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 in one-to-one conditional language modeling through a series of systematically controlled experiments with the Transformer and parallel data on the 6 diverse, official languages of the United Nations — in 30 directions, 5 sizes, and 3 primary representation types in character, byte, and word, along with 5 alternate variants for a secondary set of controls. We 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, 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. We hope our work helps open up new directions in the area of language and computing that would be fairer and more flexible.

1 INTRODUCTION

With a transdisciplinary approach to explore a space at the intersection of Deep Learning (DL) / Neural Networks (NNs), domain science, 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,

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

high-quality parallel data available for training, evaluation, and testing. This prompts the question: are some languages indeed harder to translate from or to?

1.1 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. In order to eliminate confounds associated with generation and other evaluation metrics that would differ from the one used in training and 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, 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 do not perform any translation, rather, one could think of our effort as estimating conditional probabilities with the Transformer. Furthermore, 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. Disparity in how the Transformer "sees"/classifies languages exists, if/when the differences are statistically significant.

1.2 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, 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 languages (when comparing them pairwise). 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 on 6 diverse languages such that there are no statistically significant differences between any of the 15 language pairs through lossless data transformation (or hyperparameter tuning) on the character and byte level, we show that **unless word-based methods are used, there is no complexity that is intrinsic to a language aside from its statistical properties concerning sequence length and vocabulary, irrespective of its linguistic typological, phylogenetic, historical, or geographical profile**. Language complexity is relative to and bounded by its representation level (representation relativity). The conventional expectation that languages should/must be different based on extra-statistical grounds stems from the concept of a "word" and from our tradition of word-based segmentation practice. We find explicitly debunking this expectation of disparity necessary because more diligent error analyses need to be afforded instead of simply accepting massively disparate results.
4. Bigger/overparametrized models can exacerbate the effect of data statistics. Biases that can be expressed quantitatively and lead to disparity are mitigable through hyperparameter tuning.

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 including fairness evaluation and discussion on non-monotonic behavior are in § 5, related work in § 6. We suggest some possibilities for future directions as we conclude in § 7. We refer our readers to the Appendices for more detailed descriptions/discussions and 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 while controlling 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 science and one of our higher goals is to bridge an understanding between language science 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.**

2.1 DEFINITIONS

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 ($l_{\text{src1}}, l_{\text{src2}}$) and 15 pairs of target languages ($l_{\text{trg1}}, l_{\text{trg2}}$) 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 holistic evaluation.

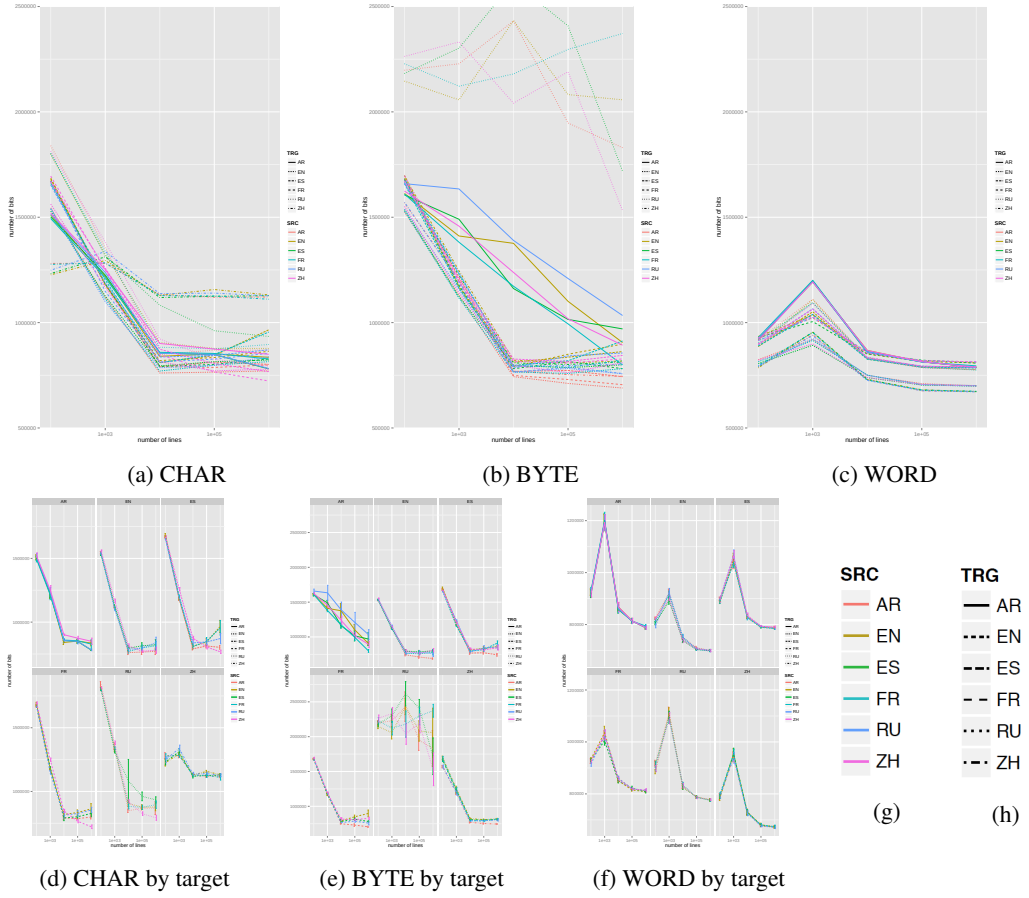
Experimental setup The systematic, identical treatment we give to our data is described as follows with further preprocessing and hyperparameter details in Appendices B and C, respectively. The distinctive point of our experiment is that the training regime is the same for all (rationale in App. N.1).

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

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

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 E. 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



(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 F.)

We should remind our readers that our goal is to investigate the presence of **relative differences** between the languages and not to directly compare absolute scores for what is "better" or "worse". There are many perspectives and set of findings possible, and our experiments here present one

perspective, one first attempt towards characterizing the behavior of the Transformer. What should be considered relevant results for our investigation is the number of language pairs with significant differences reported in Table 1, the general patterns of (non-)monotonicity and disparity in the figures, and the corresponding analyses.

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

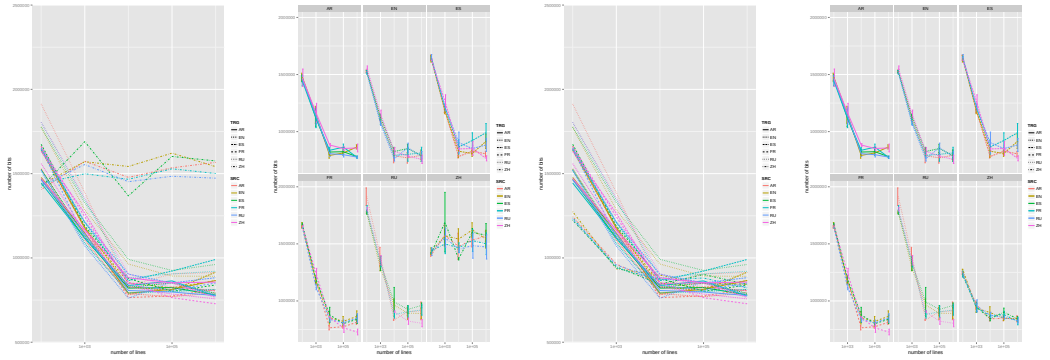
Subfigures 1d, 1e, and 1f display the corresponding information sorted into facets by target language, source languages represented as line types. Through these we see more clearly that results can be grouped rather neatly by target language (cf. figures sorted by source language in Appendix G) — as implicit in 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 for 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 and characters for AR and RU — 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)). And since the notion of word in ZH is highly contested and 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 processing in finer granularity is possible (cf. Li et al. (2019b) as well as Duanmu (2017) for a summary of 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

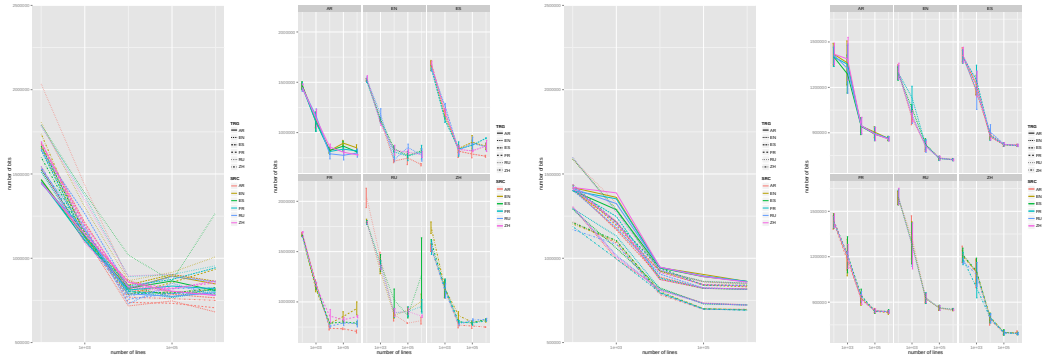
To understand why some languages show different results than others, we carried out a secondary set of control experiments with representations targeting the problematic statistical properties of the corresponding target languages. (An extended version of this section is provided in Appendix O.)

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 with 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 form a behaviorial tendency unlike those with other (phonetic) alphabetic scripts (Subfigure 2a). To the best of our knowledge, 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.

Byte level Length is the most salient statistical attribute that makes AR and RU outliers. To shorten their sequence length, 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 I 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 mitigating the word bias and 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 resembles the idea of morphological complexity. Considering there are character-internal (morphologically) meaningful units in languages with logographic script such as ZH (cf. Zhang & Komachi (2018)) that are rarely captured or studied, this goes to show that linguistic morphology, along with its complexity, as is practiced today 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 has been recognized as problematic for a very long time in language science (see Haspelmath (2011) and references therein from the past century). Since the notion of word, which has been centered on English and languages with alphabetic scripts, has negative impact on languages both morphologically rich (see Minkov et al. (2007), Seddah et al. (2010), inter alia) as well as morphologically "frugal" (Koehn, 2005), 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 "$\text{ARRU}_{s,t}$").

Double Descent (DD) 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 $\text{O}(10^{-4})$ for 10^2 lines, $\text{O}(10^{-3})$ at 10^3, $\text{O}(10^{-2})$ at 10^4, and $\text{O}(10^{-1})$ 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 $\text{O}(10^{-3})$ at 10^2 lines. Instead of thinking this spike of 10^3 as irregular, we may instead want to think of this learning curve as shifted by 1 order of magnitude to the right for characters and bytes or the performance at 10^2 lines for words and ZH-characters due to being overparametrized and hence abnormal. This would fit in with the findings by Belkin et al. (2019) and Nakkiran et al. (2020) attributing DD to overparametrization. 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. Details of our supplemental experiments on the datasets used by the Nakkiran et al. (2020) corroborating our

findings as well as results additional experiments on a non-neural LM are provided in Appendix J. Number of model parameters can be found in Appendix K.

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 L). 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 L 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 M). We observed no significant disparity across the board. It seems larger/overparametrized models can magnify and exacerbate the differences in the statistical properties in the data. That hyperparameter tuning, in this case, by changing the number of layers can mitigate effects from the data statistics, to the best of our knowledge, is 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 (source language neutralization and DD on word level).

6 Related work

Many related work have already been mentioned in our analyses in passing. One main point, however, that we find pertinent to emphasize is the (ir-)relevance of linguistic typology in multilingual NLP. Many recent work have advocated its relevance (Gerz et al., 2018b; Clark et al., 2020; Joshi et al., 2020). However, not many researchers are aware of how the modeling of many of these word-based symbolic concepts could bias NLP systems due to the modeling of a notion of "word" that is not crosslinguistically consistent and that we could be doing a disservice to many other languages. That basic data statistics being 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, representation granularity and quantitative and qualitative fairness. We make a finer distinction demarcating when linguistic typological concepts could be relevant (possibly on word level and when they are being explicitly modeled) and when they are not. 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. Other related work can be found in Appendix P.

7 Conclusion

Machine learning has enabled greater diversity in NLP (Joshi et al., 2020). Fairness, in the elimination of disparity, does not require big data. It will take everyone's effort to mitigate the bias in ourselves and to support fairness in modeling instead of sticking to a convenient standard that serves only languages that have been dominant in our academic tradition or engineering practice. Resources can be invested in e.g. building character encoding that complement languages' statistical profiles (considering the basis of many things multilingual goes back to the Multilingual Plane), finer-grained science for NLP (e.g. studying the relation between data transform and algorithm/hyperparameter adjustment, alignment of elements between logographic and (phonetic) alphabetic scripts), decomposition analyses for characters, better compression schemes, as well as the creation and curation of multitexts and stylistic/multimodal contrast sets that are non-artificial, because we need to understand the diversity of the natural statistical profiles/behaviors of the world's languages, in raw (not pretokenized) form, also for the purpose of evaluation and education in a statistical science for NLP beyond the explicit modeling of word-based concepts.

REFERENCES

Ahmed M. Alaa and Mihaela van der Schaar. Demystifying black-box models with symbolic metamodels. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 11304–11314. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/9308-demystifying-black-box-models-with-symbolic-metamodels.pdf.

Leonor Becerra-Bonache, M. Dolores Jiménez-López, Carlos Martín-Vide, and Adrià Torrens-Urrutia (eds.). *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, Santa Fe, New-Mexico, August 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-4600.

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.

Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, Thomas François, and Philippe Blache (eds.). *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://www.aclweb.org/anthology/W16-4100.

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.

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.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470, 2020. doi: 10.1162/tacl_a_00317. URL https://www.aclweb.org/anthology/2020.tacl-1.30.

Marta R. Costa-jussà, Carlos Escolano, and José A. R. Fonollosa. Byte-based neural machine translation. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pp. 154–158. Association for Computational Linguistics, 2017. doi: 10.18653/v1/W17-4123. URL http://aclweb.org/anthology/W17-4123.

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.

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.

Dan Gillick, Cliff Brunk, Oriol Vinyals, and Amarnag Subramanya. Multilingual language processing from bytes. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1296–1306. Association for Computational Linguistics, 2016. doi: 10.18653/v1/N16-1155. URL http://aclweb.org/anthology/N16-1155.

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.

Bernd Kortmann and Verena Schröter. Linguistic complexity. In M. Aronoff (ed.), *Oxford Bibliographies in Linguistics*. Oxford University Press, New York, 2020. doi: 10.1093/OBO/9780199772810-0254.

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.

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.