word models from run A0

I LANGUAGE PAIRS WITH SIGNIFICANT DIFFERENCES

15 (non-directional) language pairs total possible from 30 language directions, p=0.001.

LANG PAIR	CHAR	Pinyin	Wubi	BYTE	ARRU$_t$	ARRU$_{s,t}$	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

Language pairs with significant differences indicate that the 2 languages are *not* equally/similarly good or equally/similarly bad.

- Character models with ZH behave differently but the disparity can be eliminated with Pinyin.
- Byte models with AR and RU exhibit unstable performance due to length but this can be rectified with compression on the target side only (ARRU$_t$).
- Word-based models, including BPE, however, consistently favor EN and ZH (though it is more of a "mis-segmentation" for the latter, see § 3 and Appendix J) and disfavor AR and RU (as morphologically complex languages with higher OOV rates).

J Language complexity

In the words of Bentz et al. (2016):

> Languages are often compared with regard to their complexity from a computational, theoretical and learning perspective. In computational linguistics, it is generally known that methods mainly developed for the English language do not necessarily transfer well to other languages. The cross-linguistic variation in the amount of information encoded at the level of a word is, for instance, recognized as one of the main challenges for multilingual syntactic parsing (formulated as The Architectural Challenge (Tsarfaty et al., 2013)). Complexity of this kind is also found to influence machine translation: translating from morphologically rich languages into English is easier than the other way around (Koehn, 2005).

*****

Morphology is "the study of the formation and internal structure of words".
Morphemes are "the smallest meaningful units of language". (Bender, 2013)

*****

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), and ZH lacking morphological richness (Koehn, 2005). But this definition of morphology is predicated on the notion of word, defined primarily from an alphabetic perspective. As pointed out by Zhang & Komachi (2018), "the important differences between logographic and alphabetic writing systems have long been overlooked". In logographic languages (i.e. languages with logographic scripts), there can be units within a character that carry semantic and phonetic information that have never been accounted for in the traditional practice of morphology or in the computation of morphological complexity. For example, in the comparison of different morphological complexity measures by Bentz et al. (2016), all measures studied are defined with the notion of word.[8] Yet, there is **no universally valid definition of a "word"** — the form/idea (as in, the philosophical concept) of a "word" may be there for most languages/cultures (though that is certainly also debatable), but its instantiations are different in different languages/cultures, as well as in different genres/settings within one language. The variability in the definition of word is evident in the variation in language-specific word tokenization algorithms, along with the "indeterminacy of word segmentation" or a work-in-progress status for the definition of "word" advocated by Haspelmath (2011), as well as the contested nature of wordhood, esp. for logographic languages such as ZH (see Duanmu (2017) and Li et al. (2019b) for how some ZH speakers do indeed consider a ZH character to be a word or how "word", as conventionally used in NLP, is not a native term or does not correspond with speakers' judgement).

Our results with the Transformer indicate that a notion of morphological complexity can be modeled given our word tokenization scheme, confirming that morphological complexity is only predicated on the notion of word and bounded within the word level, and orthogonal to the performance of character or byte models. That is, unless word-based segmentation has been applied, there is no reason to attribute crosslinguistic performance disparity to differences in morphological complexity. In fact, on the character and byte level, we were able to achieve performance without disparity. Hence **disparity is not a necessary condition but an expectation that has been in mutual reinforcement with our practice of word segmentation, while the definitions of "morphological complexity" and "word" are in a circular dependency with each other.**

In this paper, we *re*solve language complexity, more specifically that of morphological complexity, in the context of computing through CLMing with the Transformer, in that we *explain away* the representation granularities and criteria relevant for such calculation.

TLDR: Up to the point of our taking up the subject of language complexity in this paper, there has been not a rigorous definition of "language complexity". Conventionally, "language complexity" is synonymous to "linguistic complexity" (with the tradition of "linguistics" being primarily word-

[8] An exception could be that of the type/token ratio (TTR). One could imagine applying TTR on the character level for ZH, and that would be indicative of its morphological richness on the character level. However, that has thus far never been practiced or recognized in NLP.

based), and people just assume linguistic complexity, e.g. morphological/syntactic complexity, to be intrinsic and necessary in languages (across representation levels). Our findings show that linguistic complexity is relative to the representation granularity, i.e. since morphology is based on words, it is bounded to the word level.

*****

An alternative perspective, with finer prints:

We have also developed a more rigorous interpretation. We take on the definition of "language complexity" as one that is related to the statistical attributes of languages. We assume and define *solving* as the elimination of statistically significant performance disparity.

In larger (6-layer) models, and according to the conventional definition of "language" — i.e. language as a whole, we solved language complexity with compression of AR and RU in byte representations. In smaller (1-layer) models, one can think of the situation as: i) no complexity has been modeled by the Transformer hence there is nothing to solve, or ii) there is no complexity between these languages to begin with, or iii) the Transformer solved the complexity.

With respect to each representation level/granularity in the larger models:

- BYTE: one can think of us as having solved complexity with byte representations or with 1-layer models — for these 6 languages empirically. Theoretically, there could be languages with longer sequence lengths than RU and AR, in those cases, we don't claim to have solved the matter empirically but only resolved it conceptually. But this is the most that anyone could do at the moment, as there is no relevant parallel data available.
- CHARACTER: one can think of us as having solved it via bytes or 1-layer models. Whether we can be considered to have solved it via Pinyin for ZH depends on whether the evaluator accepts decomposition into a *phonetic representation only* qualifies as a solution for the ZH language.
- WORD: one can think of us as having solved it via bytes or 1-layer models. It is not possible to solve it strictly within the word level without creating word segmentation criteria that would be unrelatable to native speakers. And since "word" is exclusively a human concept, we must either claim that a universal solution is undefined or undefinable for computing, or retreat to a unit that is the greatest common factor crosslinguistically. Since some ZH speakers consider ZH characters as words, we return to the character-level solution.

It is beyond the scope of our paper to solve the qualitative disparity on the word level. However, we do advocate a more inclusive evaluation and critical reflection on the possibility of discontinuing the usage of "word" as such a non-technical term biases against both "morphologically complex" and "morphologically simple" languages. The world of languages in written form can be divided into those with logographic scripts and those with (phonetic) alphabetic ones, with the unit of character being the greatest common factor of them all, from the human perspective. For technical processing, esp. for fair multilingual sequence-to-sequence modeling with the Transformer, we recommend measures that are more standardized, such as those based on bytes or characters. There is room for improvement in the design of character encoding that complements the statistical profiles, e.g. with relative rank in sequence length, of different languages. We believe there is crosslinguistic systematicity on the character level to be leveraged.

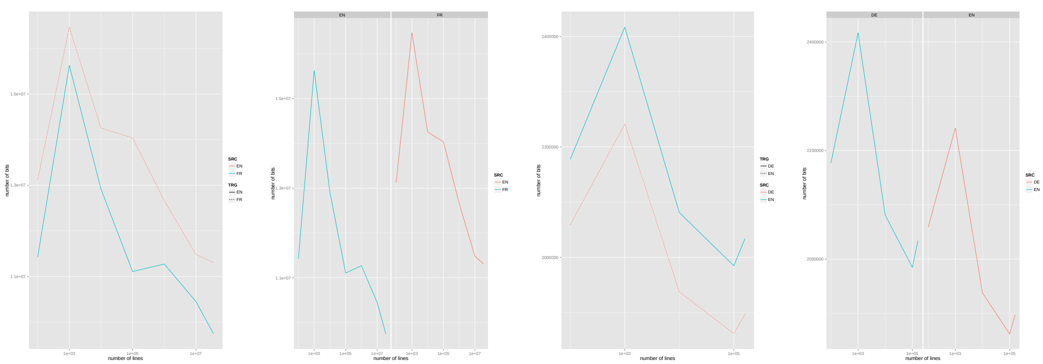
One's readiness to accept this as a solution to language complexity can be a subjective matter. One may insist that language complexity be solved exclusively with monolingual LMing (which lies outside the scope of the present work), instead of being confounded with the logic of one language being conditional on another. One may also object to the idea of (re-)solving morphological complexity being equivalent to or leading to solving language complexity as a whole, for there could also be e.g. syntactic complexity (although as substantial "information concerning syntactic units and relations is expressed at word level" in morphologically rich languages (Tsarfaty et al., 2010), the boundary between morphology and syntax is less distinct for some languages than others (Haspelmath, 2011)). If, however, our results could be extended, we wonder if syntactic complexity could be due to our sentence segmentation or a combination of word and sentence segmentation practice. That we leave for future work for those who are interested in the topic.

K Sample-wise Double Descent (DD)

K.1 Our experimental framework on DD datasets from (Nakkiran et al., 2020)

Text experiments from previous work reporting sample-wise DD involved words (Belkin et al., 2019) and BPEs (Nakkiran et al., 2020).

We applied our experimental framework — by testing data points with 10^n lines — on the datasets reported in (Nakkiran et al., 2020) to exhibit DD. WMT'14[9] EN-FR was reported to demonstrate model-wise DD and IWSLT'14 (Cettolo et al., 2012) DE-EN model-wise and sample-wise DD. We downloaded and prepared the data with scripts[10] from the FAIRSEQ Toolkit (Ott et al., 2019). The WMT data was preprocessed with 40,000 BPE operations and IWSLT 10,000. Our focus is on sample-wise DD and hence our goal was to see if the spike at 10^3 we observed with the UN data would apply also to these datasets. We used the same training regime[11] with the Transformer and Adam on SOCKEYE as before and tested both language directions on the entirety of both datasets, with no subsampling. For the IWSLT dataset, we tested data sizes with $10^2 - 10^5$ lines, then at $160,239$ as that is the total number of lines available. For the WMT dataset, we tested from 10^2 to 10^7, then at $35,762,532$.



(a) WMT'14 (b) WMT'14 by target (c) IWSLT'14 (d) IWSLT'14 by target

Figure 16: WMT'14 EN-FR and FR-EN and IWSLT'14 DE-EN and EN-DE: sample-wise DD shown at 10^3

Table 2: Target-Train-Token-to-Parameter ratio (TTT2P ratio) for WMT'14 EN-FR and FR-EN

	Number of lines						
	100	1,000	10,000	100,000	1,000,000	10,000,000	35,762,532
EN: num train tokens	3,248	33,768	313,154	3,123,129	30,852,455	308,640,462	1,174,344,513
FR: num train tokens	3,548	36,507	339,803	3,414,959	33,865,679	343,344,536	1,327,817,765
EN-FR num params	45,609,474	51,039,363	62,871,584	75,630,304	85,210,037	108,226,335	111,417,633
TTT2P ratio	0.000078	0.000715	0.005405	0.045153	0.397438	3.172468	11.917483
FR-EN num params	45,540,219	50,692,575	61,916,891	74,547,874	83,936,258	107,378,859	111,399,165
TTT2P ratio	0.000071	0.000666	0.005058	0.041894	0.367570	2.874313	10.541771

This shows that the effect we reported in § 5 also holds on these datasets: "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".

[9] `http://www.statmt.org/wmt14/translation-task.html`

[10] `https://github.com/pytorch/fairseq/blob/master/examples/translation/prepare-wmt14en2fr.sh` and `https://github.com/pytorch/fairseq/blob/master/examples/translation/prepare-iwslt14.sh`

[11] max-seq-len 300; checkpoint-frequency 4000 except for cases where 50 epochs would be reached before the first checkpoint: 400 for 10^2 lines and 3450 for 10^3 lines.

Table 3: Target-Train-Token-to-Parameter ratio (TTT2P ratio) for IWSLT'14 DE-EN and EN-DE

	Number of lines				
	100	1,000	10,000	100,000	160,239
DE: num train tokens	2,874	27,675	253,757	2,519,534	4,035,591
EN: num train tokens	2,739	26,416	245,659	2,461,879	3,949,114
DE-EN num params	45,297,348	49,410,683	53,639,825	55,189,376	55,428,584
TTT2P ratio	0.000060	0.000535	0.004580	0.044608	0.071247
EN-DE num params	45,405,078	49,809,797	54,300,056	56,245,643	56,564,366
TTT2P ratio	0.000063	0.000556	0.004673	0.044795	0.071345

K.2 Token-to-parameter ratio for non-neural monolingual LMs

We experimented also on KenLM (Heafield, 2011; Heafield et al., 2013), a non-neural LM with modified Kneser-Ney smoothing (Kneser & Ney, 1995; Chen & Goodman, 1999), on our dataset A and found that on the word level, such a spike (or a hump) is common across all languages, see Figure 17. The target-token-to-parameter ratio is under 1 for most of these smaller data sizes. This seems related to the analytical findings in Opper et al. (1990) where the pseudo-inverse solution to a simple learning problem was shown to exhibit non-monotonicity, with the peak exactly as the ratio of data to parameters (α) approaches 1.

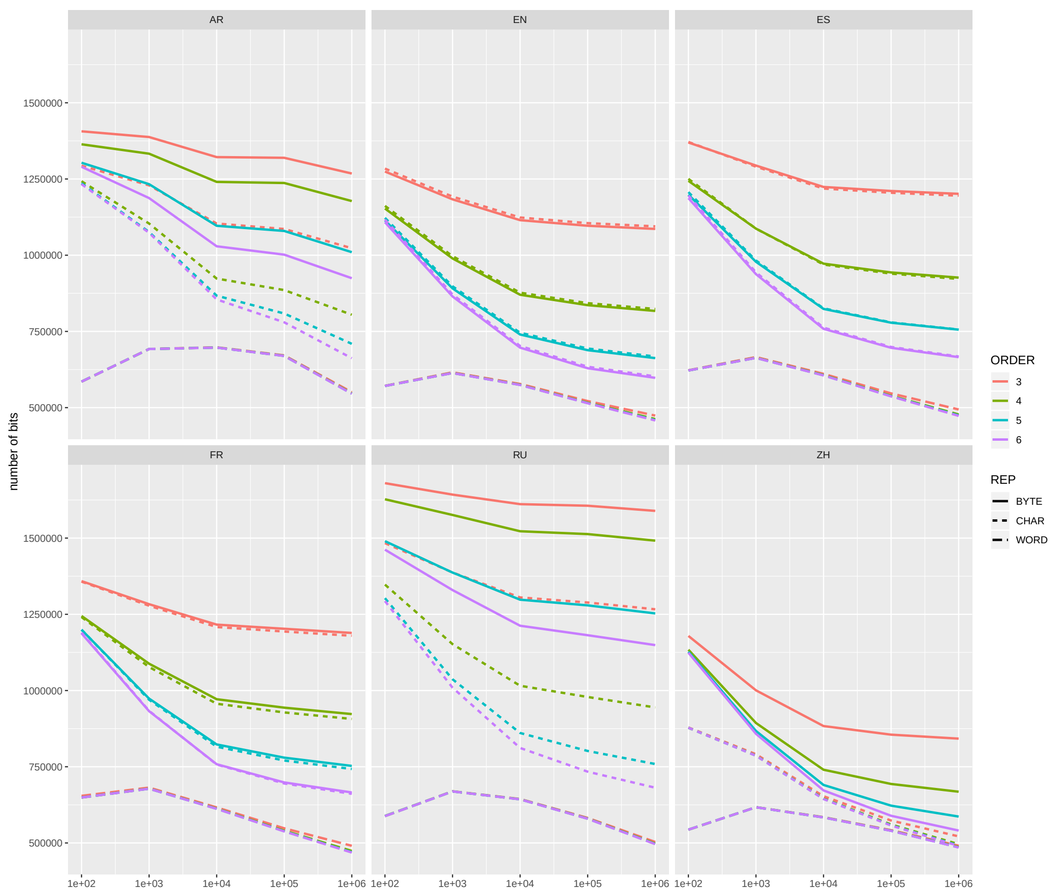


Figure 17: Kneser-Ney (monolingual) n-gram LMs on the same data (A) used for our neural CLMs

The number of parameters of a k-gram model is the number of unique n-grams, $1 \leq n \leq k$. Table 4 shows the ratios for our trigram model (all n-gram models of higher order exhibit the same effect).

On word level, where the function of number of bits to data size is not always monotonic, we observe less of a monotonic development whenever the token-to-parameter ratio is smaller than 1. This is more notably shown in the first 4 sizes in AR with a hump-like curve before the performance improves at 10^6. This is different from the sharper descent for ES and FR, where only the first two data sizes have a non-monotonic relationship and a token-to-parameter ratio less than 1. Taking the token-to-parameter ratio as a rough proxy for over- (< 1) and under-parameterization (> 1), this can be seen as an instance of non-monotonicity with respect to data size in the "critical regime", i.e. when the model transitions from being (heavily) over- to under-parameterized (Belkin et al., 2019; Nakkiran, 2019).

A remark on modeling with finer granularity Our KenLM results show the performance of bytes and characters is not on par with that of words with non-neural algorithms. NNs/DL has enabled much progress in this regard.

Table 4: Token-to-parameter ratios on non-neural monolingual trigram LMs

	lang_numlines	num_tokens	| unigrams |	| bigrams |	| trigrams |	num_params	tokens/params
CHAR	AR_100	9079	85	925	2894	3904	2.325563525
	AR_1000	123832	110	1577	8592	10279	12.04708629
	AR_10000	1083517	152	3216	21479	24847	43.60755826
	AR_100000	10625047	179	5114	44251	49544	214.4567859
	AR_1000000	102064230	242	8517	90353	99112	1029.786807
	EN_100	11730	78	806	2532	3416	3.433840749
	EN_1000	159444	84	1215	5808	7107	22.43478261
	EN_10000	1344001	125	2532	17181	19838	67.7488154
	EN_100000	13132862	170	4231	36104	40505	324.2281694
	EN_1000000	123491871	247	7126	70406	77779	1587.727677
	ES_100	12374	87	781	2398	3266	3.788732394
	ES_1000	171104	93	1210	5045	6348	26.95400126
	ES_10000	1484804	117	2534	15462	18113	81.97449346
	ES_100000	14549703	176	4261	33554	37991	382.9776263
	ES_1000000	138596036	257	7217	67280	74754	1854.02836
	FR_100	12456	89	836	2610	3535	3.523620934
	FR_1000	179048	97	1259	5711	7067	25.33578605
	FR_10000	1490983	133	2607	16282	19022	78.38203133
	FR_100000	14528593	178	4390	35051	39619	366.707716
	FR_1000000	138049189	259	7353	69522	77134	1789.732012
	RU_100	11980	98	952	3051	4101	2.921238722
	RU_1000	168156	111	1415	7106	8632	19.48053753
	RU_10000	1436078	163	3506	20478	24147	59.4723154
	RU_100000	14151728	190	5737	44071	49998	283.0458818
	RU_1000000	134706120	263	10186	94975	105424	1277.755729
	ZH_100	3318	605	2036	2634	5275	0.6290047393
	ZH_1000	42572	1239	13266	24811	39316	1.082816156
	ZH_10000	372003	2270	68178	175730	246178	1.511113909
	ZH_100000	3659617	3403	241215	968852	1213470	3.015828162
	ZH_1000000	34672612	4888	611213	3977112	4593213	7.548661906
BYTE	AR_100	16655	76	320	1163	1559	10.68313021
	AR_1000	227163	98	539	2070	2707	83.91688216
	AR_10000	1985014	133	1616	5974	7723	257.0262851
	AR_100000	19487689	148	2844	14274	17266	1128.674215
	AR_1000000	186171180	165	5219	40507	45891	4056.812447
	EN_100	11731	79	807	2533	3419	3.431120211
	EN_1000	159449	85	1219	5812	7116	22.40711074
	EN_10000	1345771	130	2527	17139	19796	67.98196605
	EN_100000	13158948	154	3971	34985	39110	336.4599335
	EN_1000000	123705128	169	6422	66606	73197	1690.030029
	ES_100	12629	88	766	2414	3268	3.864443084
	ES_1000	175286	94	1146	4901	6141	28.54355968
	ES_10000	1513782	121	2409	14894	17424	86.87913223
	ES_100000	14821495	154	3925	31905	35984	411.8912572
	ES_1000000	141276766	169	6338	62199	68706	2056.250779
	FR_100	12875	90	830	2560	3480	3.699712644
	FR_1000	185227	99	1227	5497	6823	27.14744247
	FR_10000	1542105	133	2492	15615	18240	84.54523026
	FR_100000	15055657	156	4014	33105	37275	403.9076325
	FR_1000000	143495667	175	6423	64044	70642	2031.308103
	RU_100	21751	100	475	1365	1940	11.21185567
	RU_1000	309279	113	694	2732	3539	87.39163606
	RU_10000	2636591	151	1898	8430	10479	251.607119
	RU_100000	25990263	160	3364	18321	21845	1189.757977
	RU_1000000	247098758	169	6224	45935	52328	4722.113553
	ZH_100	8559	140	1524	3532	5196	1.647228637
	ZH_1000	116667	146	2706	12857	15709	7.426761729
	ZH_10000	1019969	156	5596	36176	41928	24.32667907
	ZH_100000	9990046	167	9228	81997	91392	109.3098521
	ZH_1000000	94268840	196	13407	160359	173962	541.893287
WORD	AR_100	1776	869	1534	1669	4072	0.4361493124
	AR_1000	23460	5868	16064	20063	41995	0.5586379331
	AR_10000	206549	26108	116814	164062	306984	0.6728331118
	AR_100000	2035190	97997	776730	1383009	2257736	0.9014295737
	AR_1000000	19410502	304978	4297319	10005650	14607947	1.328763173
	EN_100	2071	682	1567	1869	4118	0.5029140359
	EN_1000	27398	3292	13148	19834	36274	0.7553068313
	EN_10000	236569	12014	83397	155493	250904	0.9428665944
	EN_100000	2339109	37264	428249	1117802	1583315	1.477349106
	EN_1000000	21943139	122457	1818166	6505850	8446473	2.59790554
	ES_100	2232	710	1605	1974	4289	0.5204010259
	ES_1000	29461	3839	13199	20634	37672	0.7820397112
	ES_10000	263024	15116	83900	160078	259094	1.01516824
	ES_100000	2588791	49499	439584	1116177	1605260	1.612692648
	ES_1000000	24654449	142809	1840029	6268684	8251522	2.987866844
	FR_100	2298	745	1737	2072	4554	0.5046113307
	FR_1000	32011	3881	14535	22608	41024	0.780299337
	FR_10000	273195	13998	86815	170729	271542	1.006087456
	FR_100000	2684982	42870	428339	1150965	1622174	1.655175092
	FR_1000000	25595487	118204	1703399	6171437	7993040	3.202221808
	RU_100	1854	886	1589	1734	4209	0.4404846757
	RU_1000	24746	5433	15511	20035	40979	0.603870275
	RU_10000	216638	23403	108516	162401	294320	0.7360627888
	RU_100000	2150746	81342	670857	1306351	2058550	1.044786865
	RU_1000000	20421965	236088	3295028	8617195	12148311	1.681053852
	ZH_100	1751	630	1434	1614	3678	0.4760739532
	ZH_1000	23568	3181	13998	19341	36520	0.6453450164
	ZH_10000	207714	13137	96829	160642	270608	0.7675826287
	ZH_100000	2038639	46941	554739	1278188	1879868	1.08445859
	ZH_1000000	19361101	134492	2527710	8401311	11063513	1.749995774

L Number of model parameters

Number of model parameters for dataset A

Representation / Number of lines	CHAR 100	CHAR 1,000	CHAR 10,000	CHAR 100,000	CHAR 1,000,000
Number of PARAMS					
AR-EN	44,226,639	44,245,589	44,309,118	44,369,067	44,480,248
AR-ES	44,235,864	44,254,814	44,300,918	44,375,217	44,490,498
AR-FR	44,237,914	44,258,914	44,317,318	44,377,267	44,492,548
AR-RU	44,247,139	44,273,264	44,348,068	44,389,567	44,496,648
AR-ZH	44,766,814	45,429,464	46,507,743	47,682,892	49,237,273
EN-AR	44,230,230	44,258,927	44,322,969	44,373,684	44,477,683
EN-ES	44,232,280	44,241,502	44,287,094	44,370,609	44,493,058
EN-FR	44,234,330	44,245,602	44,303,494	44,372,659	44,495,108
EN-RU	44,243,555	44,259,952	44,334,244	44,384,959	44,499,208
EN-ZH	44,763,230	45,416,152	46,493,919	47,678,284	49,239,833
ES-AR	44,234,838	44,263,535	44,318,873	44,376,756	44,482,803
ES-EN	44,227,663	44,236,885	44,291,198	44,367,531	44,487,928
ES-FR	44,238,938	44,250,210	44,299,398	44,375,731	44,500,228
ES-RU	44,248,163	44,264,560	44,330,148	44,388,031	44,504,328
ES-ZH	44,767,838	45,420,760	46,489,823	47,681,356	49,244,953
FR-AR	44,235,862	44,265,583	44,327,065	44,377,780	44,483,827
FR-EN	44,228,687	44,238,933	44,299,390	44,368,555	44,488,952
FR-ES	44,237,912	44,248,158	44,291,190	44,374,705	44,499,202
FR-RU	44,249,187	44,266,608	44,338,340	44,389,055	44,505,352
FR-ZH	44,768,862	45,422,808	46,498,015	47,682,380	49,245,977
RU-AR	44,240,470	44,272,751	44,342,425	44,383,924	44,485,875
RU-EN	44,233,295	44,246,101	44,314,750	44,374,699	44,491,000
RU-ES	44,242,520	44,255,326	44,306,550	44,380,849	44,501,250
RU-FR	44,244,570	44,259,426	44,322,950	44,382,899	44,503,300
RU-ZH	44,773,470	45,429,976	46,513,375	47,688,524	49,248,025
ZH-AR	44,500,054	44,850,287	45,421,209	46,028,980	46,853,875
ZH-EN	44,492,879	44,823,637	45,393,534	46,019,755	46,859,000
ZH-ES	44,502,104	44,832,862	45,385,334	46,025,905	46,869,250
ZH-FR	44,504,154	44,836,962	45,401,734	46,027,955	46,871,300
ZH-RU	44,513,379	44,851,312	45,432,484	46,040,255	46,875,400
AR-ZH_pinyin	44,230,739	44,254,814	44,290,668	44,355,742	44,808,248
EN-ZH_pinyin	44,227,155	44,241,502	44,276,844	44,351,134	44,810,808
ES-ZH_pinyin	44,231,763	44,246,110	44,272,748	44,354,206	44,815,928
FR-ZH_pinyin	44,232,787	44,248,158	44,280,940	44,355,230	44,816,952
RU-ZH_pinyin	44,237,395	44,255,326	44,296,300	44,361,374	44,819,000
AR-ZH_wubi	44,252,264	44,271,214	44,310,143	44,372,142	44,785,698
EN-ZH_wubi	44,248,680	44,257,902	44,296,319	44,367,534	44,788,258
ES-ZH_wubi	44,253,288	44,262,510	44,292,223	44,370,606	44,793,378
FR-ZH_wubi	44,254,312	44,264,558	44,300,415	44,371,630	44,794,402
RU-ZH_wubi	44,258,920	44,271,726	44,315,775	44,377,774	44,796,450

Representation / Number of lines	BYTE 100	BYTE 1,000	BYTE 10,000	BYTE 100,000	BYTE 1,000,000
AR-EN	44,223,056	44,240,470	44,304,515	44,336,795	44,360,874
AR-ES	44,232,281	44,249,695	44,295,290	44,336,795	44,360,874
AR-FR	44,234,331	44,254,820	44,307,590	44,338,845	44,367,024
AR-RU	44,244,581	44,269,170	44,326,040	44,342,945	44,360,874
AR-ZH	44,285,581	44,302,995	44,331,165	44,350,120	44,388,549
EN-AR	44,221,517	44,247,139	44,306,054	44,333,717	44,358,822
EN-ES	44,233,817	44,243,039	44,293,754	44,339,867	44,362,922
EN-FR	44,235,867	44,248,164	44,306,054	44,341,917	44,369,072
EN-RU	44,246,117	44,262,514	44,324,504	44,346,017	44,362,922
EN-ZH	44,287,117	44,296,339	44,329,629	44,353,192	44,390,597
ES-AR	44,226,125	44,251,747	44,301,446	44,333,717	44,358,822
ES-EN	44,229,200	44,238,422	44,298,371	44,339,867	44,362,922
ES-FR	44,240,475	44,252,772	44,301,446	44,341,917	44,369,072
ES-RU	44,250,725	44,267,122	44,319,896	44,346,017	44,362,922
ES-ZH	44,291,725	44,300,947	44,325,021	44,353,192	44,390,597
FR-AR	44,227,149	44,254,307	44,307,590	44,334,741	44,361,894
FR-EN	44,230,224	44,240,982	44,304,515	44,340,891	44,365,994
FR-ES	44,239,449	44,250,207	44,295,290	44,340,891	44,365,994
FR-RU	44,251,749	44,269,682	44,326,040	44,347,041	44,365,994
FR-ZH	44,292,749	44,303,507	44,331,165	44,354,216	44,393,669
RU-AR	44,232,269	44,261,475	44,316,806	44,336,789	44,358,822
RU-EN	44,235,344	44,248,150	44,313,731	44,342,939	44,362,922
RU-ES	44,244,569	44,257,375	44,304,506	44,342,939	44,362,922
RU-FR	44,246,619	44,262,500	44,316,806	44,344,989	44,369,072
RU-ZH	44,297,869	44,310,675	44,340,381	44,356,264	44,390,597
ZH-AR	44,252,749	44,278,371	44,319,366	44,340,373	44,372,646
ZH-EN	44,255,824	44,265,046	44,316,291	44,346,523	44,376,746
ZH-ES	44,265,049	44,274,271	44,307,066	44,346,523	44,376,746
ZH-FR	44,267,099	44,279,396	44,319,366	44,348,573	44,382,896
ZH-RU	44,277,349	44,293,746	44,337,816	44,352,673	44,376,746
EN-AR_cp1256	44,230,742	44,259,439	44,325,529	44,365,492	44,437,747
ES-AR_cp1256	44,235,350	44,264,047	44,320,921	44,365,492	44,437,747
FR-AR_cp1256	44,236,374	44,266,607	44,327,065	44,366,516	44,440,819
RU-AR_cp1256	44,241,494	44,273,775	44,336,281	44,368,564	44,437,747
ZH-AR_cp1256	44,261,974	44,290,671	44,338,841	44,372,148	44,451,571
AR-RU_cp1251	44,242,531	44,267,120	44,338,340	44,373,695	44,457,224
EN-RU_cp1251	44,244,067	44,260,464	44,336,804	44,376,767	44,459,272
ES-RU_cp1251	44,248,675	44,265,072	44,332,196	44,376,767	44,459,272
FR-RU_cp1251	44,249,699	44,267,632	44,338,340	44,377,791	44,462,344
ZH-RU_cp1251	44,275,299	44,291,696	44,350,116	44,383,423	44,473,096

Representation / Number of lines	WORD 100	WORD 1,000	WORD 10,000	WORD 100,000	WORD 1,000,000
AR-EN	45,247,147	50,481,885	69,784,815	132,473,233	325,770,330
AR-ES	45,275,847	51,042,560	72,964,365	145,014,108	346,631,130
AR-FR	45,311,722	51,085,610	71,818,415	138,219,383	321,411,005
AR-RU	45,456,247	52,676,410	81,458,540	177,653,183	442,242,105
AR-ZH	45,193,847	50,368,110	70,935,890	142,392,158	338,106,205
EN-AR	45,343,078	51,803,373	77,015,037	163,629,262	419,403,603
EN-ES	45,180,103	49,723,648	65,748,237	113,918,812	253,180,378
EN-FR	45,215,978	49,766,698	64,602,287	107,124,087	227,960,253
EN-RU	45,360,503	51,357,498	74,242,412	146,557,887	348,791,353
EN-ZH	45,098,103	49,049,198	63,719,762	111,296,862	244,655,453
ES-AR	45,357,414	52,083,437	78,603,261	169,893,582	429,823,827
ES-EN	45,165,739	49,443,037	64,156,911	107,642,257	242,739,802
ES-FR	45,230,314	50,046,762	66,190,511	113,388,407	238,380,477
ES-RU	45,374,839	51,637,562	75,830,636	152,822,207	359,211,577
ES-ZH	45,112,439	49,329,262	65,307,986	117,561,182	255,075,677
FR-AR	45,375,334	52,104,941	78,030,845	166,499,534	417,226,067
FR-EN	45,183,659	49,464,541	63,584,495	104,248,209	230,142,042
FR-ES	45,212,359	50,025,216	66,764,045	116,789,084	251,002,842
FR-RU	45,392,759	51,659,066	75,258,220	149,428,159	346,613,817
FR-ZH	45,130,359	49,350,766	64,735,570	114,167,134	242,477,917
RU-AR	45,447,526	52,899,565	82,846,205	186,197,198	477,582,675
RU-EN	45,255,851	50,259,165	68,399,855	123,945,873	290,498,650
RU-ES	45,284,551	50,819,840	71,579,405	136,486,748	311,359,450
RU-FR	45,320,426	50,862,890	70,433,455	129,692,023	286,139,325
RU-ZH	45,202,551	50,145,390	69,550,930	133,864,798	302,834,525
ZH-AR	45,316,454	51,746,541	77,590,013	168,583,886	425,565,523
ZH-EN	45,124,779	49,106,141	63,143,663	106,332,561	238,481,498
ZH-ES	45,153,479	49,666,816	66,323,213	118,873,436	259,342,298
ZH-FR	45,189,354	49,709,866	65,177,263	112,078,711	234,122,173
ZH-RU	45,333,879	51,300,666	74,817,388	151,512,511	354,953,273

Representation / Number of lines	BPE 100	BPE 1,000	BPE 10,000	BPE 100,000	BPE 1,000,000
AR-EN	45,127,305	49,170,732	63,907,829	86,155,852	89,242,512
AR-ES	45,143,705	49,556,132	65,812,279	87,922,952	89,628,937
AR-FR	45,207,255	49,516,157	65,451,479	87,395,077	89,514,137
AR-RU	45,329,230	50,670,307	71,872,079	89,287,227	89,999,987
AR-ZH	45,286,180	49,163,557	64,349,604	87,283,352	89,324,512
EN-AR	45,155,520	49,845,327	68,000,030	87,962,638	89,700,621
EN-ES	45,115,545	48,882,852	61,728,055	86,119,688	89,171,721
EN-FR	45,179,095	48,842,877	61,367,255	85,591,813	89,056,921
EN-RU	45,301,070	49,997,027	67,787,855	87,483,963	89,542,771
EN-ZH	45,258,020	48,490,277	60,265,380	85,480,088	88,867,296
ES-AR	45,163,712	50,037,839	68,951,326	88,845,326	89,893,645
ES-EN	45,107,337	48,689,964	60,774,901	85,235,276	88,978,320
ES-FR	45,187,287	49,035,389	62,318,551	86,474,501	89,249,945
ES-RU	45,309,262	50,189,539	68,739,151	88,366,651	89,735,795
ES-ZH	45,266,212	48,682,789	61,216,676	86,362,776	89,060,320
FR-AR	45,195,456	50,017,871	68,771,102	88,581,646	89,836,301
FR-EN	45,139,081	48,669,996	60,594,677	84,971,596	88,920,976
FR-ES	45,155,481	49,055,396	62,499,127	86,738,696	89,307,401
FR-RU	45,341,006	50,169,571	68,558,927	88,102,971	89,678,451
FR-ZH	45,297,956	48,662,821	61,036,452	86,099,096	89,002,976
RU-AR	45,256,384	50,594,383	71,978,270	89,526,798	90,078,989
RU-EN	45,200,009	49,246,508	63,801,845	85,916,748	89,163,664
RU-ES	45,216,409	49,631,908	65,706,295	87,683,848	89,550,089
RU-FR	45,279,959	49,591,933	65,345,495	87,155,973	89,435,289
RU-ZH	45,358,884	49,239,333	64,243,620	87,044,248	89,245,664
ZH-AR	45,234,880	49,841,743	68,220,702	88,525,838	89,741,581
ZH-EN	45,178,505	48,493,868	60,044,277	84,915,788	88,826,256
ZH-ES	45,194,905	48,879,268	61,948,727	86,682,888	89,212,681
ZH-FR	45,258,455	48,839,293	61,587,927	86,155,013	89,097,881
ZH-RU	45,380,430	49,993,443	68,008,527	88,047,163	89,583,731

Number of model parameters 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 PARAMS																				
AR-EN	44,230,224	44,277,863	44,328,071	44,430,017	44,680,027	44,233,811	44,276,843	44,321,419	44,356,263	44,385,977	45,583,728	51,799,608	72,558,944	137,925,984	347,956,340	45,326,100	49,857,876	64,758,280	86,794,440	89,354,238
AR-ES	44,240,474	44,282,988	44,331,146	44,430,017	44,703,602	44,244,061	44,281,968	44,319,369	44,354,213	44,390,077	45,680,078	52,442,283	75,710,819	146,661,034	361,865,590	45,356,850	50,162,301	66,575,605	88,073,640	89,654,563
AR-FR	44,243,549	44,291,188	44,344,471	44,443,342	44,689,252	44,246,111	44,294,268	44,328,594	44,355,238	44,390,077	45,664,703	52,299,808	74,587,419	140,492,584	339,724,565	45,354,800	50,126,426	66,297,830	87,741,540	89,551,038
AR-RU	44,264,049	44,321,938	44,371,121	44,445,392	44,656,452	44,267,636	44,315,793	44,346,019	44,354,213	44,384,952	45,856,378	54,230,908	85,165,419	180,046,309	461,160,415	45,540,325	51,314,401	73,041,305	89,296,465	90,003,063
AR-ZH	44,932,349	45,790,763	46,655,846	47,787,917	49,151,077	44,278,911	44,328,093	44,352,169	44,368,563	44,393,152	45,605,253	51,813,958	73,639,294	141,968,584	347,214,240	45,617,200	50,174,601	65,079,105	87,398,165	89,281,463
EN-AR	44,235,867	44,295,818	44,347,052	44,461,823	44,725,171	44,238,428	44,287,616	44,327,575	44,354,211	44,385,977	45,713,517	53,118,531	80,029,250	165,947,583	435,393,599	45,339,438	50,370,876	68,884,852	88,261,620	89,754,891
EN-ES	44,234,842	44,265,068	44,312,202	44,398,273	44,658,546	44,239,453	44,271,216	44,313,225	44,356,261	44,390,077	45,550,542	51,125,931	68,255,075	118,694,058	274,598,774	45,343,538	49,650,301	62,457,077	86,609,320	89,254,691
EN-FR	44,237,917	44,273,268	44,325,527	44,411,598	44,644,196	44,241,503	44,283,516	44,322,450	44,357,286	44,390,077	45,535,167	50,983,456	67,131,675	112,525,608	252,457,749	45,341,488	49,614,426	62,179,302	86,277,220	89,151,166
EN-RU	44,258,417	44,304,018	44,352,177	44,413,648	44,611,396	44,263,028	44,305,041	44,339,875	44,356,261	44,384,952	45,726,842	52,914,556	77,709,675	152,079,333	373,893,599	45,527,013	50,802,401	68,922,777	87,832,145	89,603,191
EN-ZH	44,926,717	45,772,843	46,636,902	47,756,173	49,106,021	44,274,303	44,317,341	44,346,025	44,370,611	44,393,152	45,475,717	50,497,606	66,183,550	114,001,608	259,947,424	45,603,888	49,662,601	60,960,577	85,933,845	88,881,591
ES-AR	44,240,987	44,298,378	44,348,588	44,461,823	44,736,947	44,243,548	44,290,176	44,326,551	44,353,187	44,388,025	45,761,645	53,439,555	81,603,650	170,310,847	442,341,439	45,354,798	50,522,940	69,792,628	88,900,596	89,904,907
ES-EN	44,229,712	44,262,503	44,310,663	44,398,273	44,646,747	44,234,323	44,268,651	44,314,251	44,357,287	44,388,025	45,502,320	50,804,280	66,677,600	114,322,272	267,637,364	45,328,148	49,497,940	61,547,528	85,969,096	89,104,382
ES-FR	44,243,037	44,275,828	44,327,063	44,411,598	44,655,972	44,246,623	44,286,076	44,321,426	44,356,262	44,392,125	45,583,295	51,304,480	68,706,075	116,888,872	259,405,589	45,356,848	49,766,490	63,087,078	86,916,196	89,301,182
ES-RU	44,263,537	44,306,578	44,353,713	44,413,648	44,623,172	44,268,148	44,307,601	44,338,851	44,355,237	44,387,000	45,774,970	53,235,580	79,284,075	156,442,597	380,841,439	45,542,373	50,954,465	69,830,553	88,471,121	89,753,207
ES-ZH	44,931,837	45,775,403	46,638,438	47,756,173	49,117,797	44,279,423	44,319,901	44,345,001	44,369,587	44,395,200	45,523,845	50,818,630	67,757,950	118,364,872	266,895,264	45,619,248	49,814,665	61,868,353	86,572,821	89,031,607
FR-AR	44,242,523	44,302,474	44,355,244	44,468,479	44,729,779	44,244,572	44,296,320	44,331,159	44,353,699	44,388,025	45,753,965	53,368,387	81,042,498	167,229,631	431,281,727	45,353,774	50,505,020	69,653,876	88,734,708	89,853,195
FR-EN	44,231,248	44,266,599	44,317,319	44,404,929	44,639,579	44,235,347	44,274,795	44,318,859	44,357,799	44,388,025	45,494,640	50,733,112	66,116,448	111,241,056	256,577,652	45,327,124	49,480,020	61,408,776	85,803,208	89,052,670
FR-ES	44,241,498	44,271,724	44,320,394	44,404,929	44,663,154	44,245,597	44,279,920	44,316,809	44,355,749	44,392,125	45,590,990	51,375,787	69,268,323	119,976,106	270,486,902	45,357,874	49,784,445	63,226,101	87,082,408	89,352,995
FR-RU	44,265,073	44,310,674	44,360,369	44,420,304	44,616,004	44,269,172	44,313,745	44,343,459	44,355,749	44,387,000	45,767,290	53,164,412	78,722,923	153,361,381	369,781,727	45,541,349	50,936,545	69,691,801	88,305,233	89,701,495
FR-ZH	44,933,373	45,779,499	46,645,094	47,762,829	49,110,629	44,280,447	44,326,045	44,349,609	44,370,099	44,395,200	45,516,165	50,747,462	67,196,798	115,283,656	255,835,552	45,618,224	49,796,745	61,729,601	86,406,933	88,979,895
RU-AR	44,252,763	44,317,834	44,368,556	44,469,503	44,713,395	44,255,324	44,307,072	44,339,863	44,353,187	44,385,465	45,849,709	54,332,995	86,326,338	186,987,199	491,940,415	45,446,446	51,098,428	73,022,324	89,511,412	90,078,987
RU-EN	44,241,488	44,281,959	44,330,631	44,405,953	44,623,195	44,246,099	44,285,547	44,327,563	44,357,287	44,385,465	45,590,384	51,697,720	71,400,288	130,998,624	317,236,340	45,419,796	50,073,428	64,777,224	86,579,912	89,278,462
RU-ES	44,251,738	44,287,084	44,333,706	44,405,953	44,646,770	44,256,349	44,290,672	44,325,513	44,355,237	44,389,565	45,686,734	52,340,395	74,552,163	139,733,674	331,145,590	45,450,546	50,377,853	66,594,549	87,859,112	89,578,787
RU-FR	44,254,813	44,295,284	44,347,031	44,419,278	44,632,420	44,258,399	44,302,972	44,334,738	44,356,262	44,389,565	45,671,359	52,197,920	73,428,763	133,565,224	309,004,565	45,448,496	50,341,978	66,316,774	87,527,012	89,475,262
RU-ZH	44,943,613	45,794,859	46,658,406	47,763,853	49,094,245	44,291,199	44,336,797	44,358,313	44,369,587	44,392,640	45,611,909	51,712,070	72,480,638	135,041,224	316,494,240	45,710,896	50,390,153	65,098,049	87,183,637	89,205,687
ZH-AR	44,586,587	45,051,530	45,509,804	46,139,135	46,958,515	44,260,956	44,313,216	44,342,935	44,360,355	44,389,561	45,724,269	53,125,699	80,568,898	167,966,911	435,022,911	45,484,846	50,529,084	69,045,108	88,563,188	89,718,539
ZH-EN	44,575,312	45,015,655	45,471,879	46,075,585	46,868,315	44,251,731	44,291,691	44,330,635	44,364,455	44,389,561	45,464,944	50,490,424	65,642,848	111,978,336	260,318,836	45,458,196	49,504,084	60,800,008	85,631,688	88,918,014
ZH-ES	44,585,562	45,020,780	45,474,954	46,075,585	46,891,890	44,261,981	44,296,816	44,328,585	44,362,405	44,393,661	45,561,294	51,133,099	68,794,723	120,713,386	274,228,086	45,488,946	49,808,509	62,617,333	86,910,888	89,218,339
ZH-FR	44,588,637	45,028,980	45,488,279	46,088,910	46,877,540	44,264,031	44,309,116	44,337,810	44,363,430	44,393,661	45,545,919	50,990,624	67,671,323	114,544,936	252,087,061	45,486,896	49,772,634	62,339,558	86,578,788	89,114,814
ZH-RU	44,609,137	45,059,730	45,514,929	46,090,960	46,844,740	44,285,556	44,330,641	44,355,235	44,362,405	44,388,536	45,737,594	52,921,724	78,249,323	154,098,661	373,522,911	45,672,421	50,960,609	69,083,033	88,133,713	89,566,839
AR-ZH_pinyin	44,225,099	44,282,988	44,337,296	44,483,317	44,769,202															
EN-ZH_pinyin	44,219,467	44,265,068	44,318,352	44,451,573	44,724,146															
ES-ZH_pinyin	44,224,587	44,267,628	44,319,888	44,451,573	44,735,922															
FR-ZH_pinyin	44,226,123	44,271,724	44,326,544	44,458,229	44,728,754															
RU-ZH_pinyin	44,236,363	44,287,084	44,339,856	44,459,253	44,712,370															
AR-ZH_wubi	44,241,499	44,303,488	44,351,646	44,488,442	44,759,977															
EN-ZH_wubi	44,235,867	44,285,568	44,332,702	44,456,698	44,714,921															
ES-ZH_wubi	44,240,987	44,288,128	44,334,238	44,456,698	44,726,697															
FR-ZH_wubi	44,242,523	44,292,224	44,340,894	44,463,354	44,719,529															
RU-ZH_wubi	44,252,763	44,307,584	44,354,206	44,464,378	44,703,145															
EN-AR_cp1256						44,237,403	44,297,866	44,349,100	44,448,511	44,642,227										
ES-AR_cp1256						44,242,523	44,300,426	44,348,076	44,447,487	44,644,275										
FR-AR_cp1256						44,243,547	44,306,570	44,352,684	44,447,999	44,644,275										
RU-AR_cp1256						44,254,299	44,317,322	44,361,388	44,447,487	44,641,715										
ZH-AR_cp1256						44,259,931	44,323,466	44,364,460	44,454,655	44,645,811										
AR-RU_cp1251						44,264,561	44,316,818	44,360,369	44,398,288	44,528,452										
EN-RU_cp1251						44,259,953	44,306,066	44,354,225	44,400,336	44,528,452										
ES-RU_cp1251						44,265,073	44,308,626	44,353,201	44,399,312	44,530,500										
FR-RU_cp1251						44,266,097	44,314,770	44,357,809	44,399,824	44,530,500										
ZH-RU_cp1251						44,282,481	44,331,666	44,369,585	44,406,480	44,532,036										

Number of model parameters 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 PARAMS																				
AR-EN	44,237,394	44,277,862	44,319,874	44,432,062	44,644,154	44,237,909	44,276,331	44,315,783	44,352,675	44,380,342	45,632,910	51,878,481	72,594,782	137,710,265	347,843,439	45,372,198	49,998,240	64,770,100	86,735,506	89,389,603
AR-ES	44,243,544	44,288,112	44,327,049	44,432,062	44,640,054	44,242,009	44,285,556	44,317,833	44,353,700	44,377,267	45,698,510	52,453,506	75,866,582	146,529,365	361,666,589	45,373,223	50,205,290	66,481,850	88,137,706	89,650,978
AR-FR	44,246,619	44,293,237	44,333,199	44,441,287	44,664,654	44,248,159	44,290,681	44,318,858	44,352,675	44,381,367	45,720,035	52,319,231	74,570,982	140,332,215	338,956,689	45,424,473	50,212,465	66,174,350	87,724,631	89,533,103
AR-RU	44,257,894	44,321,937	44,361,899	44,457,687	44,620,579	44,257,384	44,314,256	44,339,358	44,359,850	44,379,317	45,898,385	54,253,406	85,360,132	180,206,765	460,198,814	45,577,198	51,406,590	73,204,825	89,355,406	90,003,578
AR-ZH	44,960,019	45,766,162	46,694,799	47,805,337	49,141,854	44,286,084	44,326,556	44,347,558	44,371,125	44,391,617	45,608,310	51,865,156	73,693,582	141,800,015	346,543,739	45,661,248	50,236,040	65,194,450	87,429,431	89,281,978
EN-AR	44,247,141	44,297,356	44,338,342	44,470,537	44,704,175	44,243,552	44,286,591	44,322,452	44,353,188	44,379,316	45,765,777	53,237,931	80,104,076	165,772,904	434,781,036	45,404,004	50,534,838	68,840,755	88,226,797	89,768,710
EN-ES	44,233,816	44,268,656	44,308,617	44,393,662	44,580,150	44,236,377	44,275,316	44,311,177	44,353,188	44,378,291	45,565,902	51,096,706	68,371,926	118,521,429	274,898,461	45,341,479	49,669,738	62,419,130	86,649,322	89,272,610
EN-FR	44,236,891	44,273,781	44,314,767	44,402,887	44,604,750	44,242,527	44,280,441	44,312,202	44,352,163	44,382,391	45,587,427	50,962,431	67,076,326	112,324,279	252,188,561	45,392,729	49,676,913	62,111,630	86,236,247	89,154,735
EN-RU	44,248,166	44,302,481	44,343,467	44,419,287	44,560,675	44,251,752	44,304,016	44,332,702	44,359,338	44,380,341	45,765,777	52,896,606	77,865,476	152,198,829	373,430,686	45,545,454	50,871,038	69,142,105	87,867,022	89,625,210
EN-ZH	44,950,291	45,746,706	46,676,367	47,766,937	49,081,950	44,280,452	44,316,316	44,340,902	44,370,613	44,392,641	45,475,702	50,508,356	66,198,926	113,792,079	259,775,611	45,629,504	49,700,488	61,131,730	85,941,047	88,903,610
ES-AR	44,250,213	44,302,476	44,341,926	44,470,537	44,702,127	44,245,600	44,291,199	44,323,476	44,353,700	44,377,780	45,798,545	53,525,163	81,738,380	170,178,152	441,685,868	45,404,516	50,638,262	69,695,795	88,927,213	89,899,270
ES-EN	44,230,738	44,263,526	44,305,026	44,393,662	44,582,202	44,234,325	44,270,699	44,310,151	44,352,675	44,379,830	45,533,070	50,808,913	66,734,430	114,107,577	267,980,143	45,340,966	49,566,112	61,562,420	85,947,538	89,141,795
ES-FR	44,239,963	44,278,901	44,318,351	44,402,887	44,602,702	44,244,575	44,285,049	44,313,226	44,352,675	44,380,855	45,620,195	51,249,663	68,710,630	116,729,527	259,093,393	45,393,241	49,780,337	62,966,670	86,936,663	89,285,295
ES-RU	44,251,238	44,307,601	44,347,051	44,419,287	44,558,627	44,253,800	44,308,624	44,333,726	44,359,850	44,378,805	45,798,545	53,183,838	79,499,780	156,604,077	380,335,518	45,545,966	50,974,462	69,997,145	88,567,438	89,755,770
ES-ZH	44,953,363	45,751,826	46,679,951	47,766,937	49,079,902	44,282,500	44,320,924	44,341,926	44,371,125	44,391,105	45,508,470	50,795,588	67,833,230	118,197,327	266,680,443	45,630,016	49,803,912	61,986,770	86,641,463	89,034,170
FR-AR	44,251,749	44,305,036	44,344,998	44,475,145	44,714,415	44,248,672	44,293,759	44,323,988	44,353,188	44,379,828	45,809,297	53,458,091	81,091,212	167,082,600	430,341,996	45,430,116	50,641,846	69,542,195	88,720,877	89,840,390
FR-EN	44,232,274	44,266,086	44,308,098	44,398,270	44,594,490	44,237,397	44,273,259	44,310,663	44,352,163	44,381,878	45,543,822	50,741,841	66,087,262	111,012,025	256,636,271	45,366,566	49,569,696	61,408,820	85,741,202	89,082,915
FR-ES	44,238,424	44,276,336	44,315,273	44,398,270	44,590,390	44,241,497	44,282,484	44,312,713	44,353,188	44,378,803	45,609,422	51,316,866	69,359,062	119,831,125	270,459,421	45,367,591	49,776,746	63,120,570	87,143,402	89,344,290
FR-RU	44,252,774	44,310,161	44,350,123	44,423,895	44,570,915	44,256,872	44,311,184	44,334,238	44,359,338	44,380,853	45,809,297	53,116,766	78,852,612	153,508,525	368,991,646	45,571,566	50,978,046	69,843,545	88,361,102	89,696,890
FR-ZH	44,954,899	45,754,386	46,683,023	47,771,545	49,092,190	44,285,572	44,323,484	44,342,438	44,370,613	44,393,153	45,519,222	50,728,516	67,186,062	115,101,775	255,336,571	45,655,616	49,807,496	61,833,170	86,435,127	88,975,290
RU-AR	44,257,381	44,319,372	44,359,334	44,483,337	44,692,399	44,253,280	44,305,535	44,334,228	44,356,772	44,378,804	45,898,385	54,424,235	86,480,524	187,000,424	490,903,916	45,506,404	51,238,326	73,054,003	89,535,469	90,075,398
RU-EN	44,237,906	44,280,422	44,322,434	44,406,462	44,572,474	44,242,005	44,285,035	44,320,903	44,355,747	44,380,854	45,632,910	51,707,985	71,476,574	130,929,849	317,198,191	45,442,854	50,166,176	64,920,628	86,555,794	89,317,923
RU-ES	44,244,056	44,290,672	44,329,609	44,406,462	44,568,374	44,246,105	44,294,260	44,322,953	44,356,772	44,377,779	45,698,510	52,283,010	74,748,374	139,748,949	331,021,341	45,443,879	50,373,226	66,632,378	87,957,994	89,579,298
RU-FR	44,247,131	44,295,797	44,335,759	44,415,687	44,592,974	44,252,255	44,299,385	44,323,978	44,355,747	44,381,879	45,720,035	52,148,735	73,452,774	133,551,799	308,311,441	45,495,129	50,380,401	66,324,878	87,544,919	89,461,423
RU-ZH	44,960,531	45,768,722	46,697,359	47,779,737	49,070,174	44,290,180	44,335,260	44,352,678	44,374,197	44,392,129	45,608,310	51,694,660	72,575,374	135,019,599	315,898,491	45,731,904	50,403,976	65,344,978	87,249,719	89,210,298
ZH-AR	44,608,101	45,040,780	45,524,646	46,155,529	46,950,831	44,267,616	44,311,679	44,338,324	44,362,404	44,384,948	45,753,489	53,231,275	80,652,940	167,815,784	434,131,820	45,548,388	50,653,622	69,052,723	88,573,421	89,714,950
ZH-EN	44,588,626	45,001,830	45,487,746	46,078,654	46,830,906	44,256,341	44,291,179	44,324,999	44,361,379	44,386,998	45,488,014	50,515,025	65,648,990	111,745,209	260,426,095	45,484,838	49,581,472	60,919,348	85,593,746	88,957,475
ZH-ES	44,594,776	45,012,080	45,494,921	46,078,654	46,826,806	44,260,441	44,300,404	44,327,049	44,362,404	44,383,923	45,553,614	51,090,050	68,920,790	120,564,309	274,249,245	45,485,863	49,788,522	62,631,098	86,995,946	89,218,850
ZH-FR	44,597,851	45,017,205	45,501,071	46,087,879	46,851,406	44,266,591	44,305,529	44,328,074	44,361,379	44,388,023	45,575,139	50,955,775	67,625,190	114,367,159	251,539,345	45,537,113	49,795,697	62,323,598	86,582,871	89,100,975
ZH-RU	44,609,126	45,045,905	45,529,771	46,104,279	46,807,331	44,275,816	44,329,104	44,348,574	44,368,554	44,385,973	45,753,489	52,889,950	78,414,340	154,241,709	372,781,470	45,689,838	50,989,822	69,354,073	88,213,646	89,571,450
AR-ZH_pinyin	44,233,294	44,287,087	44,333,199	44,487,412	44,749,729															
EN-ZH_pinyin	44,223,566	44,267,631	44,314,767	44,449,012	44,689,825															
ES-ZH_pinyin	44,226,638	44,272,751	44,318,351	44,449,012	44,687,777															
FR-ZH_pinyin	44,228,174	44,275,311	44,321,423	44,453,620	44,700,065															
RU-ZH_pinyin	44,233,806	44,289,647	44,335,759	44,461,812	44,678,049															
AR-ZH_wubi	44,254,819	44,309,637	44,348,574	44,489,462	44,744,604															
EN-ZH_wubi	44,245,091	44,290,181	44,330,142	44,451,062	44,684,700															
ES-ZH_wubi	44,248,163	44,295,301	44,333,726	44,451,062	44,682,652															
FR-ZH_wubi	44,249,699	44,297,861	44,336,798	44,455,670	44,694,940															
RU-ZH_wubi	44,255,331	44,312,197	44,351,134	44,463,862	44,672,924															
EN-AR_cp1256						44,248,677	44,299,916	44,340,902	44,456,713	44,636,591										
ES-AR_cp1256						44,250,725	44,304,524	44,341,926	44,457,225	44,635,055										
FR-AR_cp1256						44,253,797	44,307,084	44,342,438	44,456,713	44,637,103										
RU-AR_cp1256						44,258,405	44,318,860	44,352,678	44,460,297	44,636,079										
ZH-AR_cp1256						44,272,741	44,325,004	44,356,774	44,465,929	44,642,223										
AR-RU_cp1251						44,255,334	44,315,281	44,352,683	44,405,975	44,492,067										
EN-RU_cp1251						44,249,702	44,305,041	44,346,027	44,405,463	44,493,091										
ES-RU_cp1251						44,251,750	44,309,649	44,347,051	44,405,975	44,491,555										
FR-RU_cp1251						44,254,822	44,312,209	44,347,563	44,405,463	44,493,603										
ZH-RU_cp1251						44,273,766	44,330,129	44,361,899	44,414,679	44,498,723										

M ERRATICITY

Length has been an issue since the dawn of the encoder-decoder approach for NMT (Cho et al., 2014). Most work on length bias, except for that by e.g. Sountsov & Sarawagi (2016), seems to have focused on the evaluation of generated translation output and monitored performance degradation with respect to sequence length, often arguing that beam size plays a role (Koehn & Knowles, 2017; Murray & Chiang, 2018). (Related work in Stahlberg & Byrne (2019) provides a good summary on this issue.) While there could also be confounds in search, our experiments show that a kind of length bias can surface already with CLMing, without generation taking place. To our knowledge, length bias has not been expressed as a sample-wise non-monotonicity across a large data size range as ours. While the connection between erraticity in CLMs and length bias in NMT models remains to be verified on a case-by-case basis, the knowledge of length also contributing to robustness (not just consistently poor/poorer performance) could support further experimentation/replication of any study. Failed attempts to reproduce results may be explainable by erraticity.

One may argue that erraticity may not be relevant when each model is more optimally trained (as opposed to being treated with our one-setting-for-all regime). But we do want to stress that this very stark contrast between erratic and non-erratic behavior is possible, prompting a question on fairness: is there a one-for-all setting under which the languages with non-erratic behavior shown in our study would demonstrate erraticity and vice versa?

To the best of our knowledge, the meta phenomenon of erraticity, as a sample-wise non-monotonicity measured intrinsically with cross-entropy and contributing to large variance across runs, is a novel and original discovery and contribution to research in robustness. We hope our work would inspire further evaluation on other models/architectures, reflection and theories on our assumption of unbounded computation (e.g. Xu et al. (2020)), as well as new understanding and solutions that take data statistics and realistic computational aspects into account. We defer a more comprehensive analysis of erraticity with further experiments to future work.

M.1 ERRATICITY AS LARGE VARIANCE: EVIDENCE FROM DIFFERENT RUNS OF THE SAME DATA

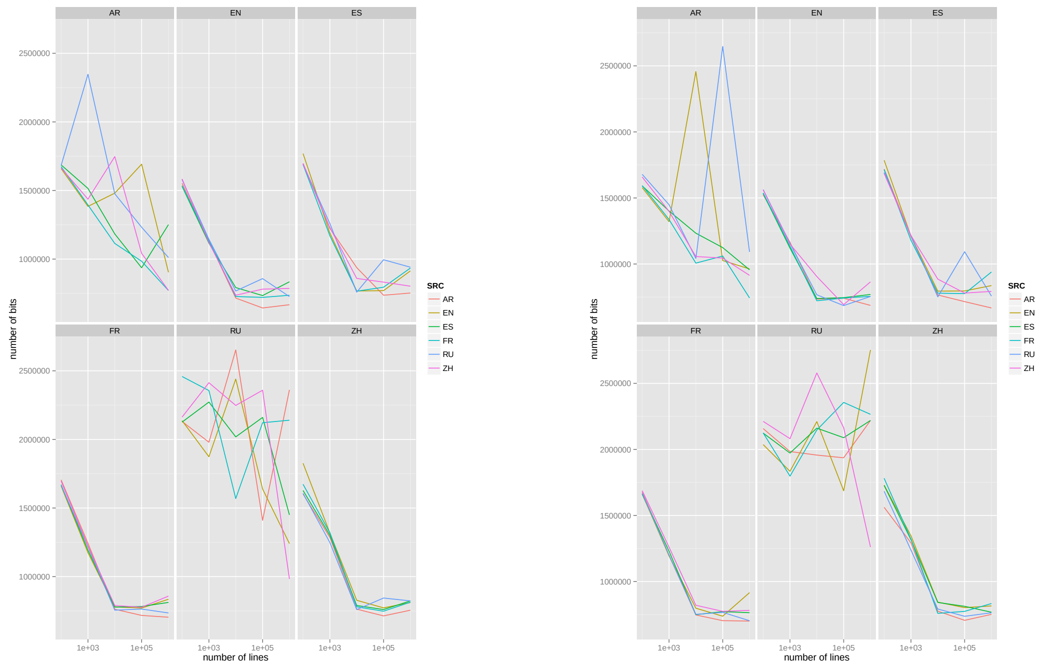
To confirm that erraticity is not due to data-specific reasons, e.g. when certain data segments might be "easier" to model than others, we show figures from 2 runs (Figs. 18a and 18b) on the same dataset of wildly differing performance that only differ in seed. Note that changes in the y-direction can vary much, indicating large variance across runs.

By establishing that high variance holds across sample sizes, we showcased how it'd be possible to just test on 2 or 3 data points of smaller sizes to get a gauge on the robustness in higher order. It serves as a signal of when the system is being "stress-tested" and hyperparameters need re-tuning. Spot-testing on a couple of smaller data sizes can indeed save much time and energy. Take our run B0 byte models as an example: the training of the 10^2-line model for EN-RU took 15 minutes, 10^3 40 minutes, 10^4 1 hour 50 minutes, and 10^5 3 hours 36 minutes. One can imagine how these would just be a fraction of training time for bigger models. (Likewise, for our ratio of target training token count to number of parameters — knowing when a representation might be prone to DD within a data size range could help prevent practitioners from prematurely declaring experimental results as negative or from unnecessarily rerunning an experiment because bigger data did not lead to better results.)

M.2 ADDITIONAL EXPERIMENT WITH LENGTH FILTERING TO 300 BYTES

Figure 19a and 19b show results of additional experiment with subset of data in byte (UTF-8) representation length-filtered to 300, including dev data:

Erraticity remains for AR and RU. Scores are lower, though they cannot be compared with the experiments in the main paper due to difference in dev data size (3,077 lines vs. 1,804 lines here). Number of total lines for train is 5,533,672 lines for each language, from which we took the initial 10^2-10^6. As in our main experiments, we filtered out only whole lines, i.e. not by discarding the tails of longer lines. 300 bytes aren't long sequences, but without data transform or hyperparameter tuning, things can look unfair. The EN translation of the longest RU line in this dataset is: "47. It is

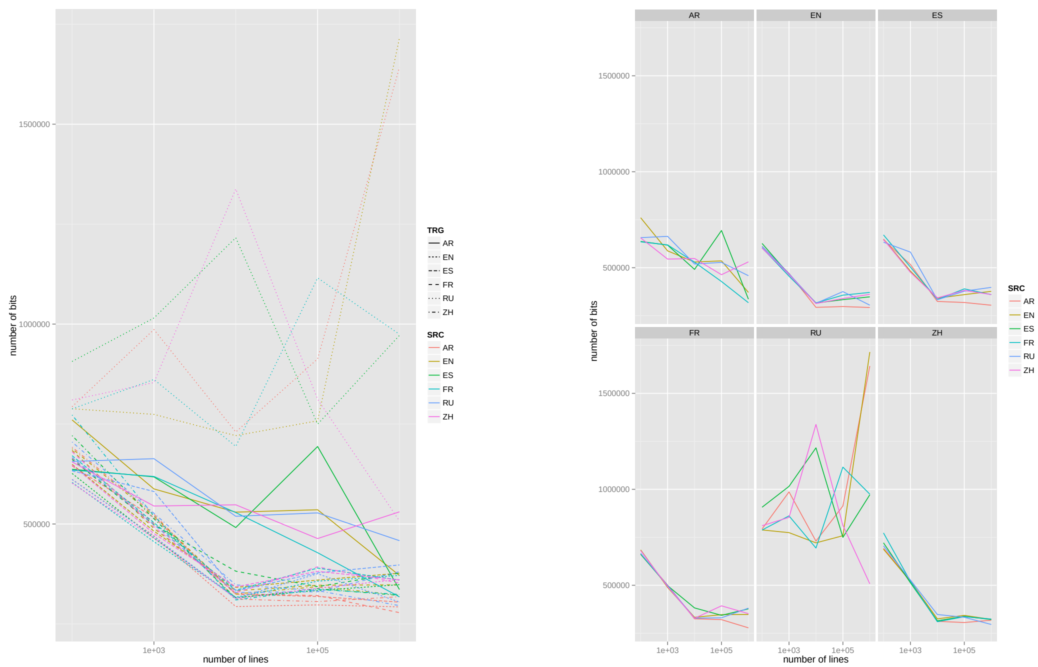


(a) run A0

(b) run A1

Figure 18: Same data with differing seeds

noted that there is a lack of information provided by the Government of Trinidad and Tobago with regard to the legal status of the Convention in the domestic legislation."



(a) (b)

Figure 19: Additional experiment with maximum length of 300 bytes (with no hyperparamter tuning, in our blind one-setting-for-all evaluation). Considering there are languages with much higher character sequence length than RU, there is food for thought for the design of next-generation Multilingual Plane.

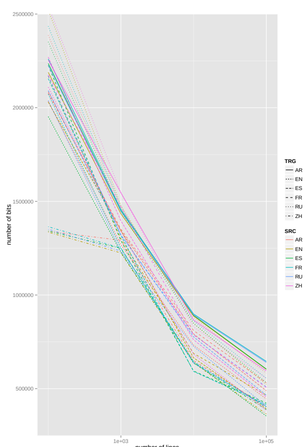
N EXPERIMENTS WITH ONE-LAYER TRANSFORMER

We performed 1 run with dataset A in 4 sizes (10^2-10^5 lines, seed=13) with the primary representations of characters, bytes, and words, on 1-layer Transformers (num-layers 1:1, all other hyperparameters remain the same as for our main experiments). We compared this against run A0 in 4 sizes with the same seed. (Based on how our null hypothesis is set up, the higher the number of runs, the more likely it is for there to be disparity. Important is that we evaluate based on an equal number of runs and on the same data for all candidates.) Results are shown in Table 5 with no statistically significant disparity observed on the models trained with 1 layer across the board.

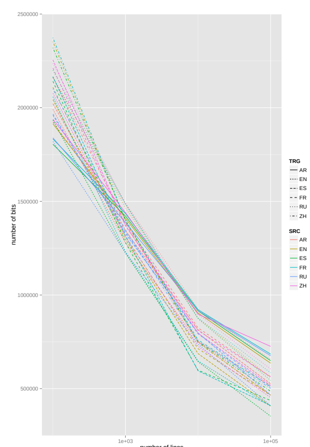
Many are under the impression that big data is the cause to the neutralization of language instances in DL/NNs. But, as this set of experiments shows, it is possible for there to be no statistically significant differences between them, with as little as our smallest data size of 100 lines.

Table 5: Number of language pairs out of 15 with significant differences, with respective p-values. BYTE$_{6layers}$ is the representation with erratic AR$_{trg}$ and RU$_{trg}$.

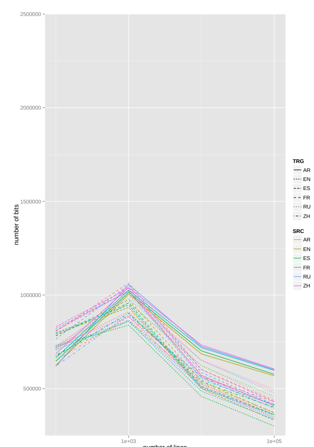
	CHAR$_{6layers}$		BYTE$_{6layers}$		WORD$_{6layers}$		CHAR$_{1layer}$		BYTE$_{1layer}$		WORD$_{1layer}$	
p-value	src	trg	src	trg	src	trg	src	trg	src	trg	src	trg
0.05	0	0	0	6	0	5	0	0	0	0	0	0
0.01	0	0	0	6	0	1	0	0	0	0	0	0
0.001	0	0	0	5	0	0	0	0	0	0	0	0



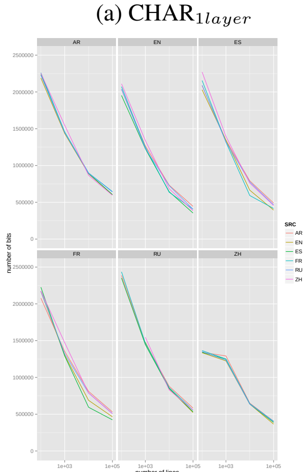
(a) CHAR$_{1layer}$



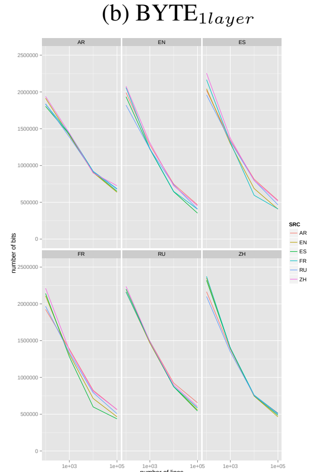
(b) BYTE$_{1layer}$



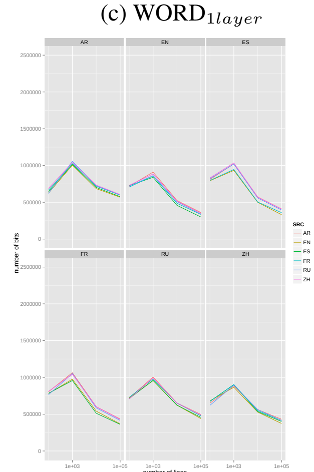
(c) WORD$_{1layer}$



(d) CHAR$_{1layer}$ by target



(e) BYTE$_{1layer}$ by target



(f) WORD$_{1layer}$ by target

Figure 20: One-layer Transformer models

O PAQS (PREVIOUSLY ASKED QUESTIONS)

O.1 ONE SETTING FOR ALL

Q: Normally, one trains a model with the objective of optimizing based on the training and evaluation data with hyperparameter tuning. The experiments here used one setting for all. Some model configurations might train better and converge close to their optima while other configurations might not reach their full potential. Can this not create a distortion in the results?

A: For conventional engineering practice, we agree that hyperparameter tuning would be a sine qua non. However, the evaluation objective is the relational distance between languages, hence we need to see it in a different light. Here is a loose analogy:

***

Assume 3 objects in 3 different locations in space.

Relative evaluation from one setting allows one to capture the distance between these objects. It does not matter whether these three objects are in their "best" states.

For example, if one were to use a camera to capture these 3 objects and one does not adjust the setting (using just one random aperture, shutter speed, and focus), i.e. no tuning to capture any of these 3 specifically, nor does one try to model these 3 to their individual bests separately, what would result could be a picture that captures one of these 3 objects more favorably than the others, or it could be that all of these would be blurred. But either way, there is a degree of blurriness to be measured, giving us an idea of the relative distance between the objects. Such relative measurement is the evaluation strategy that our paper adopts.

Now, to add to the camera analogy, say one of the objects is running water, which was extra blurry [erraticity]: we suggest freezing the water, so even from the one arbitrary angle, it could be captured better. And it worked.

Also, while one might generally like to have a "pretty" photo, one that is e.g. taken with sub-optimal lighting, say, overexposure, can have a telling effect as it can bring out details in something dark, like a black box.

***

Alternatively, one can tune hyperparameters for each model individually such that each model would be a more optimized one and then compare these models. In that case, one would be interpreting the differences between language in terms of hyperparameters, and the paper would be one that is algorithm-centric. That is of course also a possibility. Our approach, however, is a data-centric one. We would, first of all, like to understand the nature of language data, i.e. what it is about language, if there is anything at all, that makes it a different data type than other data, and what kind of structural constraints, if any, that we need to take into consideration. Then with findings from this data perspective, we try to relate back to the algorithm and make connections so to create a more holistic picture.

O.2 TRANSLATIONESE / WORD ORDER

Q: Multitexts are parallel texts or translations with the same meaning. There is little to no variation in word order, hence they are just "Translationese" (Gellerstam, 1986). That is why they turn out to be the same, with no performance disparity.

A: Our findings do show that when the semantics is properly controlled, such as in multitexts, the factors influencing performance are statistical properties related to sequence length and vocabulary, e.g. $|V|$ or TTR, and the languages tested can be different. Semantic equivalence is also not a reason why we should expect neutralization of source language instances, as that would mean we should expect equal results across target languages.

We agree that faithfulness is often a priority in producing good translations. Whether the translations are produced by humans or machines, only a single best translation can surface as the translation of choice. There may be many other competing hypotheses, but regardless of whether it is done through an automatic ranking algorithm by a machine or through a human expert, the purpose of translation is the same. However, *styles* and preferences in translations can vary. While faithfulness is generally preferred in the translations of legal texts, more freedom with skillful rearrangement of and play on words (or rather, character or sub-character sequences) or sounds being a criterion for literary texts could be appreciated by certain readers. We agree that it could be very interesting and necessary to model these variations, and we understand that languages can surface in many multimodal forms beyond the confines of texts as well. But with a data-driven perspective, to model this broader variation in language, we need corresponding datasets — we suggest contrast sets where the difference in e.g. sequential order is explicit. And for evaluation, we would require an even more systematic meta evaluation, one that spans different datasets.

But the argument that language or data *could be* different beyond how it appears in one dataset is irrelevant in the evaluation of experiments involving said dataset.

P UNDERSTANDING THE PHENOMENA WITH ALTERNATE REPRESENTATIONS (EXTENDED VERSION)

[Appendix P is an extended version of § 4.]

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

Character level On the character level, it is well known that ZH differs from the other languages in its high $|V|$, in this study it has an averaged mean$\pm$std of 2550$\pm$1449[12] across all 5 data sizes from all 3 datasets compared to 170$\pm$87 from all other 5 languages combined, may these be in Latin or Cyrillic alphabet or the Abjad script. But what is often not known is that the character sequence length of logographic languages such as ZH is typically short (think and compare the sequence length of the Ancient Egyptian hieroglyphs or the Demotic script with that of the Greek script on the Rosetta Stone). Here in our case, the averaged mean sequence length in characters for ZH is 35$\pm$19, compared to 129$\pm$71 from the other 5 languages. Heuristics to mitigate high $|V|$ often involve decomposition, which automatically resolve the problem of short sequence length. We tried 2 methods to lower character $|V|$ with representations in ASCII characters — Pinyin and Wubi. The former is a romanization of ZH characters based on their pronunciations and the latter is an input algorithm that decomposes character-internal information into stroke shape and ordering and matches these to 5 classes of radicals (Lunde, 2008). We replaced the ZH data with these formats *only on the target side* and reran the experiments involving ZH as a target language (ZH$_{trg}$) 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. But Wubi exhibits erraticity (Subfigure 2a). Wubi in our data has a maximum sequence length of 688 characters. As we shall also show in our byte-level analysis below, there are reasons to attribute length as cause to erraticity.

Decomposition into strokes may seem like a natural remedy analogous to decomposing an EN word into character sequences, but one needs to be mindful of not exceeding an optimal length given finite computation. Considering the ZH in the UN data is represented in simplified characters, decomposing traditional characters would surely complicate the problem. As there are also sub-character semantic and phonetic units (Zhang & Komachi, 2018) that can be exploited for information and aligned with character sequences of other alphabets, qualitative advances in this area can indeed be a new state of the art.

Byte level On the byte level, we observe irregularity for AR and RU. We find minimum sequence length of the target language to be one of the highest metrics correlating positively with the total number of bits ($\rho = 0.60$).[13] Our data is based on 300 characters as maximum length per line. While we wanted to retain at least 75% of the UN data after length filtering, this length still renders a maximum sequence length that exceeds 100 words (the default maximum length for the word alignment model, GIZA++ (Och & Ney, 2003), in the traditional SMT pipeline). Translated into bytes with UTF-8 encoding, data with 300 characters maximum gives us, e.g. for the 10^6-line datasets, an averaged mean$\pm$std of 185$\pm$106 in length for AR and 246$\pm$142 for RU, considerably larger than that for ZH (94$\pm$53) and for EN/ES/FR ($\approx$145.41$\pm$77). With UTF-8 encoding, each character in AR, RU, and ZH contains 2 or more bytes. ZH typically has shorter line length in characters, compensating for the total byte sequence in length, even when most ZH characters are 3 bytes each. However, AR and RU generally have long line length in characters, so when converted to bytes, the sequence length remains long even when most of the characters might be just 2 bytes each. Results from our pairwise comparisons indicate 8 (non-directional) language pairs to be significantly different (see Table 1 under "BYTE"): ES-RU, EN-RU, FR-RU, RU-ZH, AR-RU, AR-EN, AR-ZH, and AR-FR — all involving AR or RU. (Appendix I lists also the language pairs with significant differences for other representations.)

[12]Figures are rounded to whole number. Complete tables of data statistics are provided in Appendix D.
[13]Top-3 correlates for each representation can be found in Appendix F.

Leveraging language-specific code pages can be a useful practical trick, a reminder that there are alternatives to UTF-8 for analyses and back-end processing if data is clean and homogeneous and if success of larger-scale prediction is not a concern. But one more sustainable alternative is to design a more adaptive and flexible character encoding scheme in general, taking into account the statistical profiles such as length (wrt characters and bytes) and sub-character (atomic/elementary/compound) information of all (or as many as possible) of the world's languages.

Word level The main difference between word and character/byte models is the absence of length as a top contributing factor correlating with performance. Instead, what matters more are metrics concerning word vocabulary, with top correlate being OOV token rate in the target language ($\rho = 0.66$). This is understandable as word segmentation neutralizes sequence lengths — the longer lengths in phonetic alphabetic scripts are shortened through multiple-character groupings, while the shorter lengths in logographic scripts (cf. difference in length for the 3 scripts on the Rosetta Stone, logographic scripts are typically shorter than phonetic ones) are lengthened by the insertion of whitespaces. 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 for its capability to model subword units for morphologically complex languages (e.g. AR and RU). 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 ($p \leq 0.001$) reduced to 7 from 8 for word models, showing how finer-grained modeling has a positive effect on closing the disparity gap.

Version 1.1 (graphs to be updated, score tables added)