# Assessing the Impact of Typological Features on Multilingual Machine Translation in the Age of Large Language Models

**Vitalii Hirak[1], Jaap Jumelet[2], Arianna Bisazza[2]**
[1]Data & Knowledge Engineering, Heinrich Heine University
[2]Center for Language and Cognition (CLCG), University of Groningen
vitalii.hirak@hhu.de {j.w.d.jumelet, a.bisazza}@rug.nl

## Abstract

Despite major advances in multilingual modeling, large quality disparities persist across languages. Besides the obvious impact of uneven training resources, typological properties have also been proposed to determine the intrinsic difficulty of modeling a language. The existing evidence, however, is mostly based on small monolingual language models or bilingual translation models trained from scratch. We expand on this line of work by analyzing two large pre-trained multilingual translation models, NLLB-200 and Tower+, which are state-of-the-art representatives of encoder-decoder and decoder-only machine translation, respectively. Based on a broad set of languages, we find that target language typology drives translation quality of both models, even after controlling for more trivial factors, such as data resourcedness and writing script. Additionally, languages with certain typological properties benefit more from a wider search of the output space, suggesting that such languages could profit from alternative decoding strategies beyond the standard left-to-right beam search. To facilitate further research in this area, we release a set of fine-grained typological properties for 212 languages of the FLORES+ MT evaluation benchmark.

## 1 Introduction

Despite major advances in multilingual modeling, the quality of language technologies still varies widely across languages (Joshi et al., 2020; Blasi et al., 2022; Sarti et al., 2022). These inequalities are largely due to the uneven availability of training data. However, some languages also appear to be intrinsically more difficult to model than others by modern approaches, a variability that has been connected to typological properties by a rich line of work (Birch et al., 2008; Cotterell et al., 2018; Mielke et al., 2019; Bugliarello et al., 2020; Bisazza et al., 2021; Park et al., 2021; Arnett and Bergen, 2025).

To isolate intrinsic modeling difficulty from other factors, those studies strongly prioritize the comparability of training corpora. While principled, this choice constrains evaluations to a very small set of existing multi-parallel datasets, typically, Europarl (Koehn, 2005) or the Bible (Mayer and Cysouw, 2014), which are limited in typological diversity or in size and domain. Moreover, models trained from scratch on such corpora are poor representatives of current practices in MT and language modeling in general, leaving the open question: can typological properties explain state-of-the-art MT quality in the age of LLMs?

To provide an answer, we expand on previous work by evaluating two large pre-trained multilingual models in a wide, typologically diverse set of languages, while using approximations of language resourcedness to control for data size effects. Additionally, we explore how widening the search for a high-probability sequence during inference affects translation quality in languages of different typology, calling into question the optimality of using a single decoding strategy across many different languages. With a focus on word order and morphological complexity, we consider a broad set of fine-grained features, following recent trends in typology where continuous (or gradient) features are increasingly preferred over coarse-grained categorical ones (Levshina et al., 2023; Baylor et al., 2024). Our work makes the following contributions:

- Identifying specific typological features that predict translation quality of two widely used multilingual MT models — the encoder-decoder NLLB-200 model (Costa-jussà et al., 2022) and the decoder-only Tower+ model (Rei et al., 2025) — across a total of 7 source and 124 target languages.
- Analyzing the interplay between optimal beam size and typological properties of the generated language.



*Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*
*Volume 1: Long Papers*, pages 2416–2434

- Compiling a dataset of fine-grained, continuous typological features for the 212 target languages in the FLORES+ benchmark (Costa-jussà et al., 2022) to facilitate further research on language-specific decoding strategies.[1]

Leveraging a typologically diverse selection of languages, we show that target language typology drives translation quality of the NLLB-200 model, even after accounting for more trivial factors, such as data resourcedness and writing script. Additionally, we uncover a large variability in optimal beam size across target languages, and find that widening beam search yields significantly higher-probability outputs for languages of certain typologies. Important disparities in translation performance are also observed in the LLM-based Tower+ model, although across fewer significant factors, calling for further experimentation with this type of models.

## 2 Background

### 2.1 Typological Language Properties

Languages differ from one another in various aspects such as phonology, morphology, and syntax. One approach to measure this variation is to assign *discrete* categorical values to languages, as exemplified by the World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013) and GramBank (Skirgård et al., 2023). Recent literature argues, however, that using *continuous* languages properties more accurately reflects the variability of natural languages and is more appropriate in the realm of NLP (Levshina et al., 2023; Baylor et al., 2024). Continuous approaches towards measuring the morphological complexity and word order flexibility range from simple ratios (Chotlos, 1944; Xanthos et al., 2011) to applying information-theoretic principles (Juola, 1998; Bentz and Alikaniotis, 2016) to using the accuracy of machine learning (ML) models on the task of predicting inflected word forms (Cotterell et al., 2019). In line with these findings, we use a combination of continuous language properties computed in prior work as well as properties calculated ourselves on the basis of the FLORES+ dataset.

### 2.2 Typology and Modeling Difficulty

A number of studies have measured and tried to explain the intrinsic difficulty of modeling different languages using controlled setups, that is with fixed size and, where possible, comparable content of training data. Cotterell et al. (2018) and Mielke et al. (2019) study monolingual LSTM models trained on the Europarl corpus. They conclude that only general statistical properties, like raw character sequence length and the tokenization-specific word inventory size, correlate significantly with LM surprisal values. Their considered WALS features and other language properties do not. Park et al. (2021) revisit the question including more languages (from the Bible corpus) and more morphological features from WALS. They find various morphological features to correlate significantly with LM difficulty for a BPE-based model, but not for a character-level one. Arnett and Bergen (2025) analyse 22 monolingual Transformer models from the Goldfish suite (Chang et al., 2024) and find mismatches in data size calculations to better predict perplexity differences compared to morphological complexity and tokenizer quality.

In the translation domain, target language morphological complexity and language relatedness were found to be significant predictors of translation performance already in the era of pre-neural statistical MT (Birch et al., 2008). More recently, Bugliarello et al. (2020) train Transformer models on Europarl language pairs and find only type/token ratio (TTR) to correlate significantly with translation difficulty. Wan (2022) denies the role of morphological complexity in driving modeling difficulty and instead attributes disparities to mismatches in representational granularity among languages (e.g. longer encodings for some writing scripts). Bisazza et al. (2021) use synthetic versions of English to investigate whether languages with flexible word order and case marking are more difficult to translate by NMT models. They find that, in low-resource settings, such languages are harder to learn than their fixed-order, no-marking counterparts. Concurrently to this work, Ploeger et al. (2025) construct a multi-parallel dataset and use it to train bilingual NMT models and analyze translation difficulty in the scope of 8 European languages. They find that NMT difficulty varies across language pairs and can be predicted based on their genetic and syntactic similarity. While being highly controlled, their experimental setup is not necessarily representative of state-of-the-art NMT and is very limited in typological diversity.

Another set of studies attempts to uncover the inductive biases of language models by training (and evaluating) them on artificial languages of

[1] github.com/v-hirak/explaining-MT-difficulty.

different typologies (White and Cotterell, 2021; Kallini et al., 2024; Kuribayashi et al., 2024; El-Naggar et al., 2025; Yang et al., 2025).

We depart from these lines of work by studying considerably larger, pre-trained multilingual models representative of state-of-the-art MT, at the cost of control over the training data. A similarly pragmatic approach is taken by Arnett and Bergen (2025) in their analysis of multilingual models in various tasks (but not MT). Their hypotheses were later reassessed in Poelman et al. (2025), who identified confounding factors in the experimental setup that should be considered, such as languages considered, tokenization algorithms, training data characteristics, and performance indicators. Also related to our work, Sarti et al. (2022) assess the usefulness of Google Translate and mBART-50 translations by a human post-editing study. Their chosen target languages are typologically diverse, but are too few (six) to draw any reliable conclusions on the effect of typological properties.

### 2.3 Decoding Algorithms

Decoding or generation algorithms serve to search for a high-probability sequence through an intractably large output space. One of the most widely used algorithms is *beam search* (Wu et al., 2016; Brown et al., 2020; Costa-Jussà et al., 2022; Raffel et al., 2023), a simple deterministic approximation to maximum a-posteriori decoding, where a 'beam' of the $k$ most probable partial hypotheses is kept at each time step. Despite its simplicity, beam search with a small $k$ value (e.g. 3 to 5) typically yields considerably higher-probability outputs and better task performance than greedy decoding (a special case of beam search with $k = 1$) at the cost of slower inference, across tasks and settings (Junczys-Dowmunt et al., 2016; Freitag and Al-Onaizan, 2017; Kulikov et al., 2019; Park et al., 2020). Output quality has been found to deteriorate beyond a certain beam size (Koehn and Knowles, 2017; Cohen and Beck, 2019). While all these works consider very few, mostly high-resource languages, our evaluation covers a much broader and diverse set of language pairs.

Beyond beam search, many other decoding algorithms exist, including deterministic and stochastic ones (see Welleck et al. (2024) for an extensive survey). Given common practice in the MT community as well as recent evidence on the optimality of beam search for translation in state-of-the-art LLMs (Shi et al., 2024), we focus here on varying the width of beam search and leave the exploration of alternative algorithms to future work.

To our knowledge, no work has studied the interplay between optimal decoding strategy and the typological properties of the generated language.

## 3 Experimental Setup

This section describes our experimental setup, including the choice of models and datasets, languages, and evaluation metrics.

**Translation Models** We opt for two multilingual pretrained models: NLLB-200 (Costa-jussà et al., 2022) and Tower+ (Rei et al., 2025). NLLB-200 3.3B[2] is an encoder-decoder Transformer NMT model with 3.3 billion parameters, capable of translating among 202 languages. Tower+ 9B[3] is a decoder-only multilingual LLM post-trained for MT, and it was shown to often outperform larger general-purpose open-weight and proprietary models such as Llama 3.3 70B and GPT-4o in translation tasks (Rei et al., 2025). While Tower+ was explicitly post-trained only on 22 languages, its underlying LLM model, Gemma2 9B (Riviere et al., 2024), exhibits multilingual capabilities that extend to more languages. Both NLLB and Tower models use a single subword vocabulary shared across languages, as is standard practice in modern multilingual models.

**Evaluation Dataset** To define the set of languages studied and for translation material, we use the wide-coverage multi-parallel FLORES+ machine translation evaluation benchmark dataset.[4] FLORES+ is composed of English sentences equally sampled from Wikinews, Wikijunior, and Wikivoyage and manually translated into over 200 languages. For our experiments, we use the dev split of the dataset comprising 997 sentences.

**Translation Directions** Since we are primarily interested in the effect of *target* languages on translation difficulty and decoding requirements, we select 124 typologically diverse target languages from FLORES+, taking into account NLLB-200 language coverage, the availability of language properties, and computational resources at hand.[5] Our first set of experiments involves translating

[2] hf.co/facebook/nllb-200-3.3B.

[3] hf.co/Unbabel/Tower-Plus-9B. For the prompt template used to generate translations, see App. C

[4] hf.co/datasets/openlanguagedata/flores_plus.

[5] For the list of target languages, see Appendix A.

from English into a large variety of target languages. To ensure our results are not dependent on this specific source language choice, we then experiment with six additional source languages (Arabic, Italian, Dutch, Turkish, Ukrainian, and Vietnamese), which were selected in previous work (Sarti et al., 2022) to ensure typological diversity and comparable data resourcedness.

**Beam Size** In light of previous findings on the ineffectiveness of very wide beam search (Koehn and Knowles, 2017; Cohen and Beck, 2019), we limit our selection of beam sizes used in generating translations to $k \in \{1, 3, 5, 7\}$.

**Evaluation Metrics** As our translation quality metric, we use chrF++ (Popović, 2015),[6] which is based on character-level n-gram overlap between the hypothesis and reference translations. Compared to word-level metrics such as BLEU (Papineni et al., 2002), chrF++ is better suited for languages with rich morphology. However, being based on surface-level matching, it can still fail to capture semantic similarities between MT output and reference. Modern translation quality metrics addressing this issue, such as COMET (Rei et al., 2020) and MetricX (Juraska et al., 2023), correlate better with human judgments. However, their reliance on pre-trained neural encoders or LLMs makes them unreliable for low-resource languages (Falcão et al., 2024; Singh et al., 2024; Wang et al., 2024; Sindhujan et al., 2025). Given our strong focus on cross-lingual comparability, we thus opt for a simpler, model-free metric. Note that chrF++ is still used routinely as an additional metric in large-scale multilingual MT evaluations (Alves et al., 2024; Rei et al., 2025) and for low-resource pairs (Kocmi et al., 2025). Appendix D reports overall scores of the NLLB model with alternative metrics (BLEU and COMET), showing similar trends by beam size on average.

## 4 Language Properties

Here we outline our choice of language properties used to estimate the impact of a language typology on NMT (see App. B for a more detailed description of each property). Besides language resourcedness and coarse-grained source-target distances, we focus on features of morphological complexity and word order of the language being generated.

[6]We use the sacreBLEU implementation (Post, 2018).

### 4.1 Language Resourcedness

The proportion of a language in the model's training data is clearly an important factor for translation quality. Unfortunately, this information is often unavailable for large pre-trained models. Even in the case of open-source models like NLLB-200, calculating a language proportion is complicated by the size of the dataset and absent or imprecise meta-data.[7]

In light of this, we approximate the *general resourcedness of a language* using language size data from the GlotCC broad-coverage CommonCrawl corpus (Kargaran et al., 2025). Specifically, we collect content length values for 210 out of 212 languages in FLORES+. While this is a very rough approximation of what our evaluated models were exposed to, we assume that large disparities within CommonCrawl will correlate overall with large disparities of language presence on the Web, which provides the large bulk of training data for modern translation systems.

### 4.2 Source-Target Distances

Following previous work on NMT (Sarti et al., 2022) and cross-lingual transfer (Lin et al., 2019), we adopt the URIEL typological database (Littell et al., 2017) and query six types of typological distances pre-computed by aggregating broad categories of typological features: **genetic**, **geographic**, **syntactic**, **inventory**, **phonological**, and **overall features** distance.[8] Additionally, we experiment with a simple **same_script** binary feature, capturing the advantage of language pairs having the same writing script (and, potentially, shared subwords).

### 4.3 Morphological Complexity Measures

To estimate the morphological complexity of a language, we make use of eight continuous measures that were precalculated and made available by Çöltekin and Rama (2023). These measures were computed on the basis of Universal Dependencies and are available for 33 languages of the FLORES+ dataset. **Information in Word Structure** (`WS`) compares the information content (i.e. entropy) of the original text with its compressed version (Juola,

[7]We initially attempted to reconstruct NLLB-200 training data size proportions by language, but due to important details missing (such as the amount of back-translated bitext), our estimates were ultimately inaccurate and unreliable.

[8]Concurrently to our work, Goot et al. (2025) released a toolkit providing a larger variety of language distances, which could be used to extend our analysis in future work.

1998). **Word and Lemma Entropy** (WH, LH): word entropy is based on word frequency distribution of a text (Bentz et al., 2016); Çöltekin and Rama (2023) additionally calculate the entropy of lemmas. **Mean Size of Paradigm** (MSP) is calculated by dividing the number of word forms in a text by the number of lemmas (Xanthos et al., 2011). **Inflectional Synthesis** (IS) is the maximum number of inflection categories that can be expressed by a standalone verb. **Morphological Feature Entropy** (MFH) reflects the usage of morphological features (e.g. grammatical cases) and their values. **Inflection Accuracy** (IA) is the accuracy of an ML model on the task of predicting inflected forms given lemma and grammatical features.[9] Finally, we express the **Type/Token Ratio** — the number of unique words divided by total words — in three ways: **TTR** (Chotlos, 1944) (on both UD and FLORES+), *Root Type/Token Ratio* (**RTTR**) (Guiraud, 1959), and *Moving Average Type/Token Ratio* (**MATTR**), which is TTR calculated inside a sliding window (Covington and McFall, 2010) (both computed on FLORES+).

### 4.4 Gradient Word Order Measures

We include a number of gradient word order measures proposed and computed by Levshina (2019) and Levshina et al. (2023). **Average word order entropy of dependents and codependents** (h_dep/h_codep) is the entropy of different word order patterns of *dependencies* (e.g. verb-subject and noun-adposition relations) and *codependencies* (e.g. subject and object of the same verb). **Proportion of Subject-Object order** (SO_prop) is based on the frequencies of clauses where subject comes before object.[10] **Percentage of head-final phrases** (head_finality) approximates the preference of a language towards head-initial or head-final phrases.

## 5 Results

Having collected the set of language properties, we analyze the cross-lingual variability of translation quality and its improvement as a function of beam size, across a wide set of languages. We begin with a correlation study (§5.1.1) to get a general idea of which typological properties of target languages correlate significantly with chrF++ scores of the NLLB-200 model. We then use these insights to narrow down our selection of language features and use them in a more focused regression setup (§5.1.2, §5.1.3), where we evaluate their effect on predicting translation difficulty and its change at a higher beam size. Finally, we extend the regression experiments to the Tower+ model to determine whether our findings generalize to decoder-only LLMs used nowadays for MT (§5.2).

### 5.1 Encoder-Decoder NLLB-200

Due to computational constraints, we conduct the full set of analyses on the lighter-weight NLLB-200 model, which also provides official support for a larger set of languages than Tower+.

#### 5.1.1 Translation Quality

As a first exploration of language properties affecting translation performance, we measure the correlation between all our language properties (§4) and chrF++ scores for translations from English at beam size $k = 5$, commonly used in practice. We report Spearman's rank correlation as property values are not normally distributed. Due to the varying availability of linguistic features, finding a subset of languages where all the features are specified would result in a very limited sample. We therefore compute correlations on subsets of varying sizes to leverage as many data points as possible.

As shown in Fig. 1, various language properties correlate significantly with translation quality scores. The strong positive correlation with *GlotCC size* confirms that these estimates can be used to approximate language proportions in our model's training data. The source-target typological distances also behave as expected: *genetic*, *geographic*, and *syntactic* distances inversely correlate with chrF++, confirming that more disparate language pairs are harder to translate. Unsurprisingly, no correlation is observed for the phonological distances (*phonological* and *inventory*). As for target morphological complexity, 2 out of 11 continuous measures correlate significantly with translation quality, namely *word and lemma entropy* (WH, LH), which reflect the average information content of words. This implies that target languages packing more information into word structure (rather than sentence structure) tend to have lower translation quality in NLLB. Finally, two target word order measures have some of the strongest correlations with translation quality: languages with less pre-

[9] For consistency with other metrics, we report results with *negative* inflection accuracy -IA, such that higher -IA means *lower* accuracy, reflecting higher complexity.

[10] For consistency with other metrics, we report results with *negative* SO order proportion -SO_prop, such that higher -SO_prop values reflect *lower* predictability of word order.

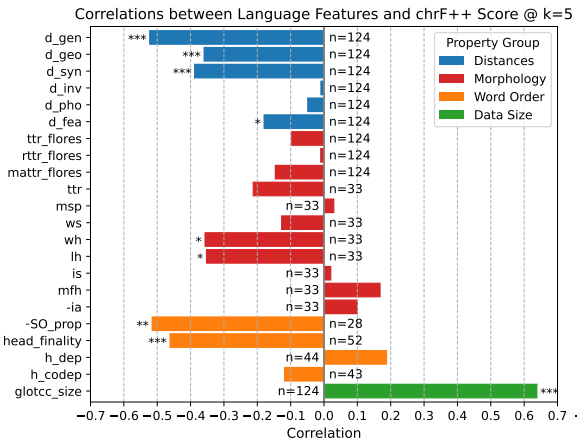


Figure 1: Spearman correlations between continuous language properties and NLLB-200 chrF++ translation quality scores (a character n-gram based translation quality metric, cf.§3) at beam size $k = 5$. Source language is English. Sample sizes (i.e. number of target languages) for each property are indicated next to their respective bars. Correlations significant at $p < 0.05$ are marked with *, at $p < 0.01$ with **, at $p < 0.001$ with ***.

*Source language*: **All** ($n = 364$)

| **Feature** | $\beta$ | $F$ | $p$**-value** | ***LR*** | **Adj.** $R^2$ |
|---|---|---|---|---|---|
| `src_language` | – | 9.02 | <0.001 | – | 0.07 |
| *Arabic* | **-3.65** | | | | |
| *Dutch* | **-10.3** | | | | |
| *Italian* | **-8.48** | | | | |
| *Turkish* | **-7.18** | | | | |
| *Ukrainian* | **-5.16** | | | | |
| *Vietnamese* | **-8.58** | | | | |
| `tgt:glotcc_size` | **3.34** | 56.5 | <0.001 | 59.1* | 0.20 |
| `is_same_script` | **5.67** | 26.8 | <0.001 | 63.6* | 0.33 |
| `d_gen` | **-1.24** | 6.2 | 0.01 | 8.75* | 0.34 |
| `tgt:mattr` | -0.67 | 2.02 | 0.16 | 7.49* | 0.36 |
| `tgt:head_fin` | **-2.58** | 27.2 | <0.001 | 27.1* | 0.40 |

Table 1: Regression results for predicting chrF++ scores at $k = 5$ for NLLB-200 translations. $\beta$ indicates the $\beta$-coefficients, $F$ indicates the $F$ scores of the ANOVA type II test, *LR* — the log-likelihood ratio of incrementally adding each variable (from top to bottom, significant ratios marked by *), and the adjusted $R^2$ — the explained variance of the incremental linear model (top to bottom). Coefficient values ($\boldsymbol{\beta}$) marked in bold are significant ($p < 0.05$).

dictable word orders (higher `-SO_prop`) and languages with a stronger preference for head-final phrases (higher `head_finality`) are more challenging to translate into.

### 5.1.2 Predicting chrF++ Scores

Based on the correlation results, we narrow down our selection of features and languages to allow for a more controlled linear regression setup. Specifically, we predict chrF++ scores at $k = 5$ including obvious factors such as the source language id,[11] the log-transformed GlotCC estimate of the target language's presence in training data,[12] source-target language genetic distance, and whether the language pair shares the same script.[13] On top of that, we incorporate *moving average type/token ratio* (MATTR) as a measure of target morphological complexity and target *head-finality* as a word order measure.[14] This feature set ensures we have data for 52 target languages, and we additionally expand our source language selection to seven languages, resulting in a total of $7 \times 52 = 364$ data points.

Table 1 summarizes the regression results. Overall, the significance and directionality of the effects on chrF++ follow our intuitions, barring target language MATTR. Namely, larger source-target genetic distance, prevalence of head-final phrases in the target language, and translating from source languages other than English all result in *lower* translation quality. Importantly, these trends are observed after controlling for data size effects and script similarity.[15]

### 5.1.3 Beam Width

We now shift our attention to the changes in translation quality as a function of beam size. First, we plot beam size against the relative increase in chrF++ for the NLLB-200 model translating from English into 124 target languages (Fig. 2). We find a large variation across languages, with roughly half of the languages benefiting from a wider beam size ($k = 7$), suggesting that a language-specific choice of decoding strategy could be beneficial in terms of quality and efficiency. Next, we investigate whether this variability can be partly explained by target typological properties. To this end, we define two measures of *gain by beam size*:

- **chrF++ gain**: for each source-target lan-

[11] The limited number of source languages (7) prevent us from properly modeling source language properties. Empirically, using src_id or source features leads to a similar fit.

[12] Because the values are positive and span orders of magnitude, we log-transform them to reduce skewness and heteroscedasticity prior to the regression analysis.

[13] As suggested later by an anonymous reviewer, we additionally conducted regression experiments to assess the effect of *token fertility* (average number of tokens per word) on chrF++ scores of both models. Ultimately, the effect emerged as non-significant after all other factors were accounted for.

[14] We also added interactions with source language MATTR and head-finality, but results were insignificant.

[15] We provide a closer look at the effects of head-finality and MATTR for individual language pairs in App. E.

| Feature | *Metric*: $\Delta chrF_{1;7}$ ($n = 364$) $\beta$ | *F* | *p*-value | *LR* | Adj. $R^2$ | *Metric*: $\Delta prob_{1;7}$ ($n = 364$) $\beta$ | *F* | *p*-value | *LR* | Adj. $R^2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| src_language | **-0.3** | 4.40 | <0.001 | – | 0.05 | **0.18** | 3.6 | 0.002 | – | 0.01 |
| tgt:glotcc_size | **0.15** | 17.2 | <0.001 | 14.2* | 0.09 | **-0.21** | 180.8 | <0.001 | 141.8* | 0.33 |
| is_same_script | -0.02 | 0.06 | 0.81 | 1.72 | 0.09 | **-0.08** | 4.5 | 0.03 | 37.2* | 0.39 |
| d_gen | 0.05 | 1.48 | 0.22 | 2.00 | 0.09 | **0.09** | 26.9 | <0.001 | 29.7* | 0.44 |
| tgt:mattr | 0.02 | 0.35 | 0.56 | 1.04 | 0.09 | **0.05** | 10.6 | 0.001 | 23.1* | 0.47 |
| tgt:head_fin | 0.06 | 2.38 | 0.12 | 2.45 | 0.09 | **0.12** | 46.3 | <0.001 | 45.0* | 0.53 |

Table 2: Regression results for predicting chrF++ gain (*left*) and probability gain (*right*) for NLLB-200 translations. The model is fitted on the results of all source languages. Coefficient values ($\beta$) marked in bold are significant ($p < 0.05$). Here $\beta$ for source language denotes a mean effect of source language with English as a reference value (see App. F.1 for effect breakdown by source language).

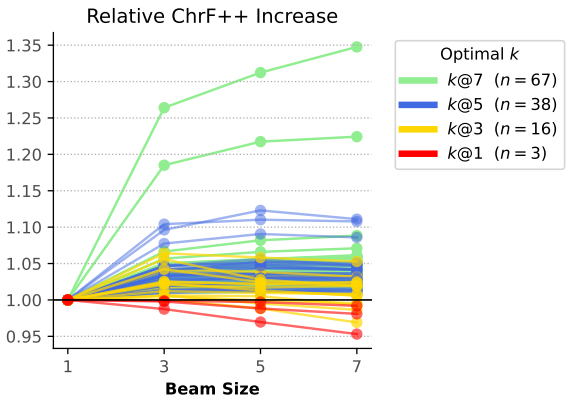


Figure 2: Relative increase in chrF++ for NLLB-200, translating from English to 124 different target languages. Curves are colored by their optimal beam size.

guage pair, we calculate the gain in chrF++ as $\Delta chrF_{1;7} = chrF_7 - chrF_1$, where $chrF_i$ denotes the chrF++ score at beam size $i$.

- **Probability gain**: we also extract the sentence generation probabilities by the model and average them over test sentences. Similar to chrF++, we calculate the probability gain as $\Delta prob_{1;7} = prob_7 - prob_1$, where $prob_i$ denotes average generation probability at beam size $i$. This allows us to estimate the optimality of a decoding strategy from the point of view of the model, independently from a specific quality metric. Since probability gains are strictly positive and vary over orders of magnitude, we apply log-transformation to these values prior to the regression analysis to reduce skewness and heteroscedasticity.

Results in Table 2 show a low model's fit (adjusted $R^2$) across the board when predicting chrF++ gain. Only source language id and target language resourcedness bear a significant effect on chrF++ gain. The model predicting probability gain paints a different picture: all explanatory variables have clear and significant predictive power. The coefficients for source languages are positive, meaning that translating from non-English languages yields a higher probability gain compared to English. Since English is arguably the most high-resourced among the languages studied, this could imply that translating not only *into*, but also *from* lower-resource languages (compared to English) may benefit from a wider beam width. All three target typological properties (genetic distance, MATTR, and head-finality) carry significant positive effects on probability gain, suggesting that narrow beam search may be a worse approximation of optimal model sequence for distant pairs, and that morphologically complex, head-final target languages may benefit from different decoding strategies.

### 5.2 Decoder-only Tower+

While NLLB-200 is an encoder-decoder Transformer designed specifically for NMT, the current state-of-the-art has shifted towards prompting decoder-only LLMs for translation. To account for this, we extend our regression experiments to the Tower+ 9B model, using the same 364 language pairs.

First, we compare chrF++ scores at beam size 7 for translations generated by NLLB-200 and Tower+ (Fig. 3). Tower+ scores higher on the vast majority of target languages it was explicitly trained on (orange region), but underperforms NLLB-200 on *all* target languages outside of Tower's coverage (blue region), reaffirming the relevance of the NLLB model for large-scale MT evaluations including low-resourced languages.

Regression results for predicting $chrF_7$ and probability gain ($\Delta prob_{1;7}$) are shown in Table 3. We use the same predictors as for NLLB-200, but add a simple binary feature `is_in_tower` denoting

| | *Metric*: $chrF_7$ ($n = 364$) | | | | | *Metric*: $\Delta prob_{1;7}$ ($n = 364$) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Feature** | $\beta$ | ***F*** | ***p*-value** | ***LR*** | **Adj. $R^2$** | $\beta$ | ***F*** | ***p*-value** | ***LR*** | **Adj. $R^2$** |
| src_language | **-4.9** | 6.67 | <0.001 | - | 0.0 | 0.15 | 1.58 | 0.15 | - | 0.01 |
| is_in_tower | **15.2** | 142 | <0.001 | 218* | 0.45 | **-1.33** | 298 | <0.001 | 302* | 0.57 |
| tgt:glotcc_size | **6.28** | 102 | <0.001 | 73.2* | 0.55 | -0.05 | 1.53 | 0.22 | 3.12 | 0.57 |
| is_same_script | **10.4** | 67.1 | <0.001 | 113* | 0.67 | -0.1 | 1.75 | 0.19 | 7.99* | 0.58 |
| d_gen | -0.04 | 0.004 | 0.95 | 0.5 | 0.67 | **0.16** | 20.2 | <0.001 | 21.7* | 0.6 |
| tgt:mattr | **-1.97** | 12.5 | <0.001 | 24.7* | 0.69 | -0.04 | 1.4 | 0.24 | 0.45 | 0.6 |
| tgt:head_fin | **-3.14** | 29.6 | <0.001 | 29.5* | 0.71 | **0.07** | 4.11 | 0.04 | 4.24* | 0.6 |

Table 3: Regression results for predicting chrF++ scores at $k = 7$ (*left*) and probability gain (*right*) for translations by Tower+ (**all source languages**). Coefficient values ($\beta$) marked in bold are significant ($p < 0.05$).

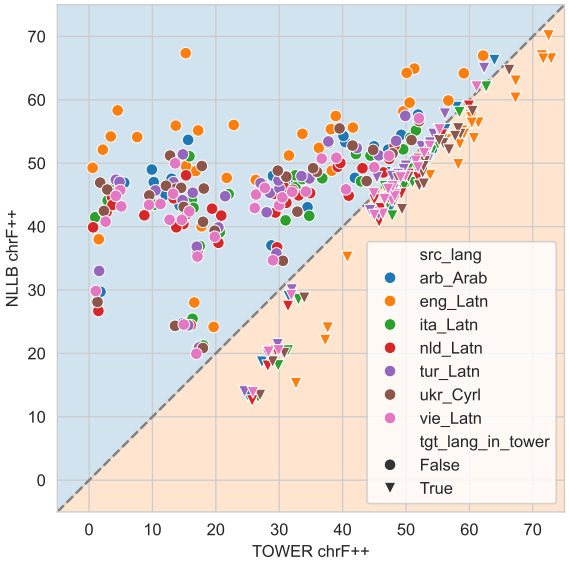


Figure 3: Tower+ 9B chrF++ scores vs. NLLB-200 3.3B chrF++ scores at beam size $k = 7$. Each point denotes a language pair and is colored by source language, while ▼ denotes target languages officially supported by Tower+. The blue and orange shaded regions indicate language pairs for which either NLLB-200 or Tower+ scores are higher, respectively. Sample size is $n = 7 \times 52 = 364$.

whether a target language is officially covered by Tower+ or not as per Rei et al. (2025). When predicting chrF++ scores (Table 3, left), almost all variables have a significant effect, barring source-target genetic distance. The standardized regression coefficients ($\beta$) suggest that disparities in target language resourcedness and model coverage dominate quality variability in the translations produced by Tower+, with the intrinsic typological features of morphology and word order playing a smaller but still significant role. The findings for $\Delta prob_{1;7}$ (Table 3, right) are less conclusive, with fewer significant features than for NLLB-200. The directionality of the Tower+ coverage factor ($-1.33$) indicates that non-supported target languages see a larger gain in modeling probability with a larger beam size. Source language id, target language resourcedness, and source-target script matching do not emerge as significant. Instead, genetic distance and head-finality of the target language have significant predictive power on probability gain and are consistent in their direction and magnitude with the NLLB-200 results.

## 6 Discussion and Conclusion

We set out to assess and explain the crosslingual variability of modern MT quality across a broad set of languages, with a focus on word order properties and morphological complexity of the generated (i.e. target) language. Leveraging the wide coverage FLORES+ MT evaluation dataset and two widely used large pre-trained multilingual MT models (NLLB-200 and Tower+), we generated translations for a variety of language pairs and beam sizes, and evaluated the models' performance and gain thereof via chrF++ and generation probabilities.

Through a set of correlation and regression experiments, we found several language properties to significantly predict quality of the encoder-decoder model (NLLB-200), even after controlling for language resourcedness and script similarity. These properties include source-target typological distances, as well as type/token ratio and head-finality of the target language.

Focusing on inference-time decoding strategy, we uncovered a large variability in chrF++ gains when widening the beam size, calling into question the common practice of using a unique decoding approach for many different language pairs and highlighting the importance of further research in language-specific decoding optimization. We failed to establish a predictive link between our selected language properties and chrF++ gains by

beam size. However, from the point of view of searching for the model's highest-probability target sequence, we did find that languages with more complex morphology and flexible word order benefit more from widening the beam size. In other words, the standard practice of decoding with a narrow beam search may be particularly suboptimal for these languages.

Finally, we replicated the regression experiments for the state-of-the-art Tower+ multilingual LLM. When predicting chrF++ score, factors like official language support by the model and resourcedness of the target language seem to dominate the variability of translation quality, but our regression analysis also revealed a strongly significant effect of target morphological complexity and word order. The findings for performance gains were somewhat less conclusive, as fewer language properties seem to have predictive effects on probability delta by beam size. Still, the features that are significant (namely, Tower+ target language coverage, genetic distance, and target head-finality) are consistent in their effects with the encoder-decoder NLLB-200 and demonstrate the significant predictive power of typological properties even when other factors like resourcedness and script emerge as non-significant.

Promising directions for future work include methods to dynamically set the beam width based on the target language indicated in the user prompt, as well as the evaluation of different strategies beyond beam search which could better account for word order flexibility and other properties of the target language. We hope that our work will inspire further research on measuring and explaining the intrinsic difficulty of translating and generating different languages in the age of LLMs.

## Limitations

While our findings offer insights into how typological features interact with decoding parameters, the scope of our analysis is subject to several constraints.

Firstly, we focus on only two translation models, which enables a controlled and large-scale study, but limits the generalizability of our findings to other models and training paradigms. Evaluating other (LLM-based) models covering different sets of languages may result in more significant factors in the regression analysis.

Relying on a surface-level evaluation metric (chrF++) remains problematic. Followup experiments could adopt newer, better metrics as more research is conducted on low-resource languages.

Furthermore, since training language proportions of our evaluated models are not publicly available, we had to estimate language resourcedness through a very rough approximation based on the CommonCrawl dataset.

We also examine only one decoding paradigm — left-to-right beam search — with four beam sizes. Alternative decoding strategies, such as sampling-based or non-autoregressive methods, may interact differently with typological properties, warranting further investigation.

Finally, our set of typological features is bounded by the availability of data for the languages in our study; for many languages, some fine-grained features remain missing. Follow-up experiments could at least use more recent versions of the typological distances (Khan et al., 2025; Goot et al., 2025).

## 7 Acknowledgements

We are thankful to the anonymous reviewers for their helpful comments. Vitalii Hirak received funding from the Erasmus Mundus Masters Programme in Language and Communication Technologies, EU grant no. 2019-1508. Jaap Jumelet and Arianna Bisazza were supported by the Talent Programme of the Dutch Research Council (grant VI.Vidi.221C.009).

## References

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks.

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

Emi Baylor, Esther Ploeger, and Johannes Bjerva. 2024. Multilingual gradient word-order typology from universal dependencies.

Christian Bentz and Dimitrios Alikaniotis. 2016. The word entropy of natural languages.

Christian Bentz, Tatyana Ruzsics, Alexander Koplenig, and Tanja Samardžić. 2016. A comparison between

morphological complexity measures: Typological data vs. language corpora. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 142–153, Osaka, Japan. The COLING 2016 Organizing Committee.

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2008. Predicting success in machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 745–754, Honolulu, Hawaii. Association for Computational Linguistics.

Arianna Bisazza, Ahmet Üstün, and Stephan Sportel. 2021. On the difficulty of translating free-order case-marking languages. *Transactions of the Association for Computational Linguistics*, 9:1233–1248.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Emanuele Bugliarello, Sabrina J. Mielke, Antonios Anastasopoulos, Ryan Cotterell, and Naoaki Okazaki. 2020. 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*, pages 1640–1649, Online. Association for Computational Linguistics.

Tyler A Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K Bergen. 2024. Goldfish: Monolingual language models for 350 languages. *arXiv preprint arXiv:2408.10441*.

John W Chotlos. 1944. Iv. a statistical and comparative analysis of individual written language samples. *Psychological Monographs*, 56(2):75.

Eldan Cohen and Christopher Beck. 2019. Empirical analysis of beam search performance degradation in neural sequence models. In *International Conference on Machine Learning*, pages 1290–1299. PMLR.

Çağrı Çöltekin and Taraka Rama. 2023. What do complexity measures measure? correlating and validating corpus-based measures of morphological complexity. *Linguistics Vanguard*, 9(s1):27–43.

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

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Ryan Cotterell, Christo Kirov, Mans Hulden, and Jason Eisner. 2019. On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics*, 7:327–342.

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

Michael A Covington and Joe D McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of quantitative linguistics*, 17(2):94–100.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.3)*. Zenodo.

Nadine El-Naggar, Tatsuki Kuribayashi, and Ted Briscoe. 2025. GCG-based artificial languages for evaluating inductive biases of neural language models. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 540–556, Vienna, Austria. Association for Computational Linguistics.

Júlia Falcão, Claudia Borg, Nora Aranberri, and Kurt Abela. 2024. COMET for low-resource machine translation evaluation: A case study of English-Maltese and Spanish-Basque. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3553–3565, Torino, Italia. ELRA and ICCL.

Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. In

*Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60, Vancouver. Association for Computational Linguistics.

Rob Van Der Goot, Esther Ploeger, Verena Blaschke, and Tanja Samardzic. 2025. DistaLs: a comprehensive collection of language distance measures. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 307–318, Suzhou, China. Association for Computational Linguistics.

Pierre Guiraud. 1959. Problèmes et méthodes de la statistique linguistique. *(No Title)*.

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

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Hieu Hoang. 2016. Is neural machine translation ready for deployment? a case study on 30 translation directions. In *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C. International Workshop on Spoken Language Translation.

Patrick Juola. 1998. Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics*, 5(3):206–213.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

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

Amir Hossein Kargaran, François Yvon, and Hinrich Schütze. 2025. Glotcc: An open broad-coverage commoncrawl corpus and pipeline for minority languages.

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

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinthór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Ilia Kulikov, Alexander Miller, Kyunghyun Cho, and Jason Weston. 2019. Importance of search and evaluation strategies in neural dialogue modeling. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 76–87, Tokyo, Japan. Association for Computational Linguistics.

Tatsuki Kuribayashi, Ryo Ueda, Ryo Yoshida, Yohei Oseki, Ted Briscoe, and Timothy Baldwin. 2024. Emergent word order universals from cognitively-motivated language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14522–14543, Bangkok, Thailand. Association for Computational Linguistics.

Natalia Levshina. 2019. Token-based typology and word order entropy: A study based on universal dependencies. *Linguistic Typology*, 23(3):533–572.

Natalia Levshina, Savithry Namboodiripad, Marc Allassonnière-Tang, Mathew Kramer, Luigi Talamo, Annemarie Verkerk, Sasha Wilmoth, Gabriela Garrido Rodriguez, Timothy Michael Gupton, Evan Kidd, et al. 2023. Why we need a gradient approach to word order. *Linguistics*, 61(4):825–883.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Luke Maurits, Dan Navarro, and Amy Perfors. 2010. Why are some word orders more common than others? a uniform information density account. *Advances in neural information processing systems*, 23.

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

Sabrina J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What kind of language is hard to language-model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989, Florence, Italy. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Chanjun Park, Yeongwook Yang, Kinam Park, and Heuiseok Lim. 2020. Decoding strategies for improving low-resource machine translation. *Electronics*, 9(10):1562.

Hyunji Hayley Park, Katherine J. Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. 2021. Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276.

Esther Ploeger, Johannes Bjerva, Jörg Tiedemann, and Robert Oestling. 2025. A cross-lingual perspective on neural machine translation difficulty. In *Proceedings of the Tenth Conference on Machine Translation*, pages 340–354, Suzhou, China. Association for Computational Linguistics.

Wessel Poelman, Thomas Bauwens, and Miryam de Lhoneux. 2025. Confounding factors in relating model performance to morphology. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7262–7287, Suzhou, China. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André F. T. Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual llms.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez,

Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size.

Gabriele Sarti, Arianna Bisazza, Ana Guerberof-Arenas, and Antonio Toral. 2022. DivEMT: Neural machine translation post-editing effort across typologically diverse languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7795–7816, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lucas Shen. 2022. LexicalRichness: A small module to compute textual lexical richness.

Chufan Shi, Haoran Yang, Deng Cai, Zhisong Zhang, Yifan Wang, Yujiu Yang, and Wai Lam. 2024. A thorough examination of decoding methods in the era of LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8601–8629, Miami, Florida, USA. Association for Computational Linguistics.

Archchana Sindhujan, Diptesh Kanojia, Constantin Orasan, and Shenbin Qian. 2025. When LLMs struggle: Reference-less translation evaluation for low-resource languages. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 437–459, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anushka Singh, Ananya Sai, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, and Mitesh Khapra. 2024. How good is zero-shot MT evaluation for low resource Indian languages? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 640–649, Bangkok, Thailand. Association for Computational Linguistics.

Hedvig Skirgård, Hannah J Haynie, Damián E Blasi, Harald Hammarström, Jeremy Collins, Jay J Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, et al. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16):eadg6175.

James P Trujillo and Judith Holler. 2024. Information distribution patterns in naturalistic dialogue differ across languages. *Psychonomic Bulletin & Review*, 31:1723–1734.

Ada Wan. 2022. Fairness in representation for multilingual nlp: Insights from controlled experiments on conditional language modeling. In *International Conference on Learning Representations*.

Jiayi Wang, David Ifeoluwa Adelani, and Pontus Stenetorp. 2024. Evaluating WMT 2024 metrics shared task submissions on AfriMTE (the African challenge set). In *Proceedings of the Ninth Conference on Machine Translation*, pages 505–516, Miami, Florida, USA. Association for Computational Linguistics.

Sean Welleck, Amanda Bertsch, Matthew Finlayson, Hailey Schoelkopf, Alex Xie, Graham Neubig, Ilia Kulikov, and Zaid Harchaoui. 2024. From decoding to meta-generation: Inference-time algorithms for large language models. *Transactions on Machine Learning Research*. Survey Certification.

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

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Aris Xanthos, Sabine Laaha, Steven Gillis, Ursula Stephany, Ayhan Aksu-Koç, Anastasia Christofidou, Natalia Gagarina, Gordana Hrzica, F Nihan Ketrez, Marianne Kilani-Schoch, et al. 2011. On the role of morphological richness in the early development of noun and verb inflection. *First Language*, 31(4):461–479.

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

## A Target Languages

Table 4 provides the list of target languages used in our experiments.

| | | |
|---|---|---|
| Acehnese (Arabic script) | **Armenian** | **Portuguese (Brazilian)** |
| Acehnese (Latin script) | Igbo | Dari |
| Mesopotamian Arabic | Ilocano | Ayacucho Quechua |
| **Afrikaans** | **Indonesian** | **Romanian** |
| Amharic | **Icelandic** | Rundi |
| **Modern Standard Arabic** | **Italian** | **Russian** |
| Moroccan Arabic | **Japanese** | Sango |
| Egyptian Arabic | Jingpho | **Sanskrit** |
| Assamese | Kamba | **Slovak** |
| Central Aymara | Kannada | **Slovenian** |
| South Azerbaijani | Georgian | Samoan |
| Bashkir | **Kazakh** | Shona |
| Bambara | Halh Mongolian | Somali |
| Balinese | Khmer (Central) | Southern Sotho |
| **Belarusian** | Kikuyu | **Spanish (Latin American)** |
| Bhojpuri | Kyrgyz | **Serbian** |
| Lhasa Tibetan | Northern Kurdish | Swati |
| **Bulgarian** | Central Kanuri (Arabic script) | Sundanese |
| **Catalan** | Central Kanuri (Latin script) | **Swedish** |
| **Czech** | **Korean** | Swahili |
| Central Kurdish | Lao | **Tamil** |
| **Mandarin Chinese (Standard Beijing)** | **Lithuanian** | Tamasheq (Latin script) |
| **Mandarin Chinese (Taiwanese)** | Ganda | Tamasheq (Tifinagh script) |
| **Welsh** | Luo | Tatar |
| **Danish** | Mizo | **Telugu** |
| **German** | Maithili | Tajik |
| **Estonian** | Malayalam | Thai |
| **Greek** | **Marathi** | Tigrinya |
| **English** | Minangkabau (Latin script) | Tswana |
| **Basque** | Meitei (Manipuri, Bengali script) | Turkmen |
| Ewe | Mossi | **Turkish** |
| Fijian | Maori | **Uyghur** |
| **Finnish** | Burmese | **Ukrainian** |
| **French** | **Dutch** | **Urdu** |
| **Scottish Gaelic** | Nepali | **Vietnamese** |
| **Irish** | Nuer | Waray |
| **Galician** | Odia | **Wolof** |
| Hausa | Pangasinan | Xhosa |
| **Hebrew** | Eastern Panjabi | Yoruba |
| **Hindi** | **Western Persian** | Yue Chinese (Hong Kong Cantonese) |
| **Croatian** | Plateau Malagasy | Zulu |
| **Hungarian** | **Polish** | |

Table 4: 125 target languages used for our *correlation* experiments (including English). Languages **in bold** are used in *regression* experiments (53 languages).

## B More Details about the Language Properties

Besides language resourcedness and coarse-grained source-target distances, we focus on features of morphological complexity and word order of the language being generated. With respect to **morphology**, current subword segmentation methods have been proved less effective for more morphologically complex languages (Park et al., 2021). **Word order** flexibility is linked to entropy and potentially less predictable sequences. Moreover, word order properties can determine how information is distributed within the sentence (Maurits et al., 2010; Trujillo and Holler, 2024),[16] which in turn could affect the optimal decoding strategy of a given target language.

### B.1 Precomputed Distances

Following previous work on NMT (Sarti et al., 2022) and cross-lingual transfer (Lin et al., 2019), we take advantage of the URIEL typological database (Littell et al., 2017) containing vector representations of numerous languages drawn from typological, geographical, and phylogenetic databases. Using the accompanying `lang2vec` library[17], we query six types of precomputed distances between each FLORES+ language and seven

[16] Some languages tend to place less predictable words and phrases in the first half of the utterance (e.g. German, Japanese), whereas others follow the opposite pattern (e.g. English, Mandarin).

[17] `github.com/antonisa/lang2vec`.

source languages: *genetic*, *geographic*, *syntactic*, *inventory*, *phonological*, and *overall features*.

### B.2 Morphological Complexity Measures

To estimate morphological complexity of a language, we make use of eight publicly available precalculated continuous measures from Çöltekin and Rama (2023), outlined below. The measures were computed on the basis of Universal Dependencies and are available for 33 languages of the FLORES+ dataset.

**Type/Token Ratio (TTR)** Number of unique word types in a text divided by the total number of word tokens in a text (Chotlos, 1944; Covington and McFall, 2010). Since TTR lies in the range of $[0; 1]$, languages where this measure is closer to 1 will have a higher number of unique word forms in part motivated by inflectional morphology, which we expect to negatively affect translation difficulty.

**Information in Word Structure (WS)** Compares the information content (i.e. entropy) of the original text with its compressed version (Juola, 1998). The expectation here is that languages with more complex morphology will have worse compression ratios.

**Word and Lemma Entropy (WH, LH)** Word entropy is calculated on the basis of word frequency distribution of a text, with morphologically complex languages having higher word entropy (Bentz et al., 2016). Çöltekin and Rama (2023) additionally calculate the entropy of lemmas. Since lemmas do not include inflectional markers, a high degree of lemma entropy would then point at more derivational morphology and compounding.

**Mean Size of Paradigm (MSP)** Calculated by dividing the number of word forms in a text by the number of lemmas (Xanthos et al., 2011). Languages with richer inflectional morphology are expected to have a higher number of paradigm cells, reflected by the MSP value.

**Inflectional Synthesis (IS)** Maximum number of inflection categories that can be expressed by a standalone verb.

**Morphological Feature Entropy (MFH)** Reflects the usage of morphological features and their values. Languages with a higher number of approximately uniformly used grammatical cases will have higher entropy values, indicating more intricate inflectional morphology.

**Inflection Accuracy (IA)** Accuracy of an ML model on the task of predicting inflected forms of words given their lemmas and grammatical features. The intuition for IA is that if a language has complex morphology, inflection accuracy on a hold-out test set will be low. Thus, for the sake of consistency with the rest of the measures, we report negative inflection accuracy (`-ia`).

**TTR Measured on FLORES+** We leverage the LexicalRichness Python module (Shen, 2022) to calculate three TTR measures on the data for the languages covered by the `dev` split of FLORES+: *TTR*, *Root Type/Token Ratio* (RTTR) (Guiraud, 1959), and *Moving Average Type/Token Ratio* (MATTR) (Covington and McFall, 2010).

### B.3 Gradient Word Order Measures

We include a number of gradient word order measures proposed in Levshina (2019) and Levshina et al. (2023).

**Average Word Order Entropy of Dependents and Codependents** Entropy of different word order patterns of *dependencies* (e.g. verb-subject and noun-adposition relations) and *codependencies* (e.g. subject and object of the same verb).

**Proportion of Subject-Object Order** Proportions based on the frequencies of phrases where subject comes before object. Proportions closer to 1 indicate strong preference of a language towards either order of subject and object, while proportions closer to 0.5 mean that a language tends to use the two orders interchangeably. For consistency with other metrics, we report results with *negative* SO order proportion `-SO_prop`, such that higher `-SO_prop` values reflect *lower* predictability of word order.

**Percentage of Head-Final Phrases** Approximates the preference of a language towards head-initial or head-final phrases. Values closer to 1 indicate a larger prevalence of head-finality.

## C Tower+ Translation Prompt

```
Translate the following {src_lang} source text to {tgt_lang}:\n{src_lang}: {src_text}\n{tgt_lang}:
```

## D Overall Results with Additional Metrics

Figure 4 shows NLLB-200 translation quality results with different metrics (BLEU, chrF++, COMET), as well as sentence generation probabilities, as a function of beam size. Scores are averaged across 124 target languages. Trends by beam size are very similar across metrics. While a wider beam generally improves translation quality and the model's confidence during output sequence generation, regardless of the source language, it is not clear whether the *degree* of this improvement varies depending on the typological properties of target languages, motivating our correlation and regression analyses.

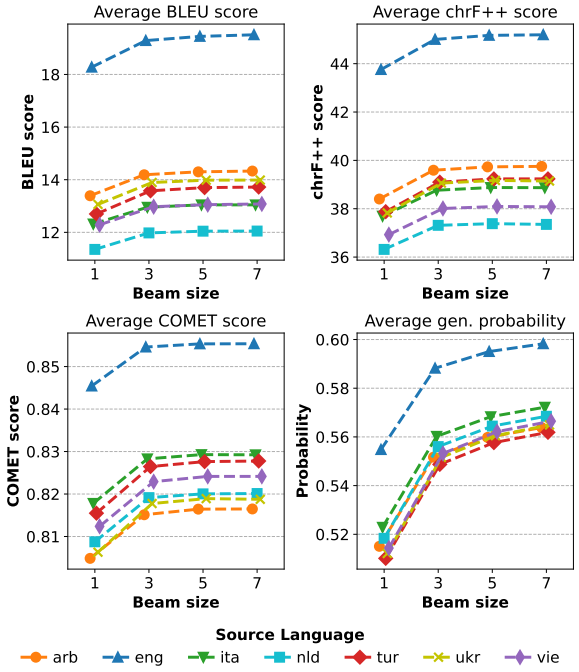


Figure 4: NLLB-200 translation performance measured at four beam sizes via translation quality metrics (BLEU, chrF++, COMET) and output sequence generation probabilities. Performance is averaged across 124 target languages for individual source languages: Arabic, English, Italian, Dutch, Turkish, Ukrainian, and Vietnamese.

## E Effect of Head-Finality and MATTR

Figure 5 provides a closer look at the negative effect of target language head finality and moving average type-token ratio on NLLB-200 chrF++ scores when translating from English.

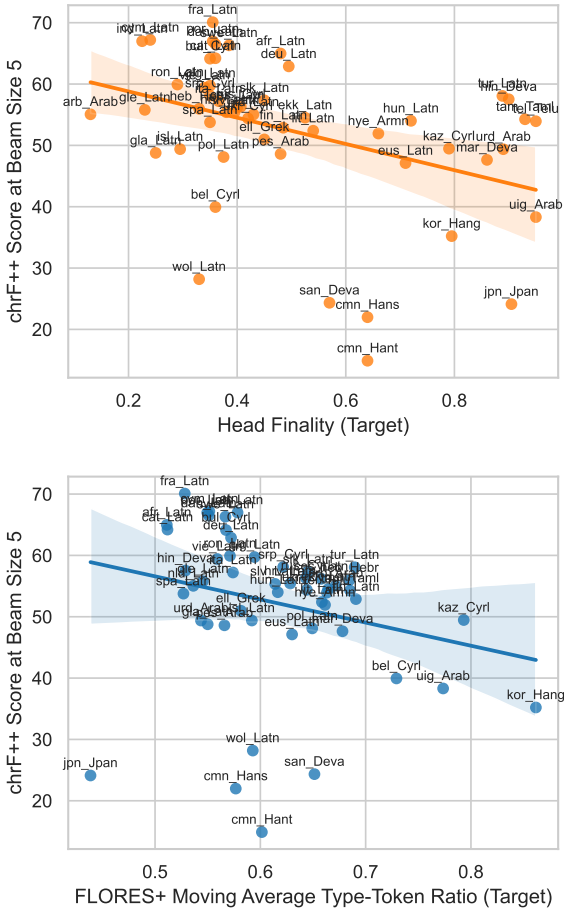


Figure 5: Target language head-finality (top) and moving average type-token ratio (bottom) vs. chrF++ scores at beam size 5. Data for translations by NLLB-200 from English into 52 target languages.

## F Additional Regression Results

### F.1 NLLB-200

Table 5 summarizes regression results for predicting chrF++ and probability gain for translations by NLLB-200. Here we indicate the coefficient values $\beta$ for each source language, with English as a reference value.

### F.2 Tower+: chrF++ gain

Table 6 showcases regression results for predicting chrF++ gain of Tower+ translations.

| | *Metric*: **chrF++ Gain** ($n = 364$) | | | | | *Metric*: **Probability Gain** ($n = 364$) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Feature** | ***β*** | ***F*-statistic** | ***p*-value** | ***LR*** | **Adj. $R^2$** | ***β*** | ***F*-statistic** | ***p*-value** | ***LR*** | **Adj. $R^2$** |
| src_language | – | 4.40 | <0.001 | – | 0.05 | – | 3.6 | 0.002 | – | 0.01 |
| *Arabic* | **-0.26** | | | | | 0.09 | | | | |
| *Dutch* | **-0.60** | | | | | **0.19** | | | | |
| *Italian* | **-0.37** | | | | | **0.19** | | | | |
| *Turkish* | -0.12 | | | | | **0.19** | | | | |
| *Ukrainian* | -0.11 | | | | | **0.21** | | | | |
| *Vietnamese* | **-0.36** | | | | | **0.19** | | | | |
| tgt:glotcc_size | **0.15** | 17.2 | <0.001 | 14.2* | 0.09 | **-0.21** | 180.8 | <0.001 | 141.8* | 0.33 |
| is_same_script | -0.02 | 0.06 | 0.809 | 1.72 | 0.09 | **-0.08** | 4.5 | 0.03 | 37.2* | 0.39 |
| d_gen | 0.05 | 1.48 | 0.224 | 2.00 | 0.09 | **0.09** | 26.9 | <0.001 | 29.7* | 0.44 |
| tgt:mattr | 0.02 | 0.35 | 0.555 | 1.04 | 0.09 | **0.05** | 10.6 | 0.001 | 23.1* | 0.47 |
| tgt:head_fin | 0.06 | 2.38 | 0.124 | 2.45 | 0.09 | **0.12** | 46.3 | <0.001 | 45.0* | 0.53 |

Table 5: Regression results for chrF++ gain (*left*) and probability gain (*right*) for NLLB-200 translations (**all source languages**); model is fitted on the results of all source languages. Coefficient values (***β***) marked in bold are significant ($p < 0.05$).

| | ***Metric***: $\Delta chrF_{1;7}$ ($n = 364$) | | | | |
|---|---|---|---|---|---|
| **Feature** | ***β*** | ***F*** | ***p*-value** | ***LR*** | **Adj. $R^2$** |
| src_language | 0.0 | 1.74 | 0.11 | - | 0.01 |
| is_in_tower | -0.48 | 1.35 | 0.25 | 0.74 | 0.01 |
| tgt:glotcc_size | **0.66** | 11.1 | <0.001 | 13.61* | 0.04 |
| is_same_script | -0.6 | 2.19 | 0.14 | 0.29 | 0.04 |
| d_gen | 0.11 | 0.38 | 0.54 | 0.12 | 0.04 |
| tgt:mattr | -0.11 | 0.36 | 0.55 | 1.77 | 0.04 |
| tgt:head_fin | **-0.47** | 6.42 | 0.01 | 6.6* | 0.05 |

Table 6: Regression results for predicting chrF++ gain for translations generated by **Tower+** (**all src langs**).

## G Correlation Results for All Source Languages

Here we include Spearman correlation results for chrF++ scores at beam size $k = 5$ and generation probability gains for NLLB-200 translation from six source languages:

- Arabic: Figure 6
- Italian: Figure 7
- Dutch: Figure 8
- Turkish: Figure 9
- Ukrainian: Figure 10
- Vietnamese: Figure 11

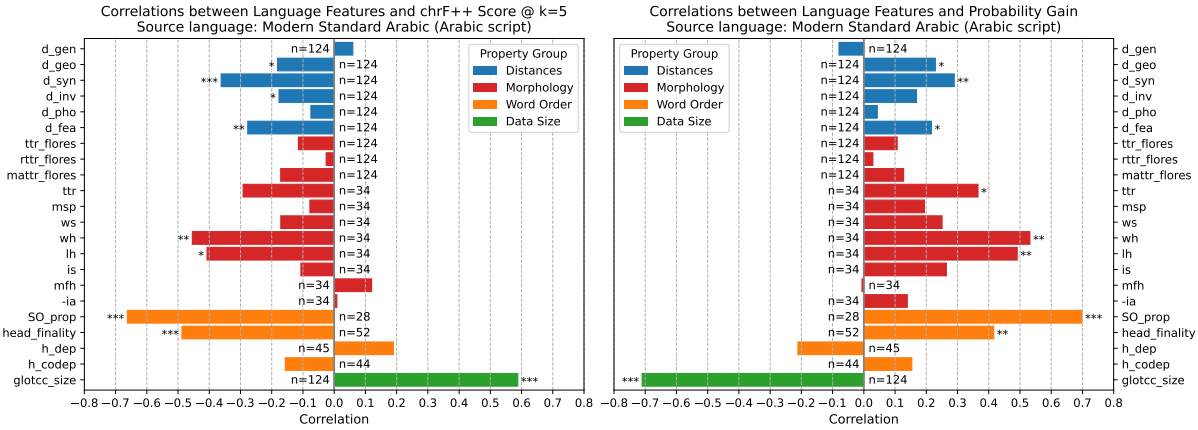


Figure 6: Spearman correlations between continuous language properties and chrF++ scores at beam size $k = 5$ (left) and probability gain for beam size $k = 7$ (right). Source language: **Arabic**.

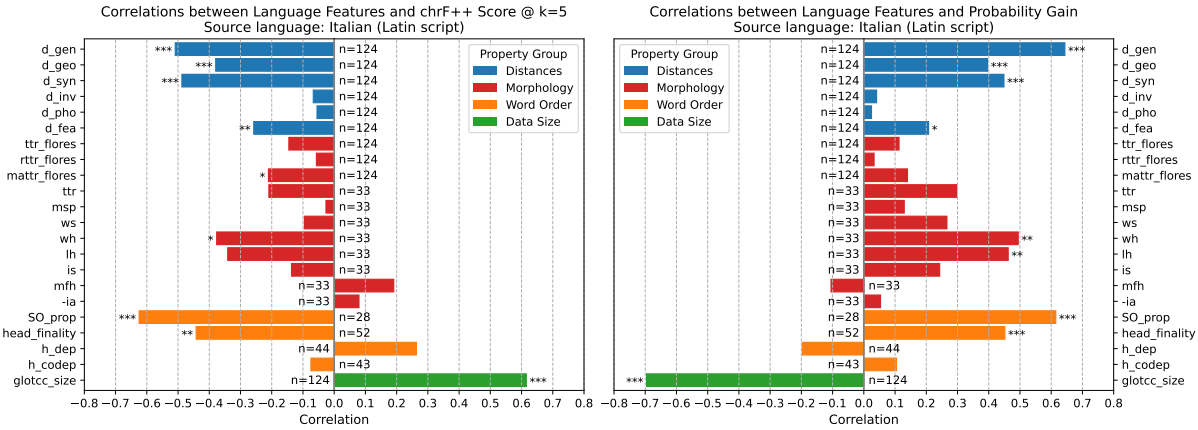


Figure 7: Spearman correlations between continuous language properties and chrF++ scores at beam size $k = 5$ (left) and probability gain for beam size $k = 7$ (right). Source language: **Italian**.

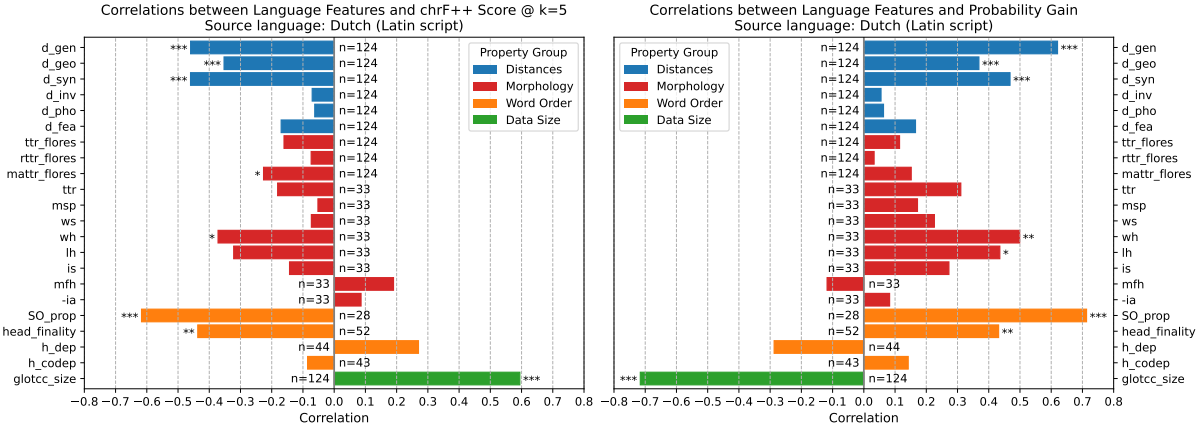


Figure 8: Spearman correlations between continuous language properties and chrF++ scores at beam size $k = 5$ (left) and probability gain for beam size $k = 7$ (right). Source language: **Dutch**.

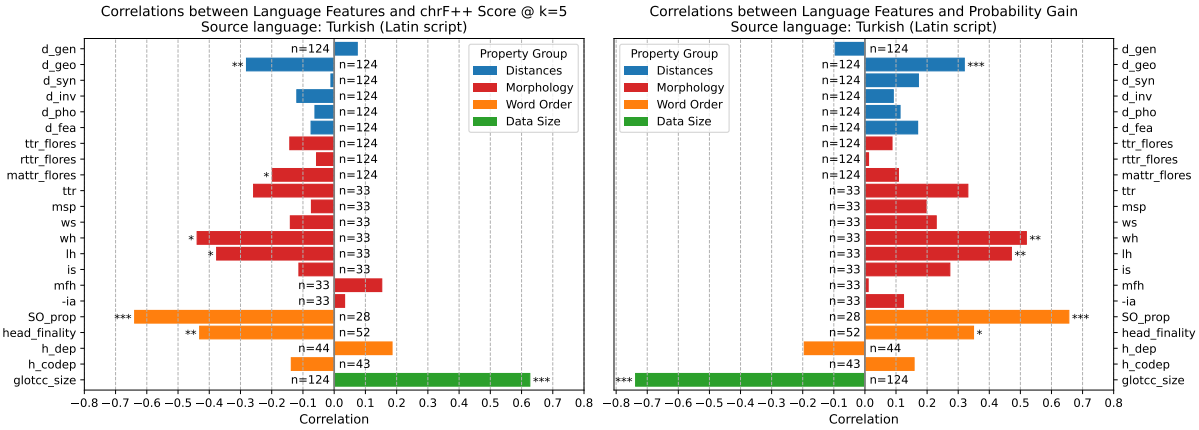


Figure 9: Spearman correlations between continuous language properties and chrF++ scores at beam size $k = 5$ (left) and probability gain for beam size $k = 7$ (right). Source language: **Turkish**.

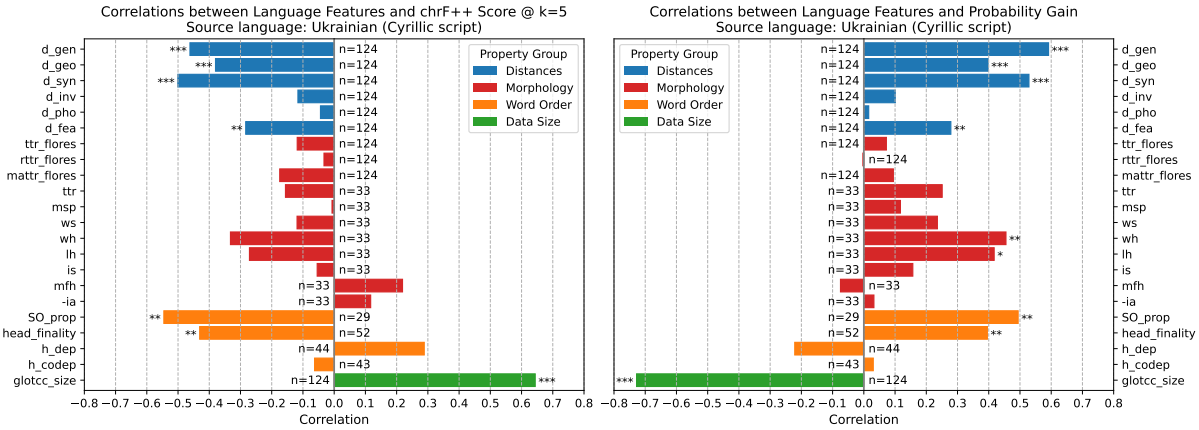


Figure 10: Spearman correlations between continuous language properties and chrF++ scores at beam size $k = 5$ (left) and probability gain for beam size $k = 7$ (right). Source language: **Ukrainian**.

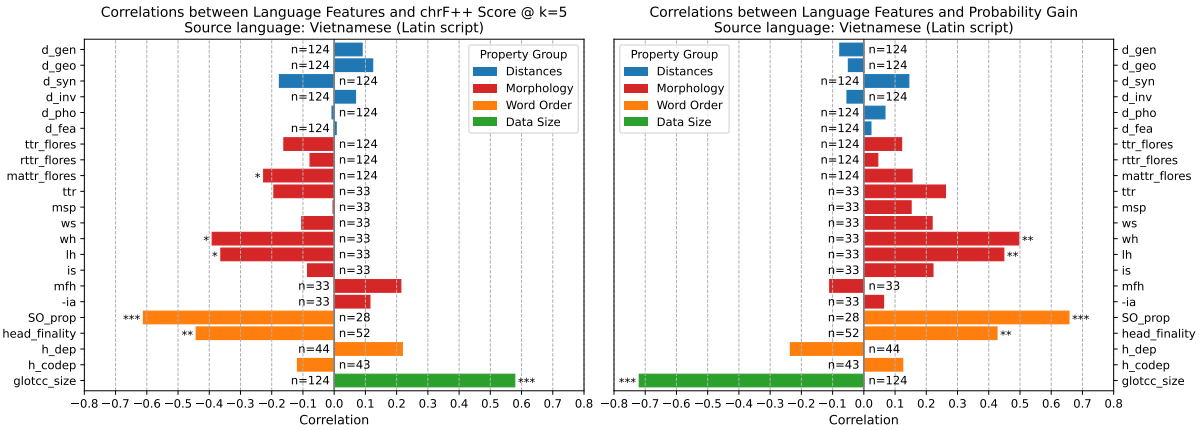


Figure 11: Spearman correlations between continuous language properties and chrF++ scores at beam size $k = 5$ (left) and probability gain for beam size $k = 7$ (right). Source language: **Vietnamese**.