university of groningen

faculty of arts

Explaining Machine Translation Difficulty in the Age of Massively Multilingual Models: A Study of the Impact of Fine-Grained Typological Features and Beam Size on State-of-the-Art NMT

Vitalii Hirak

university of groningen / faculty of arts

University of Groningen

Explaining Machine Translation Difficulty in the Age of Massively Multilingual Models: A Study of the Impact of Fine-Grained Typological Features and Beam Size on State-of-the-Art NMT

Master's Thesis

To fulfill the requirements for the degree of
Master of Arts in Linguistics (research)
at University of Groningen under the supervision of
Dr. Arianna Bisazza (Center for Language and Cognition, University of Groningen)
and
Prof. Dr. Josef van Genabith (German Research Center for Artificial Intelligence, Saarland University)

Vitalii Hirak (s5741467)

December 16, 2024

Contents

Acknowledgments

First and foremost, I want to express my sincerest gratitude to my supervisors, Dr. Arianna Bisazza and Prof. Dr. Josef van Genabith, for all their invaluable guidance, patience, and encouragement in bringing this thesis to life.

I am also incredibly grateful for the opportunity to have studied in the Erasmus Mundus European Masters in Language and Communication Technologies.

Furthermore, I am thankful to the Center for Information Technology of the University of Groningen for their support and for providing access to the Hábrók high-performance computing cluster.

Last but not least, I want to thank all my friends and family for their unwavering love and support during the time I was working on my thesis and throughout the entirety of my study program.

Abstract

Despite the impressive progress of modern neural machine translation (NMT) systems, languages with typological properties significantly different from English, such as morphologically complex languages or languages with flexible word order, have been shown to pose a bigger challenge for NMT. The primary decoding algorithm used by NMT systems to generate output sequences is beam search. While increasing beam size generally improves translation performance, the relationship between morphosyntactic properties, the width of output search space, and translation performance remains understudied. With this work, we aim to fill this gap and investigate whether languages with higher morphological complexity and/or word order freedom benefit more from widening the output search space and thus may require alternative decoding mechanisms. First, we compile and release publicly a set of robust typological properties for the languages in the FLORES+ MT evaluation benchmark. Using the state-of-the-art NLLB-200 model, we translate 997 FLORES+ English sentences into 124 languages under four beam size configurations and evaluate NMT difficulty via a combination of translation quality metrics and generation probabilities. We then perform a series of correlation studies to assess the effect of categorical and continuous morphosyntactic properties on translation quality and modeling probability gains. Our results demonstrate that languages with certain typological properties significantly different from English see a greater improvement in translation performance from increasing beam size. These findings suggest that such languages may benefit from decoding strategies other than the current de facto standard of left-to-right beam search.

1 Introduction

Machine translation (MT) – the process of automatically translating a text from a source language into a target language with a computer – has proven itself to be an integral technology in the digital age. As just one example, it helps address the problem of digital divide – the fact that the dominant language of content available online is English [1] – and facilitate access to information for users in a diverse set of languages.

Throughout its history, machine translation has undergone several paradigm shifts, and most modern MT systems such as Google Translate[1]and DeepL[2] are based on neural networks. Neural machine translation (NMT) systems are trained on massive amounts of parallel data and achieve state-of-the-art translation performance [2, 3], often surpassing that of traditional phrase-based statistical machine translation (SMT) models [4, 5]. This is especially true for morphologically complex languages and languages with significantly different word order from English [6, 7].

However, even though modern natural language processing (NLP) tools, including NMT systems, are supposed to be language-agnostic and therefore work well with any language given sufficient training data, in practice this is often not the case [8, 9]. To some extent, this inequality can be explained by the imbalance in training data, as most of NLP research is English-centric [10]. However, even with similar amounts of data, some languages still pose a bigger challenge for various NLP tasks, including language modeling (LM) [11, 12] and machine translation [13].

Whether this is caused by intrinsic language properties such as complex morphology or word order flexibility has been a subject of active discussion. While some works provide evidence in support of this hypothesis [14, 15, 16], others fail to find such an association and instead argue that more general factors like word inventory, raw character sequence length, and word segmentation are at play [11, 17]. Needless to say, the relationship between typological properties of a language and the difficulty of language modeling and machine translation is a complex one and remains poorly understood.

The primary decoding algorithm used by LM and NMT systems to generate output sequences is *beam search* [3, 18, 19, 20]. Beam search allows a model to efficiently and effectively navigate a space of

[1] `https://translate.google.com/`.

[2] `https://www.deepl.com/`.

possible output sequences (translations in the case of NMT) [21]. However, despite its ubiquity, the findings above may hint at the fact that beam search might be a suboptimal decoding strategy for target languages whose typological properties are considerably different from English.

While previous work focuses on the influence of word order flexibility and morphological complexity on the difficulty of NMT [22, 13, 9, 15], to our best knowledge, the relationship between morphosyntactic features of a language, the width of the output search space, and translation difficulty has not been studied directly. With this work, we aim to explore this relationship to better understand whether languages with more complex morphology and word order might call for different decoding strategies for optimal generation results. We begin with collecting a set of robust language properties for a large sample of typologically diverse languages included in the FLORES+[3] [18] translation evaluation benchmark dataset. These properties include corpus-based measures taken both from existing work, as well as calculated on the FLORES+ dataset. In addition, we make this dataset publicly available at `https://github.com/v-hirak/explaining-MT-difficulty`. We then leverage the recently released state-of-the-art NLLB-200 NMT model [18] to translate the sentences in the FLORES+ dataset from English into 124 languages under four different beam size configurations. The difficulty of generating the translations is measured by means of several translation quality metrics, as well as the final probabilities of the generated sequences. Finally, we look for relationships between the collected typological data and translation probability and quality scores, taking into account the beam size used for translation.

1.1 Research Question

Primarily, our experiments aim to explore whether target languages with varying degrees of morphological complexity and word order flexibility benefit more from widening the output search space. In order to answer this question, we collect a large number of language features and study their correlations with the quality of NMT using a large and typologically diverse set of languages as well as a state-of-the-art NMT model.

[3] `https://github.com/openlanguagedata/flores`.

1.2 Thesis Outline

The thesis is organized in the following way: §2 introduces important background knowledge relevant to this work; §3 describes the methodology pertaining to the collection of linguistic data, translation configuration, and translation difficulty assessment; in §4 we outline the experimental setup used for investigating relations between the collected data; §5 delves into the results obtained during our experiments; finally, §6 provides a conclusion to our work, summarizing the main findings as well as outlining the venues for potential future work.

2 Background Literature

This section introduces the context and previous research related to this work. We begin with a brief description of beam search, the most common decoding algorithm, and investigate the influence of varying beam size on the quality of NMT. After that, we look into the existing discrete and continuous approaches towards language typology and morphosyntactic complexity. Finally, we explore the previous research into the role of morphological and word order properties in language modeling and machine translation.

2.1 Beam Search

The primary goal of machine translation is to produce the most likely translation of a given input sequence [23]. To accomplish this, the model needs to determine the probability of the output sequence, which may be represented as the product the probabilities of its individual tokens, formalized as follows:

$$\prod_{t=1}^{T} P(y_t|y_1,...,y_{t-1},\mathbf{x}) \tag{1}$$

where t is a time step during decoding, T is the number of time steps, y_t is a token at time step t, and $\mathbf{x}$ is the input sequence context.

In modern NMT systems such as NLLB-200, sequence generation is handled by the *decoder*, which generates the output translation sequence one token at a time [18]. At each decoding step, the model chooses a token from its vocabulary, taking into account the input sequence and the preceding context. The most straightforward and least computationally expensive approach would be to simply pick the most likely token at each time step. If we represent the model's vocabulary as V and the input sequence context as $\mathbf{x}$, then the model's goal would be to pick a token with the highest conditional probability from the vocabulary, given the previously generated tokens and the input sequence. This strategy is known as *greedy search* [24] and can be formalized as follows:

$$y_t = \arg\max_{y \in V} P(y|y_1,...,y_{t-1},\mathbf{x}) \tag{2}$$

where V is the model's token vocabulary, y_t is the token at time step t, and $\mathbf{x}$ is the input sequence context vector.

While this strategy may seem completely reasonable, a sequence of most likely tokens would not

necessarily be the same as the most probable sequence overall. A toy example in Figure 1 illustrates this point.

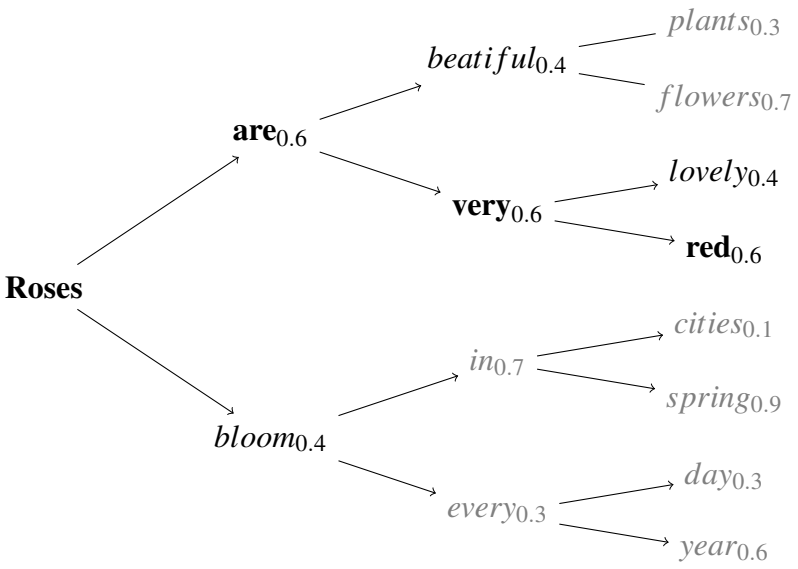

Figure 1: Example of greedy search. At each decoding step the model selects the token **in bold** (except **Roses**). The *greyed out* tokens are not considered during greedy search at all since they follow from the tokens discarded at the preceding decoding step.

If at each decoding step we were to select a token with the highest probability, we would produce the sequence *Roses are very red*. The probability of this sequence would be equal to the product of the probabilities of its individual tokens: $0.6 \times 0.6 \times 0.6 = 0.216$. However, the most probable sequence in the example in Figure 1 is actually *Roses bloom in spring* ($0.4 \times 0.7 \times 0.9 = 0.252$), and this sequence is only possible if we choose the token *bloom* instead of *are* at the first step.

The most desirable strategy (known as *exhaustive search* [24]) would then be to keep track of all potential sequences at all times and output the most probable one at the end of the generation process. However, the complexity of this approach would be $O(|V|^T)$, where T is the number of time steps (i.e. the length of the generated sequence in tokens). This makes it too computationally expensive and thus impractical to use in real-world scenarios.

Beam search acts as a compromise between the two strategies: it allows for exploring a large space of possible translations during decoding without being too computationally prohibitive. A hyperparameter known as *beam size* or *beam width* tells the model to only keep track of k best tokens at each generation step, eventually choosing between k best possible translations. Figure 2 provides a

modified example from Figure 1 to illustrate the work of beam search.

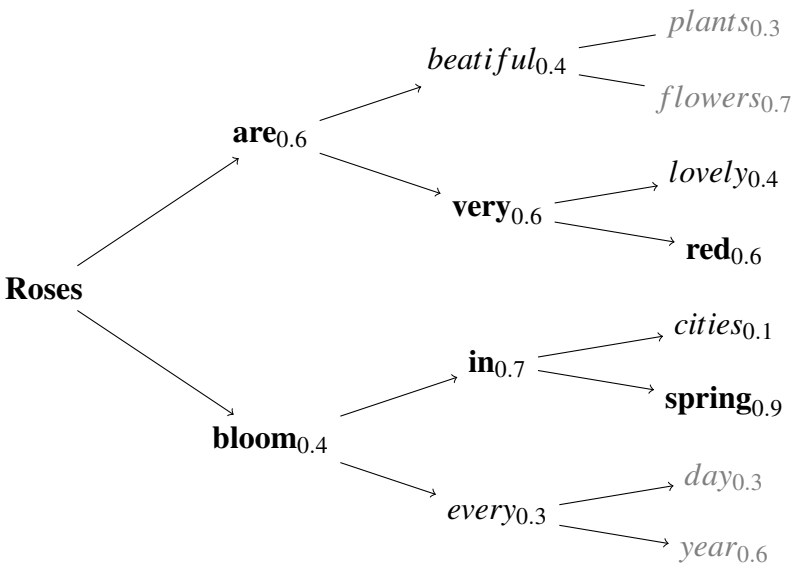

Figure 2: Example of beam search with beam size $k = 2$. At each time step the model selects two most probable tokens (**in bold**) and continues the search only from those tokens. The *greyed out* tokens are not considered during beam search at all since they follow from the tokens discarded at the preceding decoding step.

The ability of beam search to keep track of multiple potential output candidates is especially important when modeling (and translating from/into) languages with higher morphological complexity and/or word order freedom compared to English. Consider a case of translating from Polish into English: while the latter relies on a fixed word order (Subject-Verb-Object, SVO) to denote grammatical relations between syntactic arguments, Polish is characterized by higher word order flexibility and in turn often uses inflectional morphology to express head-dependent relations. Take, for example, the following simple sentence in Polish:

(3) *Kot-a* *miała* *Alicj-a.*
Cat-ACC have.3SG.F.PST Alice-NOM
'Alice had a cat.'

The sentence comes in the Object-Verb-Subject (OVS) order, where the object *Kota* and subject *Alicja* have accusative and nominal cases respectively to indicate their grammatical roles. In turn, the verb *miała* agrees with the subject in person, number, and gender. However, an English version of this sentence, *Alice had a cat*, comes with the SVO order and no case marking. In order to correctly translate the sentence from Polish into English while taking into account the change in word order,

the model has to keep track of several translation candidates at a time (e.g. one starting with *A cat* and one starting with *Alice*). Similarly, translating from English into Polish also necessitates the use of beam search, since a verb form in English may have multiple corresponding inflected forms in Polish.

An important consideration regarding beam search is that the generated sequences can have different lengths, which would influence their final probabilities: longer strings would have lower probabilities and vice versa. To mitigate this, the final sequence probabilities are often normalized, for example, by dividing them over the number of tokens in the respective sequences.

2.2 Beam Size and Translation Quality

The choice of beam width for decoding involves dealing with a trade-off between translation quality and decoding speed. Generally, a larger beam size leads to better translation accuracy since it enables exploring a broader space of candidate sequences, thus increasing the likelihood of identifying the most optimal translation. However, this improvement in quality comes at a greater computational cost, which can substantially slow down the decoding process [25, 26, 27]. For instance, as part of their experiments, Junczys-Dowmunt et al. (2016) [25] performed neural machine translation from English into French using their proprietary decoder[4] and studied the influence of beam size on translation quality (in BLEU) and decoding speed (in words per second, WPS). The relevant plot can be found in Figure 3. They observed that beam sizes beyond 5-7 do not improve translation quality substantially, but lead to a significantly slower decoding speed.

Further complicating matters is the phenomenon whereby translation quality scores tend to deteriorate beyond a certain optimal beam width. This has been demonstrated in works such as Koehn and Knowles (2017) [28] and Cohen and Beck (2019) [29]. Particularly, Koehn and Knowles (2017) perform neural machine translation for eight language pairs[5] using the Nematus NMT toolkit[6] [30] and find that for almost all language pairs, translations get worse beyond an optimal beam size (depending on the language). Normalizing sequence probabilities by the output length alleviates the decrease in quality to a certain extent, but, as Tu et al. (2017) [31] argue, does not fix it entirely. Similarly, Cohen and Beck (2019) use convolutional NMT models to translate from English into German and French

[4] `https://github.com/marian-nmt/marian`.

[5] Czech-English, English-Czech, German-English, English-German, Romanian-English, English-Romanian, Russian-English, and English-Russian.

[6] `https://github.com/EdinburghNLP/nematus`.

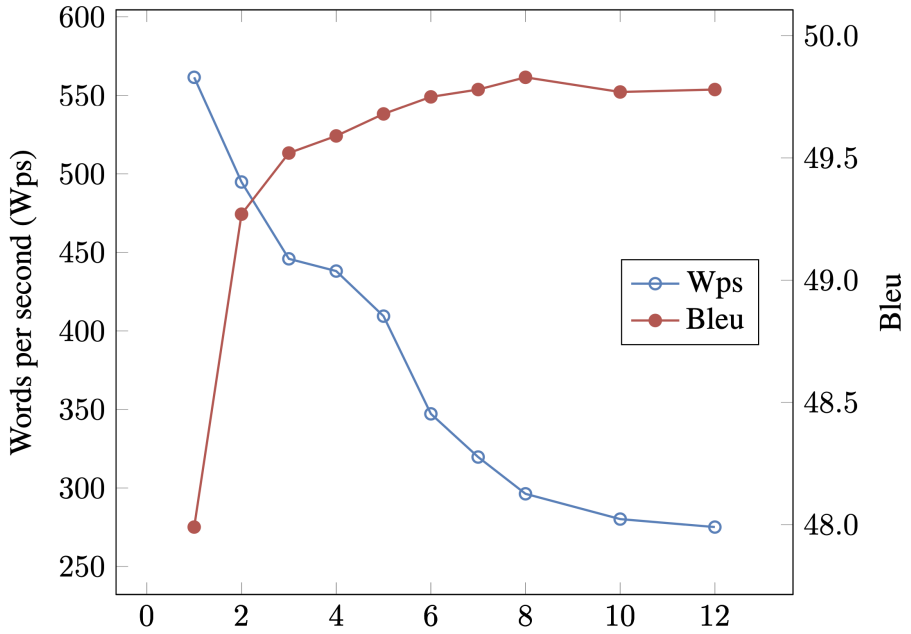

Figure 3: Beam size vs. translation quality (BLEU) and decoding speed (WPS). Figure taken from Junczys-Dowmunt et al. (2016).

under various beam width settings[7]. They find that for both translation directions the BLEU scores peak at the beam width of 5, with the quality scores worsening at bigger beam sizes.

For the reasons highlighted above, it is common to use smaller beam widths of 3 to 5 in the context of NMT [32]. Similarly, we use beam sizes of 1, 3, 5, and 7 to carry out the experiments in this work.

2.3 Language Properties

Before we delve into how morphosyntactic properties impact the difficulty of language modeling and, more importantly, machine translation, it is essential to explore how languages differ in their typology as well as how one can measure the morphosyntactic complexity of a language. We will first focus on discrete approaches and then move on to continuous measures of language properties.

2.3.1 Discrete Approach

Languages differ from one another in various aspects such as phonology, morphology, and syntax. One approach to measure this variation would be to assign categorical values to languages. This approach is followed in the World Atlas of Language Structures (WALS) [33]. The WALS online

[7] 1, 3, 5, 25, 100, and 250.

database[8] distinguishes between 192 features, each comprising between 2 and 28 different categorical values. The 2662 languages in the database were manually assigned categorical values for a number of linguistic properties by a team of 55 experts. Relevant to our work, the WALS database identifies 12 features in the Morphology category and 56 features in the Word Order category. As an example from the Morphology area, Feature 26A "Prefixing vs. Suffixing in Inflectional Morphology" [34] classifies 969 languages into one of six categorical values, illustrated in Table 1. In the realm of Word Order, Feature 81A "Order of Subject, Object, and Verb" [35] places 1376 languages into seven groups, as can be seen in Table 2.

Value	**Number of languages**
Little or no inflectional morphology	141
Predominantly suffixing	406
Moderate preference for suffixing	123
Approximately equal amounts of suffixing and prefixing	147
Moderate preference for prefixing	94
Predominantly prefixing	58

Table 1: Values and language representation of Feature 26A.

Value	**Number of languages**
Subject-Object-Verb (SOV)	564
Subject-Verb-Object (SVO)	488
Verb-Subject-Object (VSO)	95
Verb-Object-Subject (VOS)	25
Object-Verb-Subject (OVS)	11
Object-Subject-Verb (OSV)	4
No dominant word order	189

Table 2: Values and language representation of Feature 81A.

Discrete typological features of languages such as those highlighted above have been used in a number of works investigating the influence of various language properties on the difficulty of language modeling [11, 16] and neural machine translation [15]. Both Mielke et al. (2019) [11] and Park et al. (2021) [16] investigate associations between various WALS features[9] and language modeling sur-

[8] https://wals.info/.

[9] Specifically, features 26A and 81A in Mielke et al. (2019) and features 20A-29A in Park et al. (2021).

prisal, while Bisazza et al. (2021) [15] employ parallel artificial grammars with varying word order and presence/absence of case marking to study their effect on the difficulty of NMT. We explore these works in more detail in §2.5.

2.3.2 Continuous Approach

If we take a closer look at Feature 26A from the WALS database (Table 1), we can notice that its categorical values follow a certain order: excluding the first value, they start from "Predominantly suffixing" and, over a span of several discrete steps, gradually move towards "Predominantly prefixing". Levshina et al. (2023) propose to "take the next logical step" and move away from discrete categories entirely in favor of fully continuous variables [36]. From this point of view, we could imagine Feature 26A as a spectrum with predominantly suffixing languages on one end and predominantly prefixing languages on the opposite end, with the rest of the languages lying in between (languages with little or no inflectional morphology can be thought to lie outside this spectrum). Some works even attempt to quantify the WALS feature values such that the complexity of a language along a certain dimension can be measured on a continuous scale. For instance, Bentz et al. (2016) select 28 WALS features related to morphology and apply a number of transformations (e.g. binarization, reordering, recategorization, and normalization) to the categorical values of those features such that the final numerical values reflect the variation in morphological complexity of a language.

It has been argued that a gradient approach towards language typology more accurately reflects the variability of natural languages and is more appropriate in the realm of NLP [36, 37]. For example, instead of strictly assigning a certain word order to a language, as is the case with Feature 81A, Levshina et al. (2023) [36] suggest computing continuous values grounded in real-life usage data, such as the proportion of the Subject-Object to Object-Subject order or the fraction of head-final phrases. While discrete labels can still be useful as a shortcut, the authors argue that they act as a convenient simplification rather than a faithful representation of intrinsic language properties.

Extending the work of Levshina et al. (2023), Baylor et al. (2024) [37] further argue that using continuous typological values helps avoid the inconsistencies and errors in categorical data both within and between various sources, such as WALS and Grambank [38]. The authors introduce a novel seed dataset with continuous language data for five word order features calculated on the basis of Universal Dependencies [39].

2.3.3 Measures of Morphological Complexity and Word Order Flexibility

Having discussed the motivation behind using continuous typological linguistic properties, we can take a closer look at some of the existing approaches towards measuring the morphological complexity and word order flexibility of a language. These approaches range from simple calculations [40, 41] to applying information-theoretic principles [42, 43] to even using the accuracy of machine learning (ML) models on the task of predicting inflected word forms [44].

One of the differentiating morphological properties of a language is its *degree of synthesis*, which concerns the number of grammatical categories that can be expressed by a word [45]. *Analytic* languages such as Mandarin Chinese [46] tend to express grammatical information (e.g. tense, voice, agreement) using standalone units, while in *synthetic* languages like Ukrainian [47] those inflectional categories "attach" to a word as affixes. Bickel and Nichols (2013) measure the *inflectional synthesis* of verbs (also referred to as *counting complexity* [48]) by calculating their maximum category-per-word value (CPW). CPW indicates how many inflectional categories verbs of a particular language can have, and this measure is used, for example, in Shosted (2006) [49] to investigate the correlation between morphological and phonological complexity. While the author extracts the values from grammar descriptions, inflectional synthesis can also be measured using treebanks with annotated morphological structure [50].

In contrast, certain corpus-based measures do not require word structure annotation and instead estimate the morphological complexity of a language using simple statistical information about a text. For example, *type/token ratio* (TTR) measures the ratio of unique word types to word tokens in a text [40, 51]. A higher TTR value would indicate a larger number of unique word types, including word forms of the same lemma, thus pointing at higher morphological complexity. Despite their relative simplicity, TTR and its variations substantially correlate with other, more elaborate measures of morphological complexity [52] and have been used in works investigating the difficulty of language modeling [16] and neural machine translation [53].

Another quite intuitive way to estimate the morphological complexity of a language is to calculate its *mean size of paradigm* (MSP). In its simplest form, MSP is defined as the ratio of the number of word forms to the number of lemmas. It is expected that in languages with richer morphology the number of paradigm cells is higher, which is reflected by a higher MSP value.

Certain works formulate measuring morphological complexity as an information-theoretic problem [42, 43]. Juola (1998) [42], for example, speculates that texts written in more morphologically complex languages carry more information compared to languages with simpler morphology systems. In information theory, the average information of a random variable is expressed by *entropy* [54]. Given a discrete random variable X that follows from a probability distribution P with a probability density (or mass) function p_i, entropy is calculated as follows:

$$H(X) = -\sum_i p_i \log p_i \tag{4}$$

Following from the expression above, a higher degree of entropy denotes less certainty about the variable's outcomes and thus more information needed to account for them.

Juola (1998) compares the entropy of an original text with that of a distorted version with destroyed word structure. A higher difference between the two entropy values would suggest a higher degree of information expressed by the morphological tier. On a similar note, Bentz and Alikaniotis (2016) [43] propose measuring the entropy of the word frequency distribution of a text sample in order to determine the average information content of words. Languages with richer morphology would be characterized by a higher degree of word entropy and vice versa. According to Çöltekin and Rama (2023) [50], there are two morphology-related phenomena that affect the word frequency distribution and, consequently, word entropy:

- Morphologically "richer" languages tend to have more diverse and rare wordforms which leads to a longer tail of frequency distribution and, as a result, higher word entropy.
- Morphologically "poorer" languages will have a stronger reliance on highly common function words leading to more words with high likelihood and hence lower word entropy.

Information-theoretic approaches have also been used to estimate the word order variability of a language. Particularly, Levshina (2019) [55] measures the word order entropy of dependents and co-dependents using Universal Dependencies treebanks. To this end, she collects the frequencies of head and dependent elements as well as some co-dependent units (e.g. subject and object dependent on the same verbal predicate), uses their relative frequencies as probabilities, and calculates their entropies. The more dominant one pattern is over the other, the less entropy (i.e. uncertainty) there is in relation to the word order.

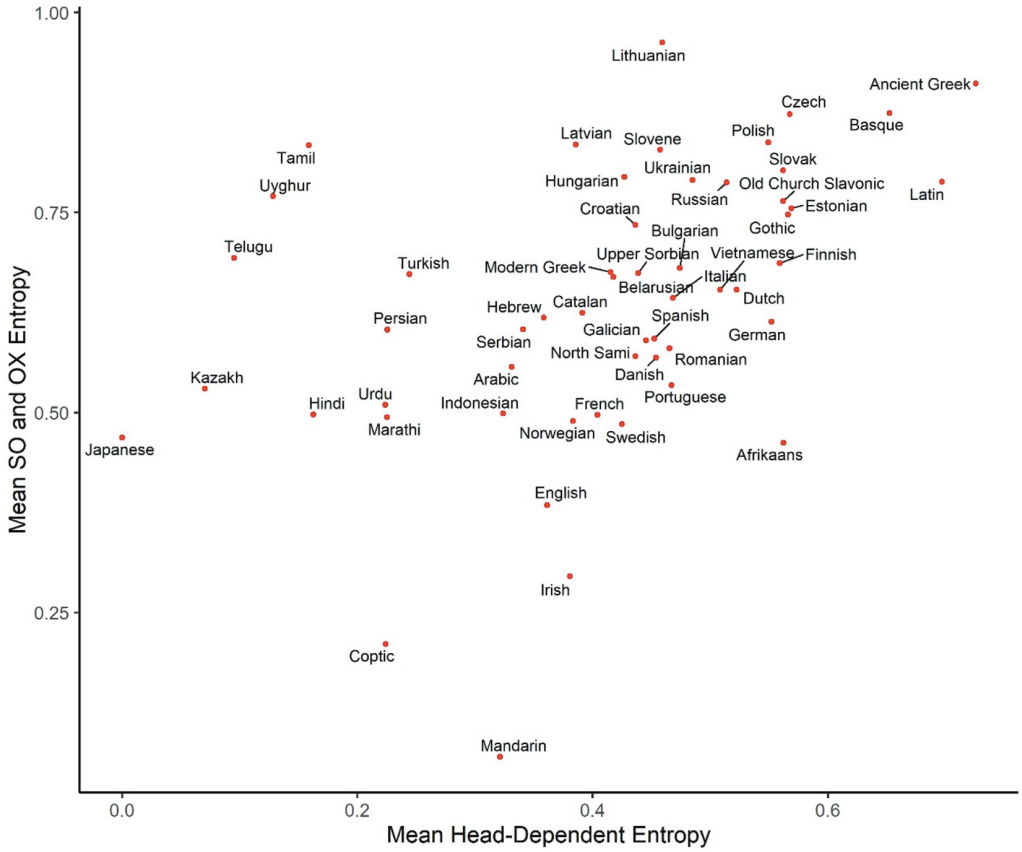

Figure 4: Average head-dependent and co-dependent word order entropy of languages in the UD corpora. Figure taken from Levshina (2019).

As part of her work, Levshina (2019) also plots the head-dependent and subject-object word order entropy. The relevant chart can be found in Figure 4. She makes the observation that many morphologically richer languages can be found to the right of the plot where head-dependent entropy values are higher, while the left part of the plot contains synthetic verb-final languages (e.g. Turkish and Japanese). In the vertical dimension, the languages located in the upper part have richer nominal morphology, suggesting that high subject-object entropy is associated with case marking. This supports the observation whereby languages with more flexible word order tend to have more case marking and vice versa [56, 57].

A novel way of estimating morphological complexity involves the use of ML models. Specifically, such models are tasked with predicting the correct inflected form of a word given its lemma and grammatical features; the accuracy achieved by the models then serves as an estimate of the morphological complexity of a language [50]. The models used for this task vary in their architectures, from simpler linear classifiers [50] to more sophisticated neural networks [58, 44].

In order to investigate the reliability of continuous morphological complexity variables, some works study their correspondence to human judgements such as those from the WALS database. Bentz et al. (2016) [52] compare four corpus-based morphological complexity measures with the quantified WALS feature values mentioned in §2.3.2. The measures are word entropy, relative entropy of word structure [59], type/token ratios, and a word alignment based measure. The authors find strong correlations between all the measures, implying that corpus-based measures can serve as reliable estimates of morphological complexity of a language. Similarly, Çöltekin and Rama (2023) [50] collect eight morphological complexity measures and find that they are related to WALS typological features, further advocating for their use when estimating morphological richness of a language.

2.4 Morphosyntactic Properties and Connectionist Modeling

Having looked at some of the existing approaches to measuring morphosyntactic variability of languages, it is crucial for our work to examine what kind of impact do languages with different typological properties have on NLP tasks. In this subsection we focus on connectionist modeling of languages with varying typological properties, while the next subsection is dedicated to language modeling and machine translation.

Although current NLP research aims to democratize access to language technologies across typologically diverse languages, some older works use connectionist simulations to draw insights about language learnability [60, 61, 62, 63]. Specifically, they speculate that language acquisition may be complicated not by linguistic universals, but rather by non-linguistic cognitive constraints associated with processing sequential structures [62].

Lupyan and Christiansen (2002), for instance, employ simple recurrent networks (SRNs) [64] to investigate the effect of word order and case marking on language learnability [62]. They argue that the prevalence of certain word order patterns (SOV, SVO, and VSO) (Greenberg 1963: Universal 1, also Table 2) is explained by the observation that they are easier to be acquired by a "general sequential-learning device". In order to test their predictions, the authors train ten SRNs with randomly initialized weights on 14 artificial grammars (six fixed word orders with and without case marking plus a flexible word order), hypothesizing that typologically rare languages will be harder for the SRNs to learn. For each of the 14 configurations, 3000 sentences were generated from a lexicon of 300 nouns and 100 verbs. The generated sentences varied from simple intransitives (e.g. *John walks*) to ditransitive

sentences with a subject, both direct and indirect objects, and genitive nouns (e.g. *Mary's friend gave Peter's key [to] John's brother*). Taking vector representations of words as input, the neural networks are tasked with correctly mapping the words to their grammatical categories (subject, direct object, indirect object, genitive noun, verb, or end-of-sentence).

Under all case-marked configurations, the trained networks correctly mapped all words to grammatical categories. Under the caseless configuration, the authors observe that the SRN performance roughly matches the frequencies of the word orders with a notable discrepancy: the results for caseless SVO and VSO orders are higher than for the caseless SOV order, even though it is the most common word order. Lupyan and Christiansen (2002) presume this is because most verb-final languages (such as SOV) rely on case-marking, since without it there is confusion between the subject and the object. In addition, they note that the performance under the case-marked free word order configuration is better than under some fixed word order settings without case marking, further confirming that flexible word order languages rely on case marking and vice versa.

Extending the work of Lupyan and Christiansen (2002), Everbroeck (2003) [63] observes the performance of neural network models when faced with various linguistic mechanisms of expressing grammatical relations in order to investigate a possible connection between language type frequency and learnability. The three primary mechanisms are word order, nominal case marking (referred to as D(ependent)-marking), and verbal agreement (referred to as H(ead)-marking). Similarly to Lupyan and Christiansen (2002), the author generates a set of artificial grammars with varying word order, its freedom, and presence/absence of D-marking and H-marking. However, in Everbroeck (2003) the lexicon used for sentence generation consists of artificially created words rather than words that exist in a particular natural language. The architecture of the SRN is also somewhat different, with larger input vector size and more output units to predict, such as the word's functional class and category. Everbroeck (2003) observes that the accuracy of the model's predictions often corresponds to the typological tendencies of certain morphosyntactic mechanisms. As just one example, the model performs well on sentences with fixed word order under all Head- and Dependent-marking configurations, except for the SOV without any markings, which corroborates the findings of Lupyan and Christiansen (2002) and supports the hypothesis that verb-final languages like SOV rely on case marking. However, Everbroeck (2003) points out that the results of the experiment do not provide an explanation as for why some language types are more common than others.

2.5 Morphosyntactic Properties and Language Modeling

The works described in §2.4 use neural networks as a simulation device to study linguistic factors influencing language learnability. Here we shift our focus to more recent studies concerned with the role of morphosyntactic properties in language modeling and machine translation (which can be thought of as a special case of LM).

Although it has been shown that some languages are more difficult for language models [8, 9], there is no clear consensus whether this is related to intrinsic language properties or rather surface-level factors. Generally, studying this phenomenon is complicated by a number of factors:

1. Cross-linguistic analysis of LM difficulty requires perfectly comparable (parallel) data in a multitude of languages, and such corpora are relatively scarce [11, 15] (notable examples include Europarl [65] and the Bible Corpus [66]). This requirement applies to both the data used for training the investigated LMs and the data used for experiments.

2. The term "difficulty" is inherently vague and should be expressed in terms of precise tasks and metrics, which there is a variety of (e.g. machine translation quality [13], grammatical feature prediction [67], and language modeling surprisal [11, 16]).

3. Measuring the difficulty of modeling may be obscured by a number of extralinguistic factors such as word segmentation [11, 16] and character encoding [14].

Despite these challenges, a considerable amount of work has been done in this direction. Cotterell et al. (2018) [14] perform cross-linguistic comparison of language models trained on Europarl (comprising 21 languages) and demonstrate that languages with complex inflectional morphology are harder to generate. Specifically, the authors train two open-vocabulary LMs: a baseline word/character n-gram model [68] and a character-level long short-term memory (LSTM) model [69]. The morphological complexity is estimated by means of *counting complexity* [48] (see §2.3.3), and the difficulty of language modeling is measured in *bits per English character* (BPEC). Cotterell et al. (2018) find that both the n-gram and the LSTM models perform worse on languages with rich inflectional morphology. Specifically, for the LSTM model, the researchers report high Spearman's rank correlation between a language's BPEC and morphological counting complexity. Additionally, in order to control for inflectional morphology, Cotterell et al. (2018) conduct the experiment with all words replaced

by lemmas and note that the correlation becomes slightly negative and insignificant. Although these results display a substantial relation between morphological richness and modeling difficulty, the authors emphasize that the origin of this difficulty is still unclear, whether it is inherent difficulty of a language or model-specific factors.

However, the researchers later extend their experiment in Mielke et al. (2019) [11] using a more typologically diverse set of 69 languages and fail to reproduce their previous findings. In fact, they show that simpler statistical factors seem important for LM difficulty as opposed to the inherent morphological language complexity. Mielke et al. (2019) train two open-vocabulary recurrent neural network (RNN) language models with character and byte-pair encoding (BPE) [70] segmentation respectively on the fully aligned segments from the Bible corpus [66]. The difficulty of language modeling is measured in terms of sentence surprisal aggregated for each language across the entire corpus. Mielke et al. (2019) also expand the set of typological language properties: aside from morphological counting complexity estimated in Cotterell et al. (2018), the authors include additional morphosyntactic features such as head-POS entropy [71], average dependency length [72], and WALS features 26A and 81A. The WALS features were chosen due to their availabitly for a large number of languages included in the study. However, none of the features significantly correlate with the estimated surprisal values; instead, general statistical properties like raw character sequence length and raw word inventory are found to be significant predictors of modeling difficulty. Wan (2021) [17] reaches similar conclusions experimenting with a six-layer Transformer LM [73] on six languages across five data sizes and three segmentation levels (character, byte, and word). The author argues for the language's representational granularity to be the culprit of modeling difficulty, with the language models influenced mainly by sequence length and vocabulary size.

A more recent work by Park et al. (2021) [16] attempts to resolve the disagreement between Cotterell et al. (2018) and Mielke et al. (2019) by increasing both the number of languages and morphological features studied. The researchers train five language models based on different segmentation methods on Bible translations in 92 languages and look for relations between modeling surprisal and various morphological properties, including all twelve WALS features in the area of morphology[10] and corpus-based measures such as TTR and mean length of words. The authors motivate their selection of morphological WALS features by the need to explore which features affect language modeling and to

[10] 20A, 21A, 21B, 22A, 23A, 24A, 25A, 25B, 26A, 27A, 28A, and 29A

what extent. The results show significant correlation of a number of morphological features with LM difficulty for the BPE-based model, while no significant association was found for the character-based model. In addition, the authors also test the models based on linguistically motivated segmentation methods (namely Morfessor [74] and Finite-State Transducers) and observe that they perform better and have lower surprisal values, suggesting that linguistically-grounded segmentation approaches may alleviate the influence of morphological complexity. However, concurrently to our thesis, Arnett and Bergen (2024) [12] published a new analysis showing that the role of morphologically-aligned tokenization in performance between agglutinative and fusional languages is insignificant. Instead, they find that this performance gap is most reduced when training on datasets of equivalent sizes scaled according to *"byte premium"* – a measure of how many bytes are required to represent the same text in different languages.

Disparities in views on the role of morphosyntactic properties in modeling difficulty are also found in the domain of neural machine translation. Belinkov et al. (2017) [13] experiment with an encoder-decoder NMT model to draw insights about what NMT models learn about inflectional morphology. Among other findings, they observe that translating into morphologically richer languages compared to English yields lower BLEU scores. On the other hand, Bugliarello et al. (2020) [9] argue that many MT quality evaluation metrics, including BLEU, are not cross-linguistically comparable. The authors propose an alternative measure of NMT difficulty – *cross-mutual information* (XMI), which makes it possible to decouple the difficulty of language generation from the inherent translation difficulty. Bugliarello et al. (2020) train 40 Transformer NMT models for translating each of the languages in Europarl from and into English. The findings show higher XMI values when translating *into* rather than *from* English, indicating that translation in this direction tends to be easier for the models. Additionally, the researchers look for correlations between translation difficulty and linguistic and data-related properties of languages. The only significant correlations they find involve the features related to type/token ratio for the source language and the distance between the source language and target language TTR values.

As was previously mentioned, conducting large-scale cross-linguistic studies on language modeling difficulty is complicated by the scarcity of typologically diverse parallel data. In light of this, some works propose using artificial language variants which differ from one another in terms of clearly defined properties [67, 15]. For instance, Ravfogel et al. (2019) [67] create synthetic versions of

English featuring different word orders, case systems, and agreement patterns, and use them to train multiple RNNs on the task of predicting subject and object agreement features of the verb. The authors observe that, among others, prediction accuracy was higher for the standard SVO version of English than for the SOV version, and that overt case marking made agreement prediction significantly easier no matter the word order. These findings further corroborate the observations made by Lupyan and Christiansen (2002) and Everbroeck (2003) (see §2.4).

Bisazza et al. (2021) [15] extend this research into the domain of neural machine translation and utilize a variety of synthetic languages to investigate whether languages with flexible word order and case marking are more difficult to translate by NMT models. Specifically, the authors compile a more controlled simple and small toy artificial grammar similar to Lupyan and Christiansen (2002) as well as a more realistic set of synthetic versions of English, inspired by Ravfogel et al. (2019). The artificial languages differ among each other by controlled parameters such as order of main constituents and case marking. The models trained to translate from these languages – a BiLSTM with Attention and a Transformer – overall perform worse on smaller amounts of data, particularly struggling more with flexible word order compared to fixed word order in the low- and medium-resource settings. In the mid-resource setup, case marking fails to improve translation quality over the fixed-order no-case language, and in the low-resource setting degrades it even compared to the language with random word order and absent case marking.

A logical next step for the research outlined in this section involves considerably expanding both the set of languages as well as language properties studied. The present work aims to accomplish exactly that, with the details described in the next section.

3 Methodology

Having looked into previous research, we now outline the methods employed in our work. We begin with describing NLLB-200, the state-of-the-art NMT model used in our translation experiments. After that, we provide a brief overview of FLORES+, an MT quality benchmark dataset which we use for language selection and translation source material. We then move on to the typological language data collected for our correlation studies. Finally, we delve into the measures we use for estimating MT difficulty, namely the metrics of translation quality and the probabilities of generated sequences.

3.1 Translation Model

While previous work mostly involves training separate NMT models from scratch, here we opt instead for a massively multilingual NLLB-200 pretrained NMT model. Despite the absence of control over the exact training data, this approach allows us to substantially increase the number of languages studied and to collect difficulty measures that are representative of state-of-the art MT.

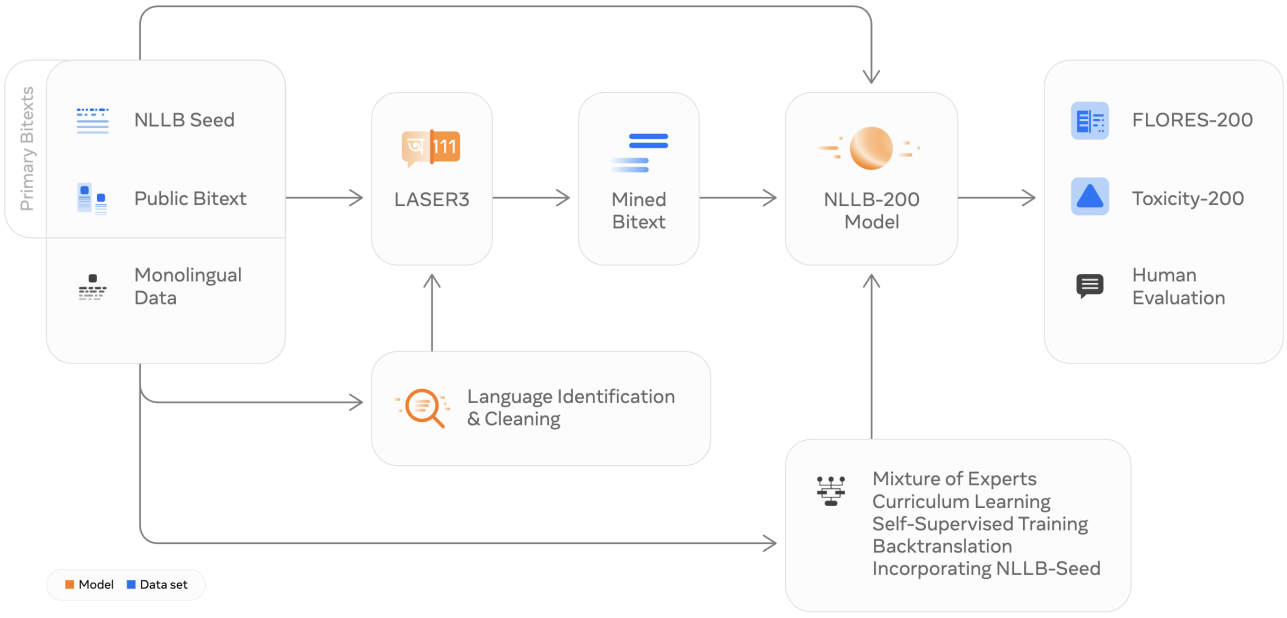

Figure 5: Components of the NLLB project. Figure taken from NLLB Team (2022) [18].

The NLLB-200 family of models was developed by Meta AI as part of their **N**o **L**anguage **L**eft **B**ehind effort to bring neural machine translation to a multitude of low-resource languages [18]. Each model in the family is capable of translating among 202 languages, resulting in over 40,000 translation

directions. The underlying architecture of NLLB-200 is based on a *Transformer encoder-decoder*: the encoder component transforms the input sequence in the source language into a series of token embeddings, while the decoder attends to those embeddings and generates the output sequence in the target language token by token. Comprising 54 billion parameters, the most sophisticated model of the family is a *Sparsely Gated Mixture-of-Experts* (MoE) model[11] [75, 76]. While standard *dense* models pass the input through all parameters, MoE models are capable of routing the input to a specific subnetwork of parameters (i.e. an *expert*), thus speeding up computation at inference time. Together with several regularization strategies (such as MoE expert output masking), this allows for reducing interference between unrelated language directions and avoiding overfitting for low-resource languages.

Due to its architecture, the MoE model weighs over 220 GB, which renders it unfeasible for our experiments. In light of this, we opt for the next largest model variant, `nllb-200-3.3B`[12]. This dense encoder-decoder Transformer model has 3.3 billion parameters, with 16 encoder and 16 decoder attention heads, as well as 24 encoder and 24 decoder layers.

The models were trained using a combination of sources and techniques[13]:

- **Primary bitext**: pre-existing pairs of translated sentences comprised of: 1) publicly available aligned textual data and 2) NLLB-Seed, a set of 6193 sentences professionally translated into 39 languages as part of the NLLB project.
- **Mined bitext**: obtained by collecting monolingual data and using LASER3 sentence encoders [77] and the stopes mining library [78] to find and collect likely translations.
- **Back-translated bitext**: monolingual data back-translated and compiled into parallel data using the stopes library.

The evaluation of the models' translation quality was performed using a combination of methods: a human-compiled benchmark called FLORES-200, human evaluation, and a novel toxicity benchmark[14]. All together, the components of the NLLB project are summarized in Figure 5.

[11] `https://huggingface.co/facebook/nllb-moe-54b`.
[12] `https://huggingface.co/facebook/nllb-200-3.3B`.
[13] `https://github.com/facebookresearch/fairseq/tree/nllb/examples/nllb/data`.
[14] `https://github.com/facebookresearch/flores/blob/main/toxicity/README.md`.

3.2 Translation Dataset

As pointed out in previous research (see §2.5), extensive cross-lingual investigation of the influence of typological features on LM and NMT difficulty requires large-scale multiparallel corpora. However, since we use an out-of-the-box NMT model for our experiments, we do not require a dataset large enough for training our own machine translation models from scratch. This allows us to take advantage of another source of multiparallel data – machine translation evaluation benchmarks. Namely, we use the FLORES+ evaluation benchmark dataset[15].

The introduction of the original **FLoRes** (**F**acebook **Lo**w **Res**ource) MT benchmark was motivated by the lack of evaluation datasets containing parallel data for low-resource languages [79]. Initially, the FLORES dataset contained sentences from Wikipedia translated into Nepali-English and Sinhala-English. Since then, a couple of expansions have been released[16]: **FLORES-101**, consisting of 3001 sentences translated into 101 languages [80] and **FLORES-200**, which was developed under the NLLB project (see §3.1) and increased the number of covered languages to 200 [18].

At the time of writing this thesis, **FLORES+** is the current updated version of FLORES-200. For this work, we use the `v2.0-rc.2` release of the dataset[17]. FLORES+ is composed of English sentences equally sampled from Wikinews, Wikijunior, and Wikivoyage and manually translated into over 200 languages. More specifically, the dataset is composed of two splits: the `dev` split contains 997 sentences in 212 languages and the `devtest` split has 1012 sentences for 205 languages. Each language is represented by a separate textual file containing the same set of sentences, and each sentence begins with a new line. We use the `dev` portion for our experiments.

3.3 Language Properties

Having determined FLORES+ as our choice of translation dataset, we begin collecting and calculating various properties of the languages covered by the dataset. While we largely take advantage of the existing measures (see §2.3), we make an effort of aggregating them from separate resources into a novel dataset, which we make publicly available as part of our contributions[18].

[15] `https://github.com/openlanguagedata/flores`.
[16] All three versions can be found at `https://github.com/facebookresearch/flores`.
[17] `https://github.com/openlanguagedata/flores/releases/tag/v2.0-rc.2`.
[18] `https://github.com/v-hirak/explaining-MT-difficulty`

3.3.1 Basic Taxonomic Properties

We first collect a set of basic identification and taxonomic features of the covered languages. Namely, we extract the languages' ISO-639-3 codes, scripts, and varieties from the FLORES+ description[19], while the WALS database codes, families, and genera are taken from the WALS database repository[20].

3.3.2 Precomputed Distances

Previous work on neural machine translation [53] and cross-lingual transfer [81] takes advantage of the URIEL typological database [82] containing vector representations of numerous languages drawn from typological, geographical, and phylogenetic databases. This approach allows us to take into consideration dataset-independent measures of language similarity based on linguistic study, in addition to dataset-dependent morphosyntactic features described later.

Using the accompanying `lang2vec` library[21], we query six types of precomputed distances between each FLORES+ language and English: *genetic*, *geographic*, *syntactic*, *inventory*, *phonological*, and *featural*. Genetic distance represents the distance between languages on the hypothesized Glottolog language tree [83], while geographic distance is calculated as the "great circle" distance between the languages on the surface of the Earth. Syntactic distance is the cosine distance between feature vectors derived from syntactic structures. Inventory and phonological distances are cosine distances between the phonological feature vectors derived from the PHOIBLE [84], WALS, and Ethnologue [85] databases. Featural distance is the cosine distance between feature vectors combining all five features described above.

All of the included distances from English are measured in the range between 0 and 1. For example, when we consider genetic distances, languages belonging to families distant from English will have values closer or equal to 1 (e.g. Arabic and Chinese), while languages of the Germanic family will have values closer to 0 (e.g. German and Danish). Similarly, in the context of geographic distance, languages which are predominantly spoken in areas closer to England will have values closer to 0 (e.g. Irish and Scottish Gaelic), while languages on distant continents will have values closer to 1 (e.g. Nepali and Samoan). If a certain language has syntactic, inventorial, phonological, or featural properties similar to English, its respective distances will be closer to 0, and vice versa.

[19] `https://github.com/openlanguagedata/flores?tab=readme-ov-file#language-coverage`.
[20] `https://github.com/cldf-datasets/wals`.
[21] `https://github.com/antonisa/lang2vec`.

In the context of this work, we expect languages with distance values closer to 1 to be more challenging for NMT and, consequently, to have lower translation quality scores. Moreover, we hypothesize that such languages will benefit more from increasing beam size, which will be reflected in higher gains in translation quality and, most importantly, generation probabilities.

3.3.3 WALS Features

Following the reasoning of Mielke et al. (2019) and Park et al. (2021), we include all 12 WALS features (20A-29A) related to morphology and one feature (81A) related to word order, as it is among the largest word order features in terms of language availability. The features are listed in Table 3.

ID	Name
20A	Fusion of Selected Inflectional Formatives
21A	Exponence of Selected Inflectional Formatives
21B	Exponence of Tense-Aspect-Mood Inflection
22A	Inflectional Synthesis of the Verb
23A	Locus of Marking in the Clause
24A	Locus of Marking in Possessive Noun Phrases
25A	Locus of Marking: Whole-language Typology
25B	Zero Marking of A and P Arguments
26A	Prefixing vs. Suffixing in Inflectional Morphology
27A	Reduplication
28A	Case Syncretism
29A	Syncretism in Verbal Person/Number Marking
81A	Order of Subject, Object and Verb

Table 3: WALS feature IDs and names.

Feature 20A describes how closely the inflectional marker is phonologically connected to its host. Features 21A and 21B are related to *exponence* – the number of grammatical categories that can be expressed by a single formative. Specifically, feature 21A measures the exponence of case markers, while feature 21B concerns Tense-Aspect-Mood markers. Feature 22A measures how many inflectional categories can be expressed by a verb (see §2.3.3). Features 23A through 25B all summarize where head/dependent marking occurs. Feature 26A estimates the preference of languages towards using prefixes or suffixes (and their proportions) in inflectional morphology (see §2.3.1 and Table 1).

Feature 27A identifies languages which use full or partial reduplication (i.e. repetition of a (part of a) word). Features 28A and 29A measure *syncretism*, which occurs when an inflectional form corresponds to multiple grammatical functions. Specifically, 28A measures syncretism of cases, while 29A measures syncretism related to verbal person/number marking. Finally, Feature 81A indicates the dominant word order of a language (see §2.3.1 and Table 2).

3.3.4 Precalculated Morphological Complexity Measures

Besides the discrete WALS morphological features outlined in §3.3.3, we make use of eight publicly available precalculated continuous measures of morphological complexity from Çöltekin and Rama (2023)[22]. The measures were computed on the basis of Universal Dependencies and are available for 33 languages of the FLORES+ dataset. Since the values for some languages were calculated on more than one treebank, we average those values to produce a single number for each language. Brief explanations for each measure are outlined below (see §2.3.3 for a more in-depth explanation).

Type/Token Ratio (TTR) In its simplest form, TTR is defined as the number of unique word types in a text divided by the total number of words in a text. This basic definition makes TTR dependent on the length of a text, so a number of modified versions exist to mitigate this factor. In particular, Çöltekin and Rama (2023) use the *moving average type/token ratio* (MATTR) [86], whereby the TTR values are calculated on fixed-length samples and later averaged across all such chunks. Since TTR lies in the theoretical 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) One way of estimating the morphological complexity of a language is to compare the information content (i.e. entropy) of the original text with its compressed version. 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. More morphologically complex languages would have higher word entropy, and vice versa. As UD datasets also include lemma annotations, Çöltekin and Rama (2023) additionally

[22] `https://github.com/coltekin/mcomplexity`.

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) As described in §2.3.3, MSP can be calculated by dividing the number of word forms in a text by the number of lemmas. Languages with richer inflectional morphology are expected to have a higher number of paradigm cells, reflected by the MSP value.

Inflectional Synthesis (IS) This feature refers to the maximum number of inflection categories that can be expressed by a standalone verb. Instead of using categorical values like in WALS, Çöltekin and Rama (2023) adopt a corpus-based approach and measure the maximum number of distinct inflectional features assigned to a lemma in a given sample.

Morphological Feature Entropy (MFH) Similarly to word and lemma entropy, MFH reflects the usage of morphological features and their values. For instance, languages with a higher number of approximately uniformly used grammatical cases will have higher entropy values, indicating a more intricate degree of inflectional morphology.

Inflection Accuracy (IA) Inflection accuracy is a relatively novel metric which refers to the accuracy of an ML model on the task of predicting inflected forms of words given their lemmas and grammatical features. Specifically, Çöltekin and Rama (2023) use linear classifiers. 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, the authors report negative inflection accuracy (`-ia`).

3.3.5 TTR Measured on FLORES+

In addition to the TTR values computed by Çöltekin and Rama (2023) on Universal Dependencies, we leverage the LexicalRichness Python module[23] [87] to calculate three TTR measures on the data for the languages covered by the `dev` split of FLORES+.

Type/Token Ratio (TTR) Basic form of TTR, defined as $TTR = t/w$, where t is the number of unique word types and w is the total number of words.

[23] `https://github.com/lsys/LexicalRichness`.

Root Type/Token Ratio (RTTR) Aimed at reducing the effect of text size [88], RTTR is similar to TTR, but the number of word types t is divided by the *square root* of the total number of words, as follows: $RTTR = t/\sqrt{w}$.

Moving Average Type/Token Ratio (MATTR) Described in §3.3.4, MATTR is calculated as an average of TTR values computed on fixed-length text chunks. Following Park et al. (2021), we use the window size of 500 word tokens.

3.3.6 Gradient Word Order Measures

Apart from WALS feature 81A, which categorizes languages into types depending on their word order preference, we also 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 Levshina (2019) calculates 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) on the basis of different corpora annotated under the Universal Dependencies approach. The author releases the calculated measures for individual types of syntactic relations, which we average to obtain a single entropy value for dependencies and codependencies for each of the available languages.

Proportion of Subject-Object Order Levshina et al. (2023) use online news corpora from the Leipzig Corpora Collection [89] annotated with Universal Dependencies to collect the frequencies of phrases where subject comes before object and vice versa. The authors then calculate the proportions of these orders: 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. In our work, we use the proportion of the Subject-Object (SO) order specifically, meaning that subject primarily comes before object in languages where this value approaches 1, indicating less flexibility in order. Thus, to ensure consistency with the rest of the continuous properties, we take the negative of this measure, such that higher values indicate *more* word order flexibility. We then expect translation to be *easier* for languages with strong preference for the SO order and *harder* for languages with less preference towards a particular order.

Percentage of Head-Final Phrases Also as part of their work, Levshina et al. (2023) count the frequencies of head-initial and head-final phrases and measure their percentages on the basis of 123 corpora annotated with surface-syntactic Universal Dependencies [90]. We use the fraction of phrases that are head-final and average these values for those languages which have multiple treebanks. We expect *lower* translation difficulty for languages with a larger percentage of head-final phrases, since this would imply less uncertainty about the location of the head in a phrase. Similarly to Subject-Object proportion, we take the negative of this value for consistency across measures.

3.3.7 Estimating Training Data Size by Language

One drawback of using a pretrained NMT model such as NLLB-200 is the lack of precise control over the data used during training. Further complicating matters is the fact that the authors of the NLLB-200 models do not provide exact numbers about the training data distribution across the 200 languages. Since training data size may serve as a potential confounding factor for correlations between language properties and MT difficulty, we attempted to approximate the training data language distribution from the partially available resources used for training NLLB-200. In particular, we counted the number of sentences in the *primary bitext* portion of the training data and the *mined bitext* metadata (see §3.1). There is no information, however, on what data was used for producing the *back-translated bitext*.

During our experiments, the estimated training size data numbers resulted in weak and mostly insignificant correlations. In light of this, we opted for Wikipedia size by language instead as a more reliable proxy for training data language distribution[24]. The sizes are represented by the total number of articles in the Wikipedia of a given language and are available for 165 languages of the FLORES+ dataset. The resulting values led to stronger and more significant correlations. Appendix B illustrates the Wikipedia sizes available for 105 out of 124 languages that we translate into.

3.4 Translation Quality Metrics

Having covered the language properties included in our experiments, we now outline the measures we use to estimate the difficulty of machine translation. This subsection is dedicated to MT quality evaluation metrics, and the next subsection explores the use of probabilities of generated sequences.

[24] `https://meta.wikimedia.org/wiki/List_of_Wikipedias`.

In order to evaluate the quality of the translations produced by NLLB-200, we opt for three of the most popular MT quality evaluation metrics: BLEU, chrF++, and COMET.

BLEU The **Bi**-**L**ingual **E**valuation **U**nderstudy [91] metric looks at n-gram overlap between the hypothesis translation and the reference translation, with the higher BLEU score indicating better translation quality. We use the sacreBLEU implementation for calculating the BLEU scores [92].

chrF++ **Cha**racter n-gram **F**-score (chrF) [93] is based on character-level n-gram overlap between the hypothesis and reference translations, and is better suited for languages with rich morphology. To demonstrate this, consider an example in Finnish, a morphologically rich language, where the reference translation is *Hän* ***kävelee*** *metsään* (*He* ***walks*** *into the forest*), and the generated hypothesis translation is *Hän* ***käveli*** *metsään* (*He* ***walked*** *into the forest*). chrF would assign partial credit for the correct root *kävel-*, while a metric like BLEU would penalize this translation heavily. chrF++ is a revised version of chrF which additionally takes into account word n-grams. As with BLEU, we use the sacreBLEU implementation.

COMET **C**rosslingual **O**ptimized **M**etric for **E**valuation of **T**ranslation [94] leverages pretrained multilingual neural models to evaluate translation quality based on deeper linguistic features rather than surface-level word or character overlap. Although COMET achieves state-of-the-art levels of correlation with human judgements for many languages, the metric is not as interpretable as BLEU and chrF++ and is less robust for low-resource languages. In our experiments, we use the official implementation by Unbabel[25].

3.5 Translation Model Probabilities

In essence, the goal of MT systems is to produce a translated sequence with the highest probability, which in turn is calculated as the product of the probabilities of individual tokens that the model chooses at each generation step from a learned vocabulary distribution. As seen in §2.1, correctly translating into languages with richer morphology and more word order freedom requires choosing among a higher number of potential candidates; the choice of tokens becomes less certain, with the tokens themselves having lower probabilities (e.g. due to several potential word orders or inflected forms).

[25] `https://github.com/Unbabel/COMET`.

In light of this, we aim to explore how output sequence probabilities change with increasing beam size depending on the morphosyntactic properties of a language. To accomplish this, we take advantage of the functionality offered by the Hugging Face transformers Python library [95] which outputs the probabilities for the tokens predicted at each time step during generation of the translations. We then use these numbers to calculate the probabilities of each sequence translated from English into a target language under each of the four beam sizes $k \in \{1,3,5,7\}$. Then, we average the sequence probabilities to obtain the overall translation probability for a given language at a specific beam width. Finally, we calculate a language's gain in probability across the four beam sizes.

It is important to point out that the decision to include the model probabilities as a measure of translation difficulty was made at a later stage. Given that the probability data can only be extracted during sequence generation, we needed to perform another round of translation. Thus, in light of the associated computational cost, we chose to generate sequence probability data for a smaller subset of 52 languages, which can be found in Appendix A.

4 Experimental Setup

Following the overview of the methods used in our work, we now describe how they are employed in our experiments. As discussed in §2.2, increasing beam size generally improves translation quality by allowing a model to explore more candidate sequences, which is particularly important for languages typologically different from English. It is not clear, however, how this improvement varies depending on the morphosyntactic properties of a language. The present work explores this novel research question in the experiments outlined in this section.

We begin with translating a set of English sentences into a selection of 124 languages under four beam size configurations (§4.1). While using English as the only source language is a limiting factor, especially given the dominance of English in NLP research, this choice offers several advantages. Firstly, English is considered a morphologically poor language, which makes it a very suitable source language for gauging the impact of morphological richness of target languages on translation difficulty. Moreover, English often serves as an actual source language for parallel data used for training the NLLB-200 model. With that being said, we argue that future work should expand the selection of source languages to ensure broader linguistic coverage.

After we perform the translations, we assess their difficulty via a combination of MT quality metrics and output sequence probabilities (§4.2). Finally, in §4.3 we look for relationships between these measures and the language properties outlined in §3.3.

4.1 Creating Translations

Source Data As the source translation material we use 997 English sentences from the `dev` split of FLORES+. Since each sentence is stored on a seperate line, we simply split the textual file into lines and pass the sentences incrementally to the NLLB-200 model (more specifically, the `nllb-200-3.3B` variant, see §3.1).

Translation Directions While FLORES+ covers over 200 languages, a number of linguistic properties described in §3.3 are only available for smaller subsets of the dataset languages. Furthermore, some languages included in FLORES+ are not supported by the NLLB-200 models. If we were only to include the languages for which all of our language properties are available (the "intersection" of all properties), we would end up with just 22 languages. In light of this, we relax our requirements

and include all languages for which at least one precomputed morphological (Çöltekin and Rama, 2023) or word order (Levshina, 2019 or Levshina et al., 2023) property is available. As a result, we end up with a selection of 124 languages to translate the English sentences into[26]. Target language is specified as a beginning-of-sequence (BOS) token.

Beam Size Configurations For our experiments, we translate the English sentences from FLORES+ into each of the selected 124 languages under four beam sizes $k \in \{1,3,5,7\}$.

4.2 Estimating Translation Difficulty

Following the setup outlined in §4.1, we generate translations for the 997 FLORES+ English sentences in 124 languages under four beam widths. Here we describe the process of estimating the (change of) translation difficulty relative to beam size.

4.2.1 Translation Quality

Raw Scores To estimate the quality of the generated translations, we use three common metrics: BLEU, chrF++, and COMET (see §3.4). For each of the 124 target languages, we score four sets of 997 translated sentences corresponding to each of the four beam sizes. As reference translations, we use the sentences from FLORES+ in the corresponding languages.

Score Gains Crucially for our research question, we estimate the improvement in translation quality by measuring the difference in quality scores between beam sizes $k = 1$ and $k = 7$. We then express this number both as *absolute* gain (e.g. 5 chrF++ points) and *relative* gain (i.e. as a percentage of the chrF++ score at beam size $k = 1$). For example, if for a given language the chrF++ score at beam size $k = 1$ is 25 points, and at beam size $k = 7$ it is 30 points, then the absolute gain is $30 - 25 = 5$ points, and the relative gain is $\frac{5}{25} \times 100 = 20\%$. We calculate gain per beam size, absolute gain, and relative gain for BLEU, chrF++, and COMET for each target language, and use all of these values to find correlations with language properties.

[26] See Appendix A for the full list of target languages.

4.2.2 Translation Probabilities

Following the procedure outlined in §3.5, we calculate translation generation probabilities for each target language and beam width setting. After that, much like with the quality scores, we measure the absolute gain and relative gain in probability. We use these metrics to test our hypothesis that morphosyntactically complex languages might benefit more from expanding beam width, which will be reflected in a larger increase in modeling probability. In addition, studying the change of generation probabilities allows us to disentangle intrinsic modeling difficulty from the effects of using a particular translation evaluation metric.

4.3 Correlation Studies

Having aggregated both data about language properties as well as translation difficulty, we perform a set of statistical tests in order to investigate a potential link between these phenomena. With these tests, we attempt to uncover which language properties (both categorical and continuous) serve as predictors of translation difficulty and, more importantly, how these properties affect the extent to which translating into a given language benefits from expanding beam width. Due to the nature of the typological data (discrete vs. continuous), our experiments are split into two sets: testing for association with categorical WALS feature values (§4.3.1) and estimating correlation with continuous morphosyntactic properties (§4.3.2).

4.3.1 WALS Features

With this set of experiments, we intend to find out whether languages belonging to different typological WALS categories will have significantly different translation quality scores and, more importantly, translation quality gains. For instance, do languages with a strong preference for suffixing see more improvement in translation quality at higher beam sizes than languages preferring prefixing? What about languages with no inflectional morphology at all?

In order to investigate these relations, we follow the methodology of Park et al. (2021). First, we measure the sample size (i.e. number of languages) of each value of each WALS feature of interest (see Table 3 in §3.3.3) from the pool of 124 target languages. To ensure the effectiveness of the statistical test, we filter out values with fewer than five languages; since this results in some features being left with only one value, we remove those features from our experiments. In the end, we are left

with 11 WALS features. The final selection of features, their values, and the corresponding sample sizes can be found in Appendix D.

Since the model probability data is only available for 52 languages, we opt for testing associations with translation quality information available for all 124 target languages to ensure the robustness of our findings. We perform the Kruskal-Wallis test (one-way ANOVA on ranks) and determine the features where there is a significant difference in quality scores among the feature's values ($p < 0.05$). For these features, we additionally perform the Dunn's post-hoc analysis with Benjamini-Hochberg p-value correction [96] to determine the exact categories which were significantly different. We report our findings in §5.1.

4.3.2 Continuous Measures

In addition to translation quality association with WALS categorical values, we perform correlation tests for the language properties measured on a continuous scale[27]. Once again, our goal here is to determine whether certain continuous morphosyntactic properties can serve as predictors of translation difficulty, and whether these properties influence the degree to which translating into a given language benefits from increasing beam size.

In contrast to the experiments in 4.3.1, we also include modeling probability data alongside translation quality metrics as estimates of translation difficulty. Since the values of the continuous language properties are not normally distributed, we opt for Spearman's rank correlation as our choice of statistical test. We report our findings in §5.2.

[27] See Appendix C for the overview of continuous language properties.

5 Results

In this section, we analyze the results of our experiments outlined in §4. To recap, our primary research question is whether target languages with higher morphological complexity and/or word order flexibility benefit more from expanding beam width, which is reflected by larger gains in translation quality or modeling probability. Additionally, we aim to discover which language properties serve as predictors of translation quality in general. We investigate the effect of categorical morphosyntactic features in §5.1 and the effect of continuous typological properties in §5.2.

5.1 WALS Features

These experiments are mainly inspired by the findings of Park et al. (2021), which are based on the same set of morphological WALS features, but involve a different NLP task (LM) and a smaller, more controlled experimental setup (a multiparallel corpus of Bible translations). In turn, we are interested in the role of these same features in the difficulty of state-of-the-art NMT.

5.1.1 Translation Quality Scores

We begin our experiments by investigating which typological language categories have a significant impact on the quality of NMT. Table 4 summarizes the WALS features where the difference in translation quality scores across the feature's values is significant. For the sake of brevity, we report the numbers for the translations at beam size $k = 5$.

As can be seen from the table, a total of seven WALS features were found to have significant differences in quality scores among their values. For the features with significant impact on translation quality, we additionally perform post-hoc analysis to identify the exact feature values where the quality scores differ significantly. We report our findings below for each quality metric.

BLEU For feature **26A**, the median score for predominantly suffixing languages is significantly higher than for predominantly prefixing languages and for languages with little or no inflectional morphology. This is partially in line with the findings of Mielke et al. (2019), who showed that the average language modeling surprisal for strongly suffixing languages is slightly smaller than for strongly prefixing languages. However, they showed this difference to be insignificant. For feature **27A**, the median BLEU scores for languages with productive full and partial reduplication and with

Metric	Feature	Sample Size	*H*-statistic	*p*-value	η^2
BLEU	26A: Prefixing vs. Suffixing	99	14.6	0.006	0.11
	27A: Reduplication	72	18.8	<0.001	**0.24**
	28A: Case Syncretism	40	7.39	0.02	**0.15**
	81A: Order of Subj, Verb, Obj	103	8.06	0.04	0.05
chrF++	21A: Exponence	31	5.54	0.02	**0.16**
	27A: Reduplication	72	18.0	<0.001	**0.23**
	28A: Case Syncretism	40	9.8	0.007	**0.21**
	29A: Person/Number Syncretism	42	9.31	0.01	**0.19**
COMET	23A: Locus of Marking in Clause	43	8.85	0.03	**0.15**
	26A: Prefixing vs. Suffixing	99	26.2	<0.001	**0.24**
	27A: Reduplication	72	12.9	0.002	**0.16**
	28A: Case Syncretism	40	6.87	0.03	0.13

Table 4: WALS features with significant effect on translation quality scores across three metrics. Results are reported for beam size $k = 5$. Sample size is measured in number of languages. Large effect sizes ($\eta^2 \geq 0.14$) are marked **in bold**. See Appendix D for overview of features and their values.

full reduplication only are significantly lower than for languages with no reduplication. For **28A**, the median score for languages where inflectional case marking is syncretic is significantly higher than for languages with no case syncretism. Finally, for feature **81A**, languages with SOV order are shown to have a significantly lower median BLEU score than languages with SVO order.

chrF++ For feature **21A**, languages with monoexponential case have significantly lower chrF++ scores than languages without the case system, which corroborates the BPE-level LM results of Park et al. (2021) when training on a multiparallel corpus (the Bible). The findings in the context of **27A** match those of BLEU, whereby languages with no productive reduplication have significantly higher scores. The results for feature **28A** demonstrate that languages with case syncretism have a significantly higher median score than languages with minimal or absent case marking. Similarly, for **29A**, languages without subject person/number marking score significantly lower than languages where this type of marking is present and syncretic, which is consistent with Park et al. (2021).

COMET Despite the significant outcome of the Kruskal-Wallis test for **23A**, the post-hoc analysis failed to reveal any significant differences among the feature's values in pairwise comparisons. For feature **26A**, predominantly suffixing languages score significantly higher than languages with

minimal inflectional morphology, languages which prefer prefixing and suffixing equally, as well as predominantly prefixing languages. Additionally, languages with moderate preference for suffixing also have a higher median score than primarily prefixing languages. In the context of **27A**, languages with productive full and partial reduplication have a significantly lower median score compared to languages with no productive reduplication. Lastly, concerning feature **28A**, languages with minimal or absent case marking have a significantly lower median score than languages with case syncretism.

5.1.2 Translation Quality Gains

When examining the increase in translation quality across languages as beam size expands, the Kruskal-Wallis test identifies a total of only three WALS features for which the differences in quality gain across languages of different categories are statistically significant. We report the results for absolute translation quality gain from beam size $k = 1$ to $k = 7$ in Table 5.

Metric	Feature	Sample Size	H-statistic	p-value	η^2
BLEU	26A: Prefixing vs. Suffixing	99	19.3	<0.001	**0.16**
	27A: Reduplication	72	9.29	0.01	0.11
	28A: Case Syncretism	40	8.66	0.01	**0.18**
	81A: Order of Subj, Verb, Obj	103	1.68	0.64	-0.01
chrF++	21A: Exponence	31	0.62	0.43	-0.01
	26A: Prefixing vs. Suffixing	99	12.4	0.01	0.09
	27A: Reduplication	72	5.21	0.07	0.05
	28A: Case Syncretism	40	8.99	0.01	**0.19**
	29A: Person/Number Syncretism	42	2.19	0.33	0.005
COMET	23A: Locus of Marking in Clause	43	7.4	0.06	0.11
	26A: Prefixing vs. Suffixing	99	16.7	0.002	**0.14**
	27A: Reduplication	72	7.48	0.02	0.08
	28A: Case Syncretism	40	5.94	0.05	0.11

Table 5: WALS features with significant effect on translation quality improvement across three metrics. Results are reported for absolute gain in quality scores between $k = 1$ and $k = 7$. For completeness, we also include features which had significant effect on translation quality scores (Table 4), but insignificant effect on quality gains (marked in grey). Sample size is measured in number of languages. Large effect sizes ($\eta^2 \geq 0.14$) are marked **in bold**. See Appendix D for overview of features and their values.

Similarly to §5.1.1, we use the Dunn's test to determine the feature values whose translation quality gains differ significantly. The results described below for each feature apply to all metrics where that feature is significant.

For feature **26A**, the median translation quality change is significantly higher for predominantly suffixing languages than for languages with little or no inflectional morphology. For **27A**, languages with both types of productive reduplication see a significantly lower median gain in quality compared to languages lacking productive reduplication. Lastly, in terms of feature **28A**, languages with absent or minimal inflectional case marking have a significantly lower median quality gain than languages with case syncretism.

5.1.3 Impact of Wikipedia Size

Overall, the results of our initial experiments provide partial evidence that morphological complexity and word order typology of target languages have a significant effect on the (gain in) quality of NMT. However, while these findings fall in line with our intuitions as well as previous works, upon further investigation, we have noted that some of the feature values that had significantly lower/higher translation quality scores and gains also had significantly lower/higher numbers of Wikipedia articles (referred to as Wikipedia size). This implies that another factor impacting the differences in quality scores and quality gains may be the degree of language representation in the NLLB-200 training data approximated via the number of Wikipedia articles.

In order to minimize the impact of this potential confounder, for each significant feature from Tables 4 and 5, we focused on the groups of languages with significantly different translation quality measures according to the Dunn post-hoc tests. We then downsampled one group to match the Wikipedia sizes of the other group using nearest neighbor matching. Finally, we used a Mann-Whitney *U* or a Kruskal-Wallis test (depending on the number of groups) to determine whether the significant difference in translation quality measures still holds after matching Wikipedia sizes.

Translation Quality Scores Of all the significant features in Table 4, after matching languages by Wikipedia size, only feature 27A (Reduplication) still had a significant effect on BLEU scores. For chrF++ scores, all four features (21A, 27A, 28A, and 29A) remained significant, while for COMET none of the features significantly impacted the scores anymore after matching Wikipedia sizes.

Translation Quality Gains After the matching procedure, the only feature that still had significant effect on the gain in translation quality was feature 26A (Preference for Suffixing vs. Prefixing vs. No Inflectional Morphology).

5.1.4 Analysis of Results

The initial set of experiments revealed that of the 11 WALS features related to morphology, a total of seven features had significant impact on translation quality *scores*, and only a total of three features had a significant effect on translation quality *gain*. Matching the languages by their Wikipedia size narrowed down the significant features even further: just four features were significant for quality *scores*, and only one remained significant for quality *gains*. Below we highlight our key findings.

Disparity among Metrics In the context of quality scores, it is worth noting the disparity in the number of significant features among the three metrics after matching languages based on their Wikipedia sizes. In the end, four features emerged as significant contributors to the chrF++ scores, while only one feature remained significant for the BLEU scores. Notably, none of the features had a significant impact on the COMET scores. While BLEU relies on exact word n-gram matches and COMET focuses on semantic similarity, chrF++ is more suitable for morphologically rich languages due to a finer granularity[28]. Coupled with the greater number of significant features for chrF++, these observations suggest that chrF++ may better capture the morphological richness of target languages and is more sensitive to the differences in morphological complexity.

Exponence and Reduplication The findings for features 21A and 27A both highlight that inflectional morphology presents a challenge for neural machine translation. Specifically, in the context of 21A, translating into languages with monoexponential case yields significantly lower chrF++ scores than the languages with no case marking, which is in line with the findings of Park et al. (2021). Similarly, in the context of 27A, languages with productive reduplication have significantly lower translation quality measured in BLEU and chrF++.

Syncretism Our findings for both features concerning syncretism – 28A and 29A – are similar to those of Park et al. (2021). Specifically, languages with syncretism are characterized with higher chrF++ scores compared to languages with no syncretism. While it is possible to assume that syncretism leads to a smaller number of inflected forms and, consequently, a smaller vocabulary for a model to learn (compared to the languages with no syncretism, where each grammatical category value has a unique inflectional form), this assumption does not explain why languages with syncretism – a marker of inflectional morphology – also obtain better translation quality scores than the

[28] See §3.4 for an illustrative example.

languages with *no* inflectional case and/or person/number marking.

Prefixing vs. Suffixing As we shift our attention towards translation quality *gains* (which have the advantage of being less subject to confounders like training data size), feature 26A is the only feature with significant influence on the total gain in BLEU and chrF++ points when increasing beam size from $k = 1$ to $k = 7$. Namely, languages with strong preference for suffixing see a higher gain in translation quality than those with strong preference for prefixing and, more notably, with little or no inflectional morphology at all. This finding reveals that languages with a strong presence of inflectional morphology (suffixing in particular) benefit more from a larger search space than languages with minimal inflectional morphology.

Overall, the results of our experiments with morphological WALS features provide inconclusive evidence regarding the impact of inflectional morphology and word order on the quality of NMT. While several features (21A, 26A, and 27A) provide strong support for our hypothesis, most of the features either do not have a significant effect on the translation quality scores and gains or contradict our assumptions to a certain degree (as is the case with 28A and 29A).

5.2 Continuous Measures

Having discussed the results of translation quality associations with various WALS features, we now switch to investigating the relationships between continuous language properties, beam width, translation quality, and generation probability change. As discussed in §2.3.2, continuous measures have been argued to better account for typological properties compared to categorical approaches such as those found in WALS [36, 37].

According to our hypotheses:

- We expect *negative* correlations between (i) translation quality scores and (ii) continuous measures of morphological complexity and word order freedom, which would suggest that such languages are harder to translate into (at least when the source language is English).

- We expect *positive* correlations between (i) the gain in translation quality and generation probability, and (ii) morphological complexity and word order freedom. This would imply that the languages with more complex typological features and higher freedom of word order benefit

more from increasing beam size compared to languages with simple morphology and more fixed order.

The findings of this set of experiments will help us understand more clearly whether languages with more complex morphology and word order require alternative decoding strategies to yield optimal generation results.

5.2.1 Translation Quality Scores

We start with the Spearman correlations between continuous language properties and the scores for translation quality metrics. As the correlations practically do not vary across the four beam sizes, we choose to report our results for beam size $k = 5$. In addition, while the results for BLEU and chrF++ are fairly similar and align with our expectations, many correlations with COMET scores contradict our intuitions. For instance, a number of continuous morphological complexity measures exhibit positive correlations, which would suggest that target languages with more complex morphosyntactic phenomena obtain higher COMET scores. However, as mentioned in §3.4, due to its nature, COMET is less reliable for low-resource languages, which constitute a substantial portion of our experiments. In light of this, we choose to focus instead on the results for the more interpretable BLEU and chrF++ metrics for our discussion. However, we still include the correlations with COMET in Appendix E and Appendix F.

Figure 6 summarizes correlation results for BLEU and chrF++. As can be seen from the figure, most of the continuous measures of typological complexity exhibit negative correlations with both BLEU and chrF++ scores, in line with our expectations. However, only nine of those correlations are significant for both metrics. Below we highlight our findings for each group of properties.

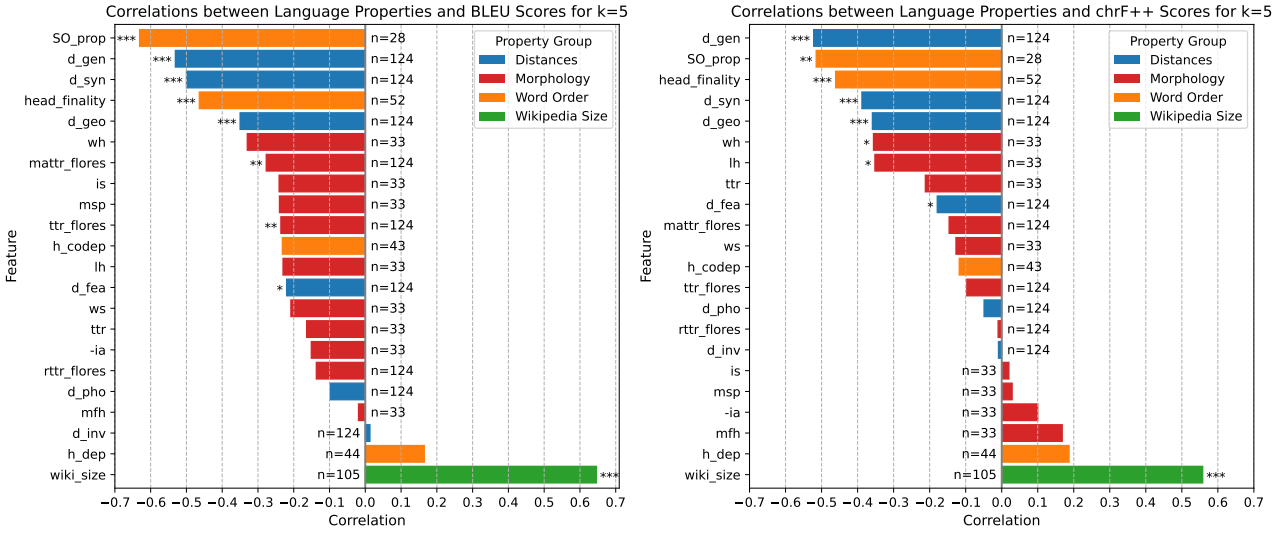

Figure 6: Spearman correlations between continuous language properties and translation quality scores at beam size $k = 5$ for BLEU (left) and chrF++ (right). Properties are categorized into four groups (see Appendix C for breakdown). Sample sizes for each property are indicated next to their respective bars. Correlations significant at $p < 0.05$ are marked with "*", at $p < 0.01$ – with "**", at $p < 0.001$ – with "***".

Wikipedia Size Our Spearman correlation tests for both BLEU and chrF++ reveal a highly significant moderate to strong link between the number of Wikipedia articles (`wiki_size`) and translation quality scores. This points at a significant effect of language representation in training data on the performance of the NMT model, which is consistent with the observations of Bisazza et al. (2021) as well as our findings for WALS features in §5.1. In an effort to reduce the effect of Wikipedia size, we calculated the correlations for a number of significant features on more balanced subsets of languages. The subsets are made up of languages whose Wikipedia sizes lie in the interquartile ranges (i.e. middle 50%) of the original samples, which helped us avoid languages with too many or too few Wikipedia articles while still providing reasonable sample sizes for each language property in a transparent and standardized manner.

Typological Distances from English Figure 6 showcases highly significant weak to moderate correlations with *genetic* (`d_gen`), *geographic* (`d_geo`), and *syntactic* (`d_syn`) distances from English, which is in line with our hypothesis that the languages typologically far from English are harder to translate into. Figure 7 provides a closer look into how these correlations change when measured on a more balanced subset in terms of the number of Wikipedia articles. Overall, the correlations with genetic distance and Wikipedia size get slightly weaker, while the correlations with geographic

distance become somewhat stronger. Syntactic distance correlations stay practically the same.

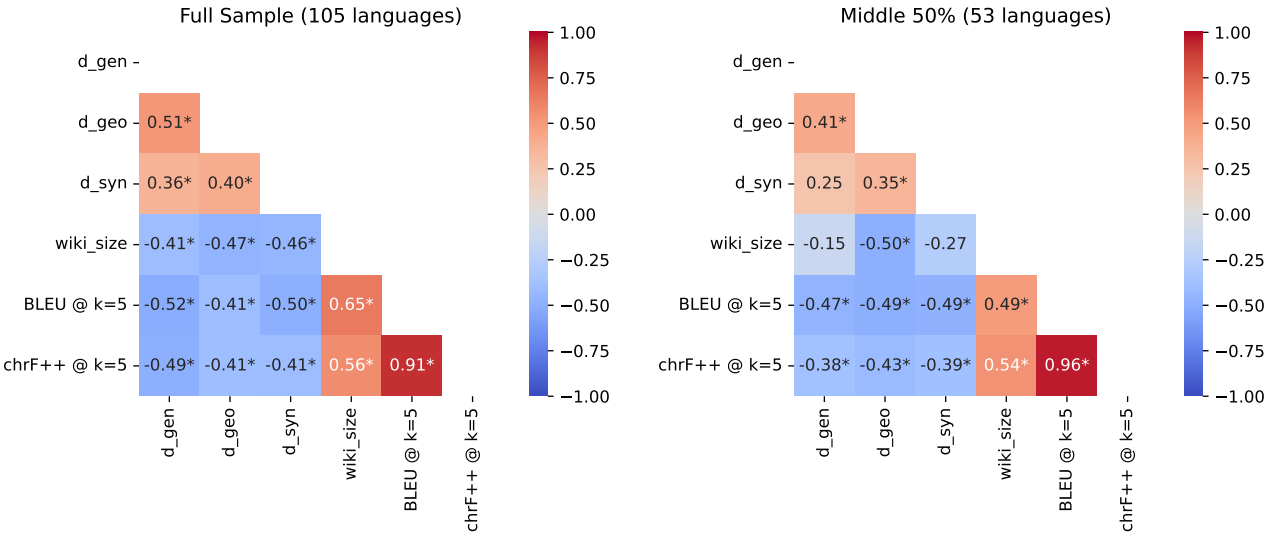

Figure 7: Spearman correlations between typological distances, Wikipedia size, and BLEU/chrF++ scores at beam size 5. Correlations are measured on the full language sample (left) and a more balanced interquartile range subset (right).

Morphological Complexity Out of all continuous morphological complexity measures employed in our study, only *type/token ratio* (`ttr_flores`) and *moving average type/token ratio* (`mattr_flores`) measured on FLORES+ have very significant weakly negative correlations with BLEU scores at $k = 5$, and only *word entropy* (`wh`) and *lemma entropy* (`lh`) display significant weakly negative correlations with chrF++ scores at $k = 5$. While the correlation coefficients for both TTR measures get stronger for BLEU with more balanced Wikipedia sizes (Figure 8), the correlations between word and lemma entropies and chrF++ scores lose significance on a more balanced dataset.

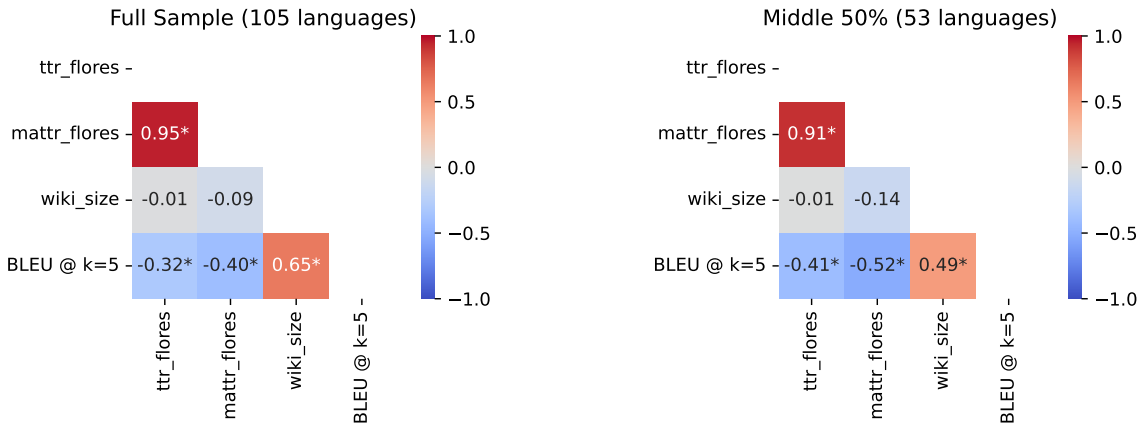

Figure 8: Spearman correlations between TTR and MATTR measured on FLORES+, Wikipedia size, and BLEU scores at beam size 5. Correlations are measured on the full language sample (left) and a more balanced interquartile range subset (right).

The results of the correlations with the two TTR measures are in line with the observations of Mielke et al. (2019) and Park et al. (2021): both those works find that TTR (and MATTR in Park et al., 2021) is highly correlated with LM surprisal in BPE models on the multiparallel Bible corpus. At the same time, while our correlations are significant for TTR of target languages specifically, Bugliarello et al. (2020) only report significant correlations for TTR of source languages (and TTR difference between source and target languages).

Word Order Freedom When looking at Figure 6, we can observe that the negative correlations with two measures of word order freedom – *proportion of Subject-Object order* (`SO_prop`) and *percentage of head-final phrases* (`head_finality`) – are among the strongest and most significant. Recall from §3.3.6 that we invert the sign of the original measures from Levshina et al. (2023), such that *higher* values indicate *less* certainty about the order of subject and object/head and dependent. Thus, our correlations suggest that more flexibility in the order of main constituents poses a bigger challenge for NMT. In fact, as seen in Figures 9 and 10, these correlations get even stronger when measured on more balanced subsets, while the correlations with Wikipedia size get weaker and become insignificant. Our observations are similar to Bisazza et al. (2021), who find that, in low- and mid-resource settings, synthetic free-order languages require more data for accurate NMT than their fixed-order counterparts.

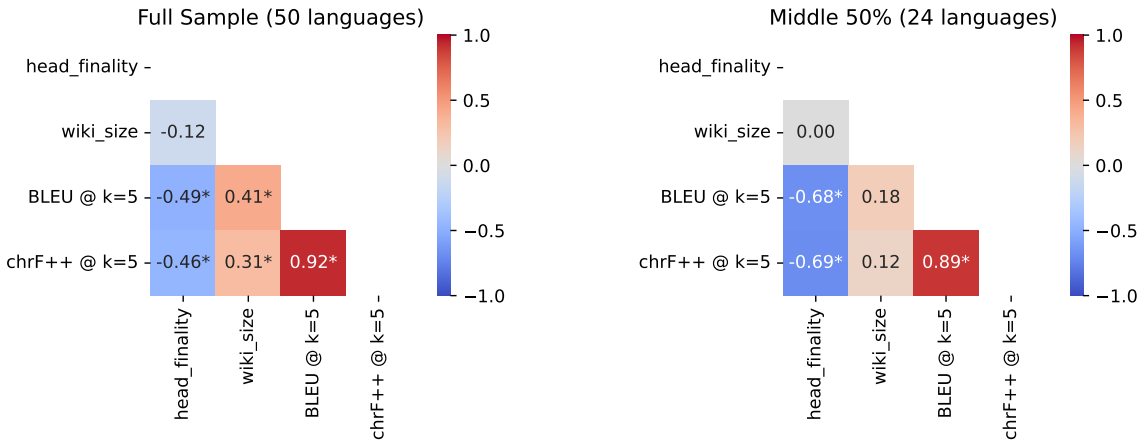

Figure 9: Spearman correlations between percentage of head-final phrases, Wikipedia size, and BLEU/chrF++ scores at beam size 5. Correlations are measured on the full language sample (left) and a more balanced interquartile range subset (right).

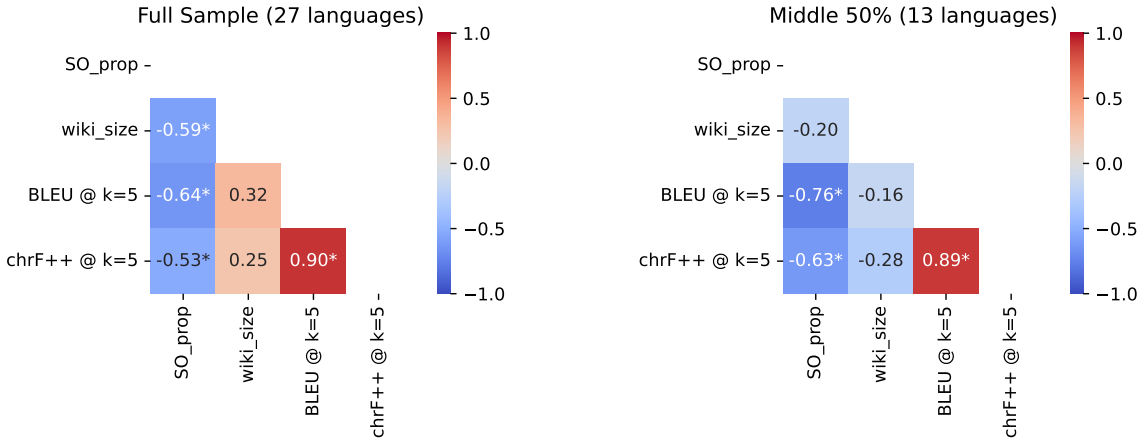

Figure 10: Spearman correlations between proportion of Subject-Object order, Wikipedia size, and BLEU/chrF++ scores at beam size 5. Correlations are measured on the full language sample (left) and a more balanced interquartile range subset (right).

In general, the morphosyntactic features determined significant in our correlation studies confirm our expectation that languages with richer morphology and flexible word order are more challenging for an NMT model to translate into (at least when the source language is English). The number of Wikipedia articles – an approximation of language representation in the NLLB-200 training data – was shown to have a significant link to MT performance, which corroborates established research. Genetic, geographic, and syntactic distances from English, our source language, also exhibit highly significant correlations with the BLEU and chrF++ scores. However, it is hard to decouple these relations from the effect of Wikipedia size even on the basis of a more balanced subset, as seen in

Figure 7. Conversely, the correlations with word order freedom measures, such as Subject-Object order proportion and percentage of head-final phrases, are among the strongest and most significant and get even stronger after controlling for Wikipedia size. Finally, only a couple of morphological complexity measures result in statistically significant weak correlations for each quality metric. Of those measures, the more significant ones are TTR and MATTR, which (i) are not as linguistically motivated as word/lemma entropy and (ii) depend not only on the language itself, but also on the specific tokenization scheme used in calculation. Taken together, these facts point at an overall weaker effect of morphology when compared to typological distance and word order flexibility.

5.2.2 Generation Probability Gains

Having investigated which continuous language properties serve as predictors of translation quality, we now focus on the crux of our research question: how these properties correlate with the change in translation difficulty as beam width expands. We conducted statistical studies for the gain in both translation quality and model generation probability. However, for our discussion, we choose to focus on the latter, since the results for modeling probability paint a more coherent picture and are independent from the effects of translation quality metrics. Still, we include correlations with translation quality gains in Appendices E and F.

We recall that a larger gain in generation probability can be interpreted as the need for the NMT model to explore a larger search space in order to generate a translation with high confidence. A large gain for a given language could thus signal that left-to-right beam search is a suboptimal decoding strategy for that language, or that the model is in general less confident in its own generated translations.

In Figure 11, we report correlations between language properties and generation probability gain, where the latter is represented as *absolute* gain, yielded by increasing beam size from $k = 1$ to $k = 7$, and *relative* gain, expressed in percents relative to probability at $k = 1$. For example, if a language's average generated sequence probability at $k = 1$ is 0.4 and at $k = 7$ it is 0.6, then the absolute probability gain is $0.6 - 0.4 = 0.2$, and the relative probability gain is $\frac{0.2}{0.4} \times 100 = 50\%$.

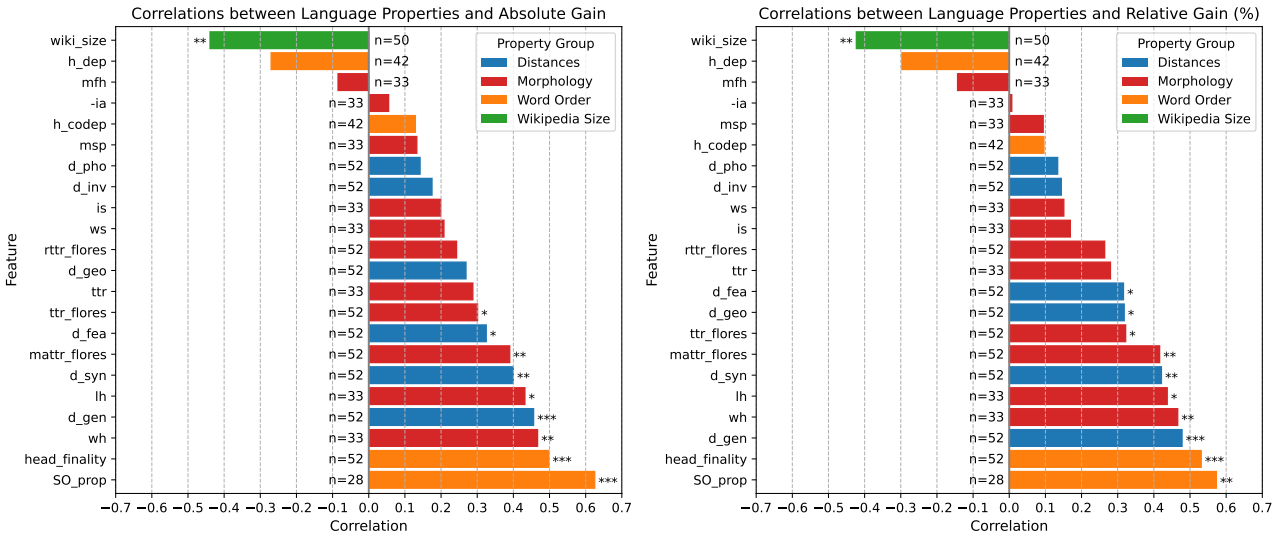

Figure 11: Spearman correlations between continuous language properties and absolute (left) and relative (right) generation probability gains. Properties are categorized into four groups (see Appendix C for breakdown). Sample sizes for each property are indicated next to their respective bars. Correlations significant at $p < 0.05$ are marked with "*", at $p < 0.01$ – with "**", at $p < 0.001$ – with "***".

As can be seen in the figure, continuous properties with statistically significant effect on probability gain are largely the same as in the case with translation quality scores (Figure 6). Overall, the positive correlations with the measures of morphosyntactic complexity point to the fact that languages with richer morphology and more flexible word order obtain a larger gain in modeling probability with expanded beam size, which is in line with our hypothesis.

Wikipedia Size Figure 11 highlights a very significant moderately negative correlation between Wikipedia size (`wiki_size`) and probability gain. This finding suggests that, when translating into a high-resource language, the model's confidence is already high at a small beam size and does not change much with expanded search space, whereas translating into a low-resource language benefits more from increasing beam width. Much like in §5.2.1, we attempted to minimize the potential confounding effect of the number of Wikipedia articles on the correlations with typological features by recalculating them on more balanced subsets of languages whose Wikipedia sizes fall in the interquartile ranges of the respective original language samples.

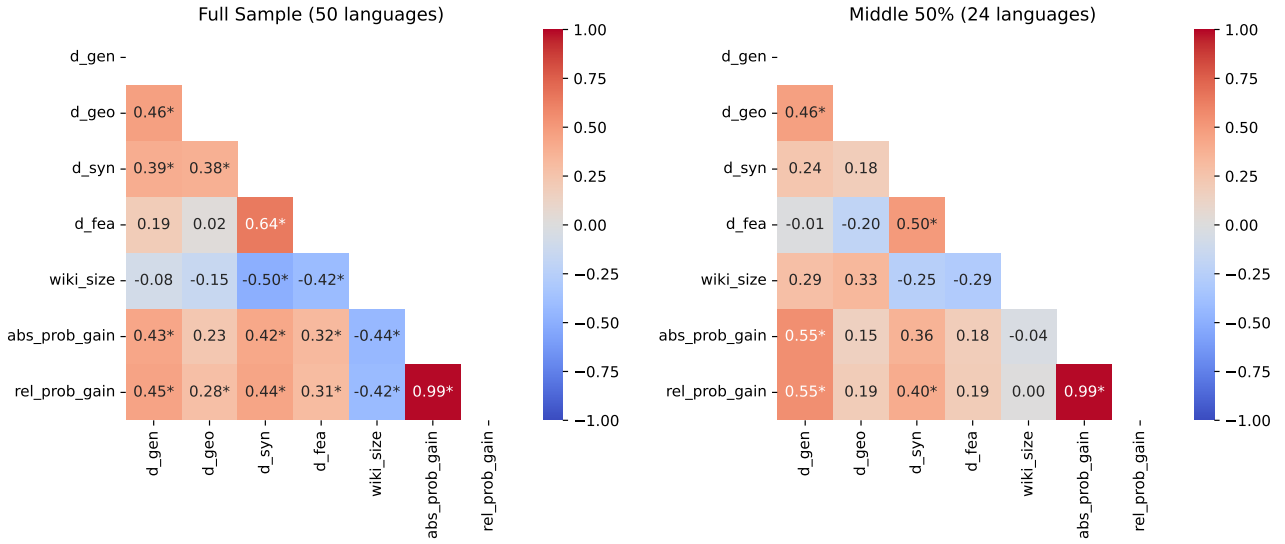

Figure 12: Spearman correlations between typological distances, Wikipedia size, and absolute/relative probability gain at $k = 7$. Correlations are measured on the full language sample (left) and a more balanced interquartile range subset (right).

Typological Distances from English Four measures of typological distance from English emerged as significant predictors of generation probability gain: *genetic* (`d_gen`), *geographic* (`d_geo`), *syntactic* (`d_syn`), and *featural* (`d_fea`). When calculated on a more data size-balanced subset (Figure 12), the effect of genetic distance got stronger, while the effects of other distances diminished and practically lost their significance. In other words, even when controlled for Wikipedia size, languages that are genetically far from English see a larger increase in generation probabilities with increased beam size compared to the languages closer to English.

Morphological Complexity Similarly to our findings in §5.2.1, the only measures of morphological complexity that correlate significantly with generation probability gain are *type/token ratio* (`ttr_flores`) and *moving average type/token ratio* (`mattr_flores`) measured on FLORES+ as well as *word entropy* (`wh`) and *lemma entropy* (`lh`). The behaviors of these correlations diverge when controlling for Wikipedia size: whereas the effects of word entropy and lemma entropy get weaker and become statistically insignificant (Figure 13), the effects of the TTR measures get substantially stronger (Figure 14). While these two trends contradict each other, the correlations with the TTR measures were calculated on a larger sample of languages and as such are more robust than the findings for the entropy measures. However, as pointed out in §5.2.1, TTR is less linguistically motivated and depends on the choice of a tokenization scheme (which in turn can vary a lot for morphologically

complex languages). In light of this, while the correlations suggest that expanding beam width leads to a larger increase in the model's confidence when translating into languages with richer morphology, these results should be interpreted with caution.

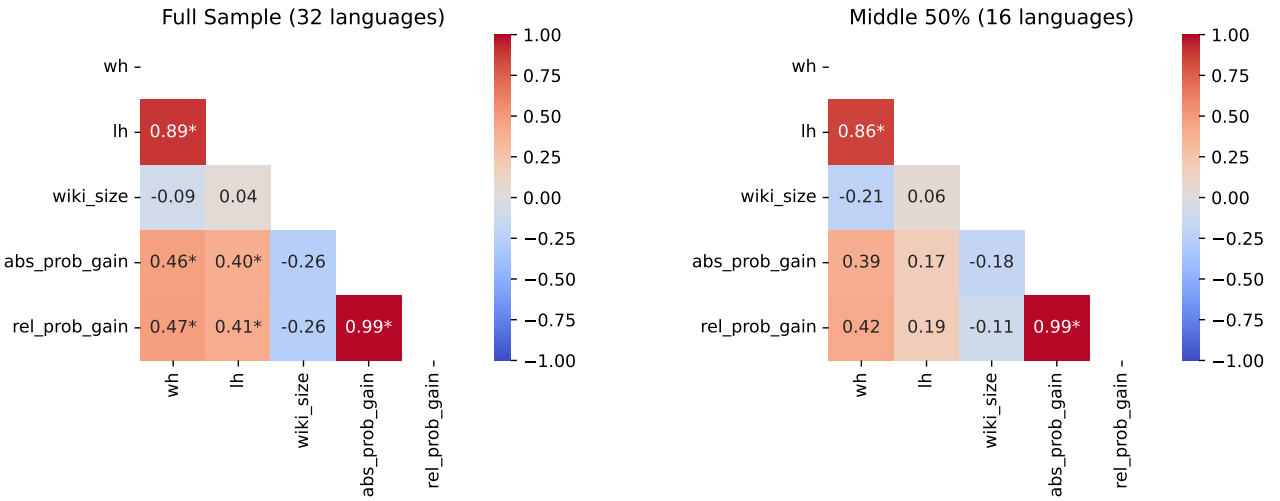

Figure 13: Spearman correlations between word and lemma entropies, Wikipedia size, and absolute/relative probability gain at $k = 7$. Correlations are measured on the full language sample (left) and a more balanced interquartile range subset (right).

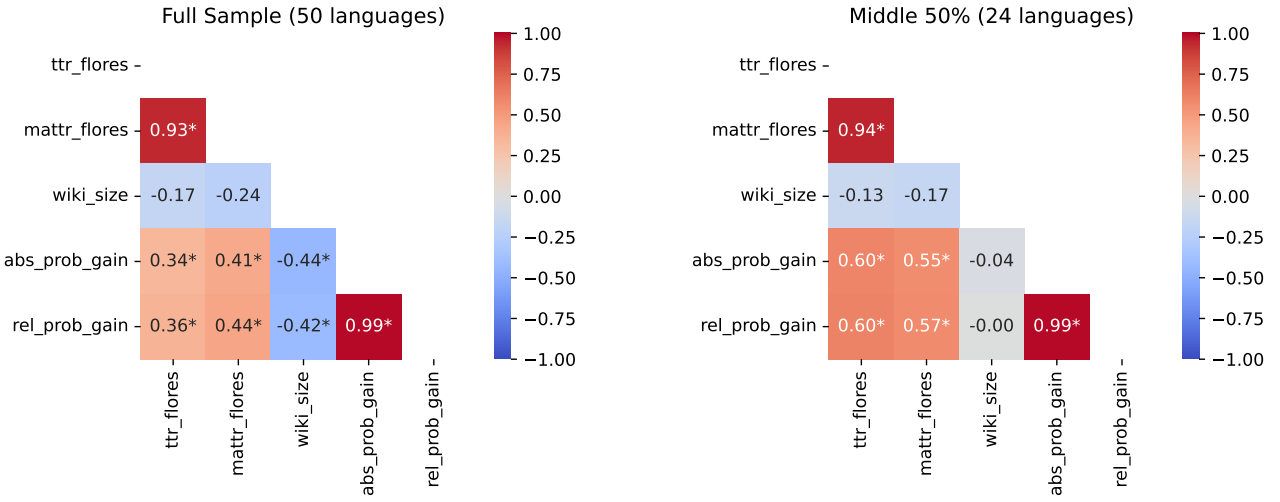

Figure 14: Spearman correlations between TTR and MATTR measures on FLORES+, Wikipedia size, and absolute/relative probability gain at $k = 7$. Correlations are measured on the full language sample (left) and a more balanced interquartile range subset (right).

Word Order Freedom From Figure 11 we can observe that the strongest positive correlations with generation probability gain belong to two measures of word order freedom: *proportion of Subject-*

Object (SO_prop) and *percentage of head-final phrases* (head_finality). After minimizing the factor of Wikipedia size, the effect of head_finality is amplified even further, as seen in Figure 15, Interestingly, the correlation coefficients of SO_prop stay the same even on a more balanced sample in terms of Wikipedia size, as illustrated in Figure 16.

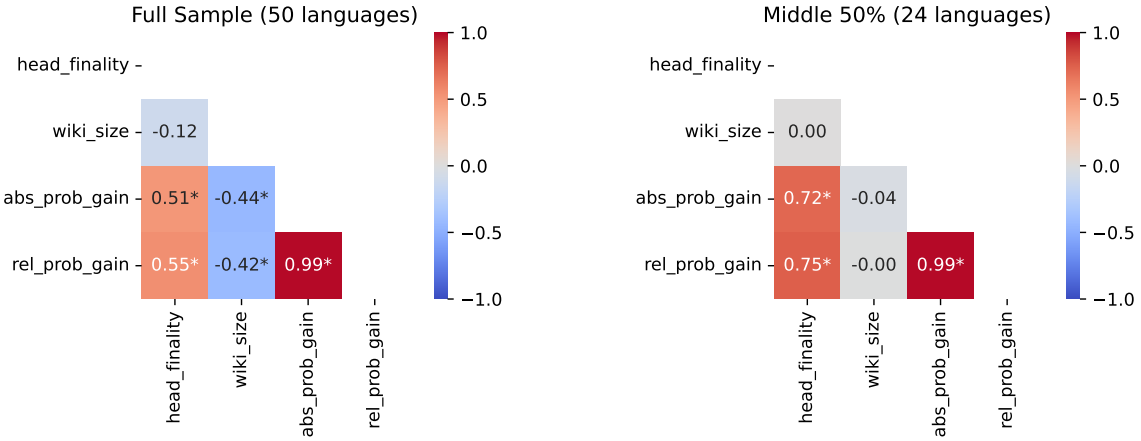

Figure 15: Spearman correlations between percentage of head-final phrases, Wikipedia size, and absolute/relative probability gain at $k = 7$. Correlations are measured on the full language sample (left) and a more balanced interquartile range subset (right).

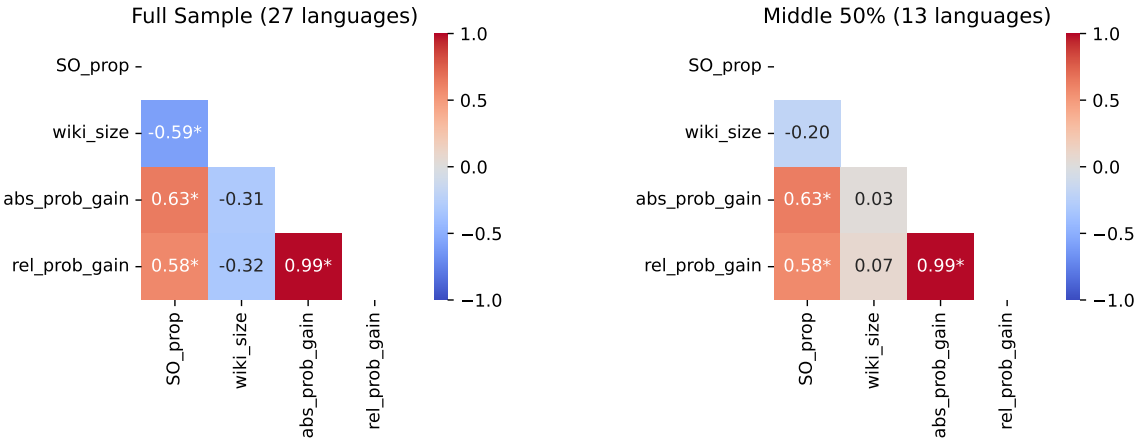

Figure 16: Spearman correlations between proportion of Subject-Object order, Wikipedia size, and absolute/relative probability gain at $k = 7$. Correlations are measured on the full language sample (left) and a more balanced interquartile range subset (right).

The results of our correlation studies with generation probability largely provide evidence that languages which have complex morphology, flexible word order, and are typologically distant from English see a greater increase in modeling probability at larger beam widths. Language representa-

tion in the NLLB-200 model's training data, approximated via Wikipedia size, exhibits a moderate negative effect on probability gain. This implies that widening search space improves the model's confidence in a generated translation more so for high-resource languages than for low-resource languages. A number of typological distance measures – particularly genetic distance from English – also play a crucial role in the amount of generation probability gain. The effects of two word order flexibility measures came out the strongest in our studies, especially after controlling for Wikipedia size: generation probability gain yielded by increasing beam size is much higher for languages with more freedom in the order of main constituents. Lastly, our observations for two TTR measures suggest that target language morphological complexity is also a key factor in the degree of translation confidence improvement at higher beam size values. This finding, however, is less robust compared to typological distances and word order measures due to a smaller number of statistically significant morphological complexity features and the susceptibility of TTR to extralinguistic factors such as tokenization schemes.

6 Conclusion

In this thesis, we investigated the relationship between morphosyntactic properties of a language, width of output search space, and the difficulty of neural machine translation. Our primary research aim was to determine how NMT performance improvement, driven by beam size expansion, varies depending on typological properties of target languages. More generally, we also wanted to identify which language properties serve as predictors of translation quality using a state-of-the-art NMT model and a wider range of language properties compared to previous works in this area.

Leveraging the FLORES+ MT evaluation dataset and a state-of-the-art NLLB-200 model, we translated 997 English sentences into 124 typologically diverse languages under four beam size settings and evaluated the difficulty of NMT using a combination of translation quality metrics and modeling probability measures. We then conducted a series of correlation studies to evaluate the effect of morphosyntactic properties on translation difficulty and the change thereof. Our statistical tests reveal a number of morphosyntactic features with significant effect on both translation quality and the degree of translation difficulty change. These findings suggest that languages with typological properties significantly different from English may benefit from decoding strategies other than the current de facto standard of left-to-right beam search.

6.1 Summary of Main Contributions

The main contributions of our work can be summarized as follows:

1. Choosing the FLORES+ translation evaluation dataset as our initial language set, we compiled diverse linguistic properties from multiple sources into a comprehensive dataset, which we make publicly available to facilitate future research: `https://github.com/v-hirak/explaining-MT-difficulty`. Additionally, to compensate for the lack of official information on language distribution in NLLB-200 training data, we used Wikipedia sizes in article counts per language as an approximate estimation, which we incorporated alongside intrinsic language properties into our statistical studies and the publicly released dataset.

2. We performed the first set of correlation studies involving categorical WALS features of target languages using a new task (MT) and a state-of-the-art model. Our findings point at a significant effect of formative exponence, reduplication, and syncretism on the quality of NMT. In addition,

we showed that the gain in translation quality associated with beam size increase is significantly higher for languages with a strong preference for suffixing, compared to languages with no inflectional morphology.

3. Our second set of correlation experiments involved the use of novel continuous language properties, which are argued by recent literature to better reflect the typology of languages. Our experiments revealed statistically significant effects of several such typological features on translation quality.

4. We also tested the effect of continuous language properties on the gain in modeling probability achieved by beam width expansion, which, to our best knowledge, has not been studied before. The significant features include typological distance from English, type/token ratios, flexibility of Subject-Object order, and head-directionality. The observed effects persisted even when Wikipedia size (proxy for training data availability) was accounted for, which serves as evidence that languages with higher degrees of morphological complexity and word order freedom benefit more from expanded search space and thus may require alternative decoding mechanisms.

6.2 Future Work

The present study paves the way for several possible research directions that would expand our findings and address existing limitations. For instance, while we use English as our only source language, future research should expand the selection of source languages to properly disentangle the impact of intrinsic language properties (such as morphology and word order) from the effect of source-target language distance. Moreover, future studies could employ methods such as linear mixed models to measure the effect of intrinsic linguistic properties while isolating the factor of training data imbalance. Finally, further research should explore alternative decoding strategies to assess their benefits for languages with rich morphology and flexible word order.

We hope this thesis sheds new light on the challenges of machine translation and serves as a valuable contribution to the ongoing efforts towards achieving language equality in NLP.

Bibliography

[1] L. Lu, "Digital divide: Does the internet speak your language?," in *EdMedia+ Innovate Learning*, pp. 4022–4025, Association for the Advancement of Computing in Education (AACE), 2010.

[2] L. Benkova, D. Munkova, L. Benko, and M. Munk, "Evaluation of english–slovak neural and statistical machine translation," *Applied Sciences*, vol. 11, no. 7, p. 2948, 2021.

[3] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[4] S. Jean, K. Cho, R. Memisevic, and Y. Bengio, "On using very large target vocabulary for neural machine translation," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (C. Zong and M. Strube, eds.), (Beijing, China), pp. 1–10, Association for Computational Linguistics, July 2015.

[5] T. Luong, I. Sutskever, Q. Le, O. Vinyals, and W. Zaremba, "Addressing the rare word problem in neural machine translation," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (C. Zong and M. Strube, eds.), (Beijing, China), pp. 11–19, Association for Computational Linguistics, July 2015.

[6] M.-T. Luong and C. Manning, "Stanford neural machine translation systems for spoken language domains," in *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, (Da Nang, Vietnam), pp. 76–79, Dec. 3-4 2015.

[7] L. Bentivogli, A. Bisazza, M. Cettolo, and M. Federico, "Neural versus phrase-based machine translation quality: a case study," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, (Austin, Texas), pp. 257–267, Association for Computational Linguistics, Nov. 2016.

[8] D. Ataman and M. Federico, "An evaluation of two vocabulary reduction methods for neural ma-

chine translation," in *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)* (C. Cherry and G. Neubig, eds.), (Boston, MA), pp. 97–110, Association for Machine Translation in the Americas, Mar. 2018.

[9] E. Bugliarello, S. J. Mielke, A. Anastasopoulos, R. Cotterell, and N. 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* (D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, eds.), (Online), pp. 1640–1649, Association for Computational Linguistics, July 2020.

[10] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, and C. Leahy, "The pile: An 800gb dataset of diverse text for language modeling," 2020.

[11] S. J. Mielke, R. Cotterell, K. Gorman, B. Roark, and J. Eisner, "What kind of language is hard to language-model?," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 4975–4989, Association for Computational Linguistics, July 2019.

[12] C. Arnett and B. K. Bergen, "Why do language models perform worse for morphologically complex languages?," *arXiv preprint arXiv:2411.14198*, 2024.

[13] Y. Belinkov, N. Durrani, F. Dalvi, H. Sajjad, and J. Glass, "What do neural machine translation models learn about morphology?," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (R. Barzilay and M.-Y. Kan, eds.), (Vancouver, Canada), pp. 861–872, Association for Computational Linguistics, July 2017.

[14] R. Cotterell, S. J. Mielke, J. Eisner, and B. 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)*, (New Orleans, Louisiana), pp. 536–541, Association for Computational Linguistics, June 2018.

[15] A. Bisazza, A. Üstün, and S. Sportel, "On the difficulty of translating free-order case-marking languages," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1233–1248, 2021.

[16] H. H. Park, K. J. Zhang, C. Haley, K. Steimel, H. Liu, and L. Schwartz, "Morphology matters: A multilingual language modeling analysis," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 261–276, 2021.

[17] A. Wan, "Fairness in representation for multilingual nlp: Insights from controlled experiments on conditional language modeling," in *International Conference on Learning Representations*, 2022.

[18] N. Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, and J. Wang, "No language left behind: Scaling human-centered machine translation," 2022.

[19] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020.

[20] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," 2023.

[21] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, 2014.

[22] A. Birch, M. Osborne, and P. Koehn, "Predicting success in machine translation," in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 745–754, 2008.

[23] D. Jurafsky and J. H. Martin, *Speech and Language Processing (2nd Edition)*. USA: Prentice-Hall, Inc., 2009.

[24] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, *Dive into deep learning*. Cambridge University

Press, 2023.

[25] M. Junczys-Dowmunt, T. Dwojak, and H. Hoang, "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, Dec. 8-9 2016.

[26] M. Freitag and Y. Al-Onaizan, "Beam search strategies for neural machine translation," in *Proceedings of the First Workshop on Neural Machine Translation*, (Vancouver), pp. 56–60, Association for Computational Linguistics, Aug. 2017.

[27] C. Park, Y. Yang, K. Park, and H. Lim, "Decoding strategies for improving low-resource machine translation," *Electronics*, vol. 9, no. 10, p. 1562, 2020.

[28] P. Koehn and R. Knowles, "Six challenges for neural machine translation," in *Proceedings of the First Workshop on Neural Machine Translation*, (Vancouver), pp. 28–39, Association for Computational Linguistics, Aug. 2017.

[29] E. Cohen and C. Beck, "Empirical analysis of beam search performance degradation in neural sequence models," in *International Conference on Machine Learning*, pp. 1290–1299, PMLR, 2019.

[30] R. Sennrich, O. Firat, K. Cho, A. Birch, B. Haddow, J. Hitschler, M. Junczys-Dowmunt, S. Läubli, A. V. Miceli Barone, J. Mokry, and M. Nădejde, "Nematus: a toolkit for neural machine translation," in *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, (Valencia, Spain), pp. 65–68, Association for Computational Linguistics, Apr. 2017.

[31] Z. Tu, Y. Liu, L. Shang, X. Liu, and H. Li, "Neural machine translation with reconstruction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, 2017.

[32] Y. Yang, L. Huang, and M. Ma, "Breaking the beam search curse: A study of (re-)scoring methods and stopping criteria for neural machine translation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, eds.), (Brussels, Belgium), pp. 3054–3059, Association for Computational Lin-

guistics, Oct.-Nov. 2018.

[33] M. S. Dryer and M. Haspelmath, eds., *WALS Online (v2020.3)*. Zenodo, 2013.

[34] M. S. Dryer, "Prefixing vs. suffixing in inflectional morphology (v2020.3)," in *The World Atlas of Language Structures Online* (M. S. Dryer and M. Haspelmath, eds.), Zenodo, 2013.

[35] M. S. Dryer, "81 order of subject, object, and verb," *The world atlas of language structures, ed. by Martin Haspelmath et al*, pp. 330–333, 2005.

[36] N. Levshina, S. Namboodiripad, M. Allassonnière-Tang, M. Kramer, L. Talamo, A. Verkerk, S. Wilmoth, G. G. Rodriguez, T. M. Gupton, E. Kidd, *et al.*, "Why we need a gradient approach to word order," *Linguistics*, vol. 61, no. 4, pp. 825–883, 2023.

[37] E. Baylor, E. Ploeger, and J. Bjerva, "Multilingual gradient word-order typology from universal dependencies," 2024.

[38] H. Skirgård, H. J. Haynie, D. E. Blasi, H. Hammarström, J. Collins, J. J. Latarche, J. Lesage, T. Weber, A. Witzlack-Makarevich, S. Passmore, *et al.*, "Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss," *Science Advances*, vol. 9, no. 16, p. eadg6175, 2023.

[39] J. Nivre, D. Zeman, F. Ginter, and F. Tyers, "Universal Dependencies," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, (Valencia, Spain), Association for Computational Linguistics, Apr. 2017.

[40] J. W. Chotlos, "Iv. a statistical and comparative analysis of individual written language samples.," *Psychological Monographs*, vol. 56, no. 2, p. 75, 1944.

[41] A. Xanthos, S. Laaha, S. Gillis, U. Stephany, A. Aksu-Koç, A. Christofidou, N. Gagarina, G. Hrzica, F. N. Ketrez, M. Kilani-Schoch, *et al.*, "On the role of morphological richness in the early development of noun and verb inflection," *First Language*, vol. 31, no. 4, pp. 461–479, 2011.

[42] P. Juola, "Measuring linguistic complexity: The morphological tier," *Journal of Quantitative Linguistics*, vol. 5, no. 3, pp. 206–213, 1998.

[43] C. Bentz and D. Alikaniotis, "The word entropy of natural languages," 2016.

[44] R. Cotterell, C. Kirov, M. Hulden, and J. Eisner, "On the complexity and typology of inflectional morphological systems," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 327–342, 2019.

[45] B. Bickel and J. Nichols, "Inflectional synthesis of the verb (v2020.3)," in *The World Atlas of Language Structures Online* (M. S. Dryer and M. Haspelmath, eds.), Zenodo, 2013.

[46] A. Lyovin, *An Introduction to the Languages of the World*. Oxford University Press, USA, 1997.

[47] V. Starko and A. Rysin, "Vesum: A large morphological dictionary of ukrainian as a dynamic tool.," in *COLINS*, pp. 61–70, 2022.

[48] B. Sagot, "Comparing complexity measures," in *Computational approaches to morphological complexity*, 2013.

[49] R. K. Shosted, "Correlating complexity: A typological approach," 2006.

[50] Ç. Çöltekin and T. Rama, "What do complexity measures measure? correlating and validating corpus-based measures of morphological complexity," *Linguistics Vanguard*, vol. 9, no. s1, pp. 27–43, 2023.

[51] M. Templin, "Certain language skills in children: Their development and interrelationships," 1957.

[52] C. Bentz, T. Ruzsics, A. Koplenig, and T. Samardžić, "A comparison between morphological complexity measures: Typological data vs. language corpora," in *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, (Osaka, Japan), pp. 142–153, The COLING 2016 Organizing Committee, Dec. 2016.

[53] G. Sarti, A. Bisazza, A. Guerberof-Arenas, and A. Toral, "DivEMT: Neural machine translation post-editing effort across typologically diverse languages," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (Y. Goldberg, Z. Kozareva, and Y. Zhang, eds.), (Abu Dhabi, United Arab Emirates), pp. 7795–7816, Association for Computational Linguistics, Dec. 2022.

[54] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[55] N. Levshina, "Token-based typology and word order entropy: A study based on universal dependencies," *Linguistic Typology*, vol. 23, no. 3, pp. 533–572, 2019.

[56] K. Sinnemäki, "Complexity trade-offs in core argument marking," *Language complexity*, pp. 67–88, 2008.

[57] R. Futrell, K. Mahowald, and E. Gibson, "Quantifying word order freedom in dependency corpora," in *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, (Uppsala, Sweden), pp. 91–100, Uppsala University, Uppsala, Sweden, Aug. 2015.

[58] R. Cotterell, C. Kirov, J. Sylak-Glassman, G. Walther, E. Vylomova, A. D. McCarthy, K. Kann, S. J. Mielke, G. Nicolai, M. Silfverberg, *et al.*, "The conll–sigmorphon 2018 shared task: Universal morphological reinflection," *arXiv preprint arXiv:1810.07125*, 2018.

[59] A. Koplenig, P. Meyer, S. Wolfer, and C. Müller-Spitzer, "The statistical trade-off between word order and word structure–large-scale evidence for the principle of least effort," *PloS one*, vol. 12, no. 3, p. e0173614, 2017.

[60] M. H. Christiansen and J. T. Devlin, "Recursive inconsistencies are hard to learn: A connectionist perspective on universal word order correlations," in *Proceedings of the 19th annual cognitive science society conference*, pp. 113–118, Lawrence Erlbaum Associates Mahwah, NJ, 1997.

[61] M. R. Ellefson and M. H. Christiansen, "Subjacency constraints without universal grammar: Evidence from artificial language learning and connectionist modeling," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 22, 2000.

[62] G. Lupyan and M. H. Christiansen, "Case, Word Order, and Language Learnability: Insights from Connectionist Modeling," *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 24, no. 24, 2002.

[63] E. v. Everbroeck, "Language type frequency and learnability from a connectionist perspective,"

2003.

[64] J. L. Elman, "Finding structure in time," *Cognitive science*, vol. 14, no. 2, pp. 179–211, 1990.

[65] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proceedings of Machine Translation Summit X: Papers*, (Phuket, Thailand), pp. 79–86, Sept. 13-15 2005.

[66] C. Christodouloupoulos and M. Steedman, "A massively parallel corpus: the bible in 100 languages," *Language resources and evaluation*, vol. 49, pp. 375–395, 2015.

[67] S. Ravfogel, Y. Goldberg, and T. Linzen, "Studying the inductive biases of RNNs with synthetic variations of natural languages," 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)*, (Minneapolis, Minnesota), pp. 3532–3542, Association for Computational Linguistics, June 2019.

[68] M. Bisani and H. Ney, "Open vocabulary speech recognition with flat hybrid models.," in *Interspeech*, pp. 725–728, 2005.

[69] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," 2015.

[70] Y. Shibata, T. Kida, S. Fukamachi, M. Takeda, A. Shinohara, T. Shinohara, and S. Arikawa, "Byte pair encoding: A text compression scheme that accelerates pattern matching," 1999.

[71] M. Dehouck and P. Denis, "A framework for understanding the role of morphology in universal dependency parsing," in *EMNLP 2018-Conference on Empirical Methods in Natural Language Processing*, 2018.

[72] H. Liu, "Dependency distance as a metric of language comprehension difficulty," *Journal of Cognitive Science*, vol. 9, no. 2, pp. 159–191, 2008.

[73] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[74] M. Creutz and K. Lagus, "Unsupervised models for morpheme segmentation and morphology learning," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 4, no. 1, pp. 1–

34, 2007.

[75] N. Du, Y. Huang, A. Dai, S. Tong, D. Lepikhin, Y. Xu, M. Krikun, Y. Zhou, A. Yu, O. Firat, *et al.*, "Glam: Efficient scaling of language models with mixture-of-experts. arxiv," *arXiv preprint arXiv:2112.06905*, 2021.

[76] C. Hwang, W. Cui, Y. Xiong, Z. Yang, Z. Liu, H. Hu, Z. Wang, R. Salas, J. Jose, P. Ram, *et al.*, "Tutel: Adaptive mixture-of-experts at scale," *Proceedings of Machine Learning and Systems*, vol. 5, pp. 269–287, 2023.

[77] K. Heffernan, O. Çelebi, and H. Schwenk, "Bitext mining using distilled sentence representations for low-resource languages," 2022.

[78] P. Andrews, G. Wenzek, K. Heffernan, O. Çelebi, A. Sun, A. Kamran, Y. Guo, A. Mourachko, H. Schwenk, and A. Fan, "stopes-modular machine translation pipelines," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 258–265, 2022.

[79] F. Guzmán, P.-J. Chen, M. Ott, J. Pino, G. Lample, P. Koehn, V. Chaudhary, and M. Ranzato, "The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English," 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)*, (Hong Kong, China), pp. 6098–6111, Association for Computational Linguistics, Nov. 2019.

[80] N. Goyal, C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, S. K. Da Ju, F. G. Marc'Aurelio Ranzato, and A. Fan, "The flores-101 evaluation benchmark for low-resource and multilingual machine translation. corr, abs/2106.03193," 2021.

[81] Y.-H. Lin, C.-Y. Chen, J. Lee, Z. Li, Y. Zhang, M. Xia, S. Rijhwani, J. He, Z. Zhang, X. Ma, A. Anastasopoulos, P. Littell, and G. Neubig, "Choosing transfer languages for cross-lingual learning," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 3125–3135, Association for Computational Linguistics, July 2019.

[82] P. Littell, D. R. Mortensen, K. Lin, K. Kairis, C. Turner, and L. Levin, "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*, (Valencia, Spain), pp. 8–14, Association for Computational Linguistics, Apr. 2017.

[83] H. Hammarström, R. Forkel, M. Haspelmath, and S. Bank, "Glottolog 5.0," 2024. Available online at `http://glottolog.org`, Accessed on 2024-09-19.

[84] S. Moran and D. McCloy, eds., *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History, 2019.

[85] L. Campbell and V. Grondona, "Ethnologue: Languages of the world," *Language*, vol. 84, no. 3, pp. 636–641, 2008.

[86] M. A. Covington and J. D. McFall, "Cutting the gordian knot: The moving-average type–token ratio (mattr)," *Journal of quantitative linguistics*, vol. 17, no. 2, pp. 94–100, 2010.

[87] L. Shen, "LexicalRichness: A small module to compute textual lexical richness," 2022.

[88] P. Guiraud, "Problèmes et méthodes de la statistique linguistique," *(No Title)*, 1959.

[89] D. Goldhahn, T. Eckart, U. Quasthoff, *et al.*, "Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages.," in *LREC*, vol. 29, pp. 31–43, 2012.

[90] T. Osborne and K. Gerdes, "The status of function words in dependency grammar: A critique of universal dependencies (ud)," *Glossa: a journal of general linguistics (2016-2021)*, 2019.

[91] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, (Philadelphia, Pennsylvania, USA), pp. 311–318, Association for Computational Linguistics, July 2002.

[92] M. Post, "A call for clarity in reporting BLEU scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*, (Brussels, Belgium), pp. 186–191, Association for Computational Linguistics, Oct. 2018.

[93] M. Popović, "chrF: character n-gram F-score for automatic MT evaluation," in *Proceedings of the Tenth Workshop on Statistical Machine Translation*, (Lisbon, Portugal), pp. 392–395, Association for Computational Linguistics, Sept. 2015.

[94] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, "Comet: A neural framework for mt evaluation," 2020.

[95] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, (Online), pp. 38–45, Association for Computational Linguistics, Oct. 2020.

[96] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

Appendices

A Target Languages

Code	Name	Code	Name	Code	Name
ace_Arab	Acehnese (Arabic script)	kan_Knda	Kannada	sun_Latn	Sundanese
ace_Latn	Acehnese (Latin script)	kat_Geor	Georgian	swe_Latn	Swedish*
acm_Arab	Mesopotamian Arabic	kaz_Cyrl	Kazakh*	swh_Latn	Swahili
afr_Latn	Afrikaans*	khk_Cyrl	Halh Mongolian	tam_Taml	Tamil*
amh_Ethi	Amharic	khm_Khmr	Khmer (Central)	taq_Latn	Tamasheq (Latin script)
arb_Arab	Modern Standard Arabic*	kik_Latn	Kikuyu	taq_Tfng	Tamasheq (Tifinagh script)
ary_Arab	Moroccan Arabic	kir_Cyrl	Kyrgyz	tat_Cyrl	Tatar
arz_Arab	Egyptian Arabic	kmr_Latn	Northern Kurdish	tel_Telu	Telugu*
asm_Beng	Assamese	knc_Arab	Central Kanuri (Arabic script)	tgk_Cyrl	Tajik
ayr_Latn	Central Aymara	knc_Latn	Central Kanuri (Latin script)	tha_Thai	Thai
azb_Arab	South Azerbaijani	kor_Hang	Korean*	tir_Ethi	Tigrinya
bak_Cyrl	Bashkir	lao_Laoo	Lao	tsn_Latn	Tswana
bam_Latn	Bambara	lit_Latn	Lithuanian*	tuk_Latn	Turkmen
ban_Latn	Balinese	lug_Latn	Ganda	tur_Latn	Turkish*
bel_Cyrl	Belarusian*	luo_Latn	Luo	uig_Arab	Uyghur*
bho_Deva	Bhojpuri	lus_Latn	Mizo	ukr_Cyrl	Ukrainian*
bod_Tibt	Lhasa Tibetan	mai_Deva	Maithili	urd_Arab	Urdu*
bul_Cyrl	Bulgarian*	mal_Mlym	Malayalam	vie_Latn	Vietnamese*
cat_Latn	Catalan*	mar_Deva	Marathi*	war_Latn	Waray
ces_Latn	Czech*	min_Latn	Minangkabau (Latin script)	wol_Latn	Wolof*
ckb_Arab	Central Kurdish	mni_Beng	Meitei (Manipuri, Bengali script)	xho_Latn	Xhosa
cmn_Hans	Mandarin Chinese (Standard Beijing)*	mos_Latn	Mossi	yor_Latn	Yoruba
cmn_Hant	Mandarin Chinese (Taiwanese)*	mri_Latn	Maori	yue_Hant	Yue Chinese
cym_Latn	Welsh*	mya_Mymr	Burmese	zul_Latn	Zulu
dan_Latn	Danish*	nld_Latn	Dutch*		
deu_Latn	German*	npi_Deva	Nepali		
ekk_Latn	Estonian*	nus_Latn	Nuer		
ell_Grek	Greek*	ory_Orya	Odia		
eus_Latn	Basque*	pag_Latn	Pangasinan		
ewe_Latn	Ewe	pan_Guru	Eastern Panjabi		
fij_Latn	Fijian	pes_Arab	Western Persian*		
fin_Latn	Finnish*	plt_Latn	Plateau Malagasy		
fra_Latn	French*	pol_Latn	Polish*		
gla_Latn	Scottish Gaelic*	por_Latn	Portuguese (Brazilian)*		
gle_Latn	Irish*	prs_Arab	Dari		
glg_Latn	Galician*	quy_Latn	Ayacucho Quechua		
hau_Latn	Hausa	ron_Latn	Romanian*		
heb_Hebr	Hebrew*	run_Latn	Rundi		
hin_Deva	Hindi*	rus_Cyrl	Russian*		
hrv_Latn	Croatian*	sag_Latn	Sango		
hun_Latn	Hungarian*	san_Deva	Sanskrit*		
hye_Armn	Armenian*	slk_Latn	Slovak*		
ibo_Latn	Igbo	slv_Latn	Slovenian*		
ilo_Latn	Ilocano	smo_Latn	Samoan		
ind_Latn	Indonesian*	sna_Latn	Shona		
isl_Latn	Icelandic*	som_Latn	Somali		
ita_Latn	Italian*	sot_Latn	Southern Sotho		
jpn_Jpan	Japanese*	spa_Latn	Spanish (Latin American)*		
kac_Latn	Jingpho	srp_Cyrl	Serbian*		
kam_Latn	Kamba	ssw_Latn	Swati		

Table 6: 124 languages that we translate into and include quality scores for. For languages with "*" we additionally compute generation probabilities.

B Number of Wikipedia Articles for 105 Target Languages

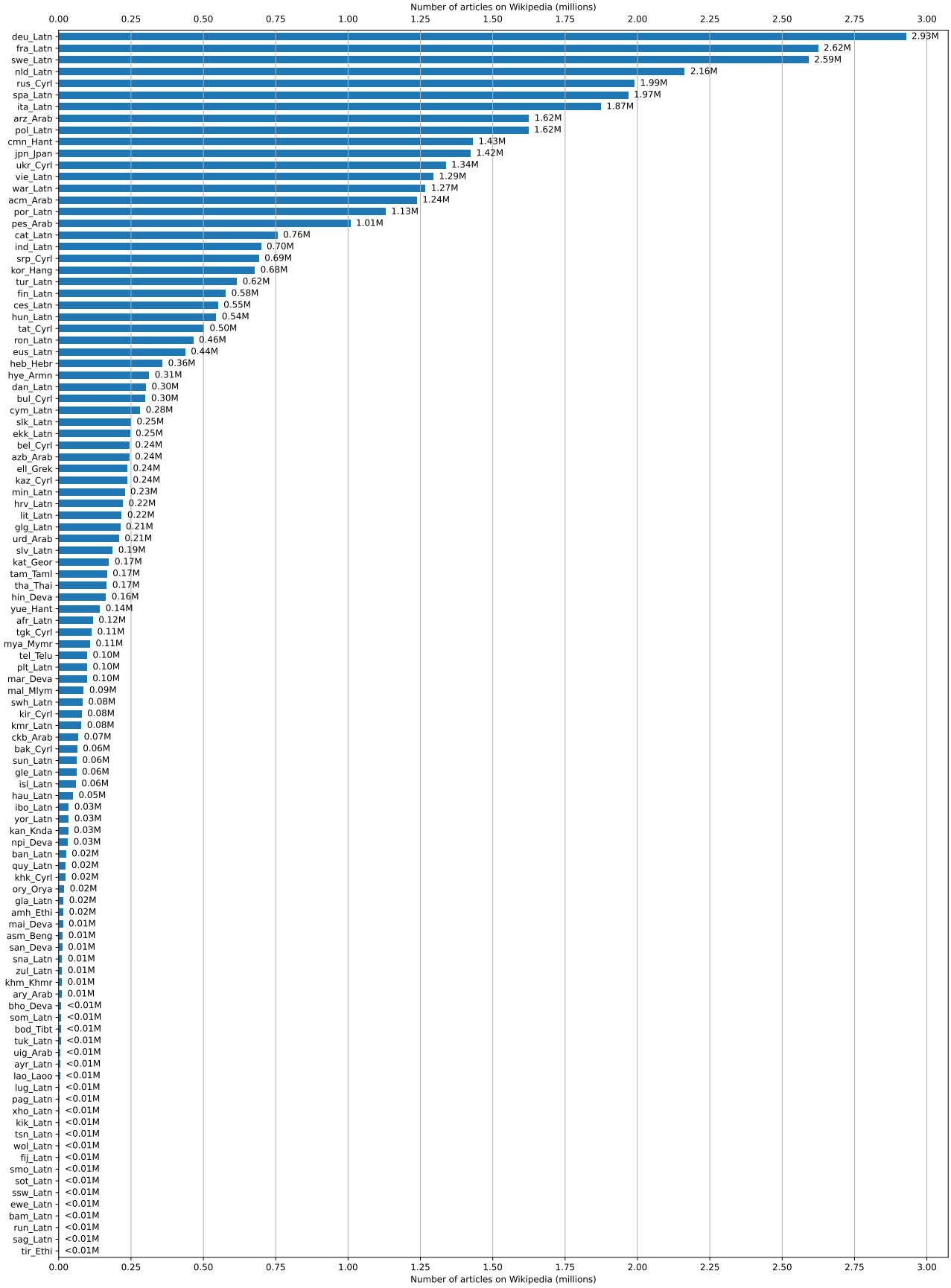

Figure 17: Wikipedia size in number of articles for 105 out of 124 target languages. We use this data to approximate language availability in NLLB-200 training data.

C Continuous Language Properties

Group	Property	Code	*N*
Typological distance from English	Genetic distance	d_gen	124
	Geographic distance	d_geo	
	Syntactic distance	d_syn	
	Inventory distance	d_inv	
	Phonological distance	d_pho	
	Featural distance	d_fea	
Morphological complexity (FLORES+)	Type/token ratio	ttr_flores	124
	Root type/token ratio	rttr_flores	
	Moving average type/token ratio	mattr_flores	
Morphological complexity (Çöltekin and Rama, 2023)	Type/token ratio	ttr	33
	Information in word structure	ws	
	Word entropy	wh	
	Lemma entropy	lh	
	Mean size of paradigm	msp	
	Inflectional synthesis	is	
	Morphological feature entropy	mfh	
	Inflection accuracy	-ia	
Word order freedom (Levshina, 2019 Levshina et al., 2023)	Dependency entropy	h_dep	44
	Codependency entropy	h_codep	43
	Subect-Object order proportion	SO_prop	28
	Head-final phrase percentage	head_finality	52
NLLB-200 training data representation estimate	Wikipedia size by language	wiki_size	105

Table 7: Continuous language properties and their categories used in our experiments. ***N*** indicates the number of languages for which a given property or a group of properties is available.

D WALS Features and Their Values

Values	# languages
21A: Exponence of Selected Inflectional Formatives	
Monoexponential case	20
No case	11
21B: Exponence of Tense-Aspect-Mood Inflection	
Monoexponential TAM	25
TAM + agreement	6
22A: Inflectional Synthesis of the Verb	
2-3 categories per verb	8
4-5 categories per verb	15
6-7 categories per verb	6
23A: Locus of Marking in the Clause	
P is head-marked	5
P is dependent-marked	20
P is double-marked	11
P has no marking	7
24A: Locus of Marking in Possessive Noun Phrases	
Possessor is head-marked	6
Possessor is dependent-marked	29
25A: Locus of Marking: Whole-language Typology	
Consistently dependent-marking	15
Inconsistent marking or other type	21
26A: Prefixing vs. Suffixing in Inflectional Morphology	
Little or no inflectional morphology	18
Predominantly suffixing	56
Moderate preference for suffixing	11
Approximately equal amounts of suffixing and prefixing	7
Predominantly prefixing	7
27A: Reduplication	
Productive full and partial reduplication	50
Full reduplication only	10
No productive reduplication	12
28A: Case Syncretism	
Inflectional case marking is absent or minimal	23
Inflectional case marking is syncretic for core and non-core cases	10
Inflectional case marking is never syncretic	7
29A: Syncretism in Verbal Person/Number Marking	
No subject person/number marking	17
Subject person/number marking is syncretic	12
Subject person/number marking is never syncretic	13
81A: Order of Subject, Object and Verb	
Subject-Object-Verb (SOV)	36
Subject-Verb-Object (SVO)	45
Verb-Subject-Object (VSO)	9
Lacking a dominant word order	13

Table 8: Final list of WALS features and their values which we include in our correlation studies.

E Correlations with Continuous Language Properties on Different Language Samples

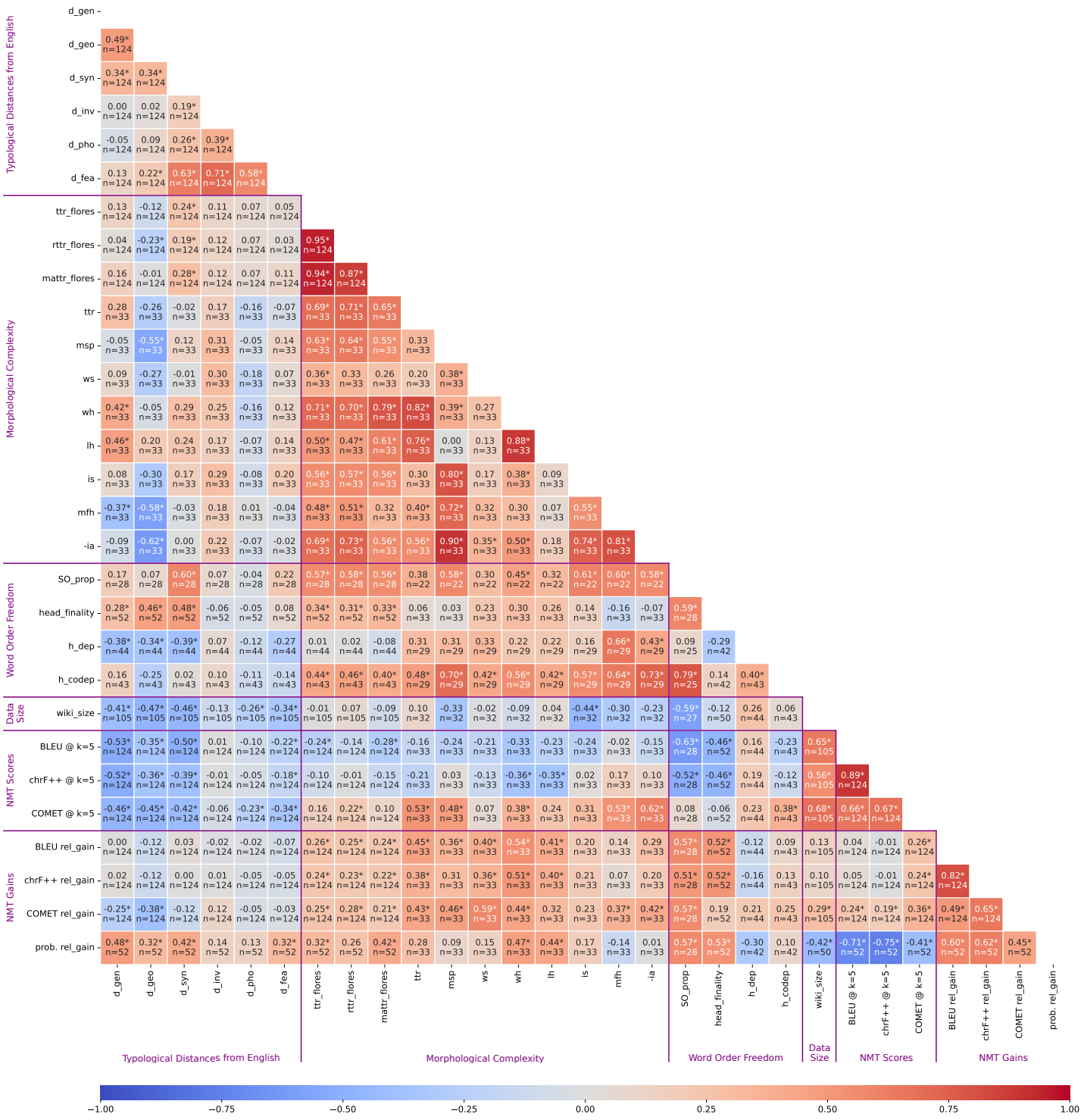

Figure 18: Spearman correlations between continuous language properties, translation quality scores at beam size $k = 5$ (BLEU, chrF++, COMET), and relative gains (BLEU, chrF++, COMET, generation probability). Language sample sizes are indicated under each correlation coefficient.

F Correlations with Continuous Language Properties on the Same Language Sample

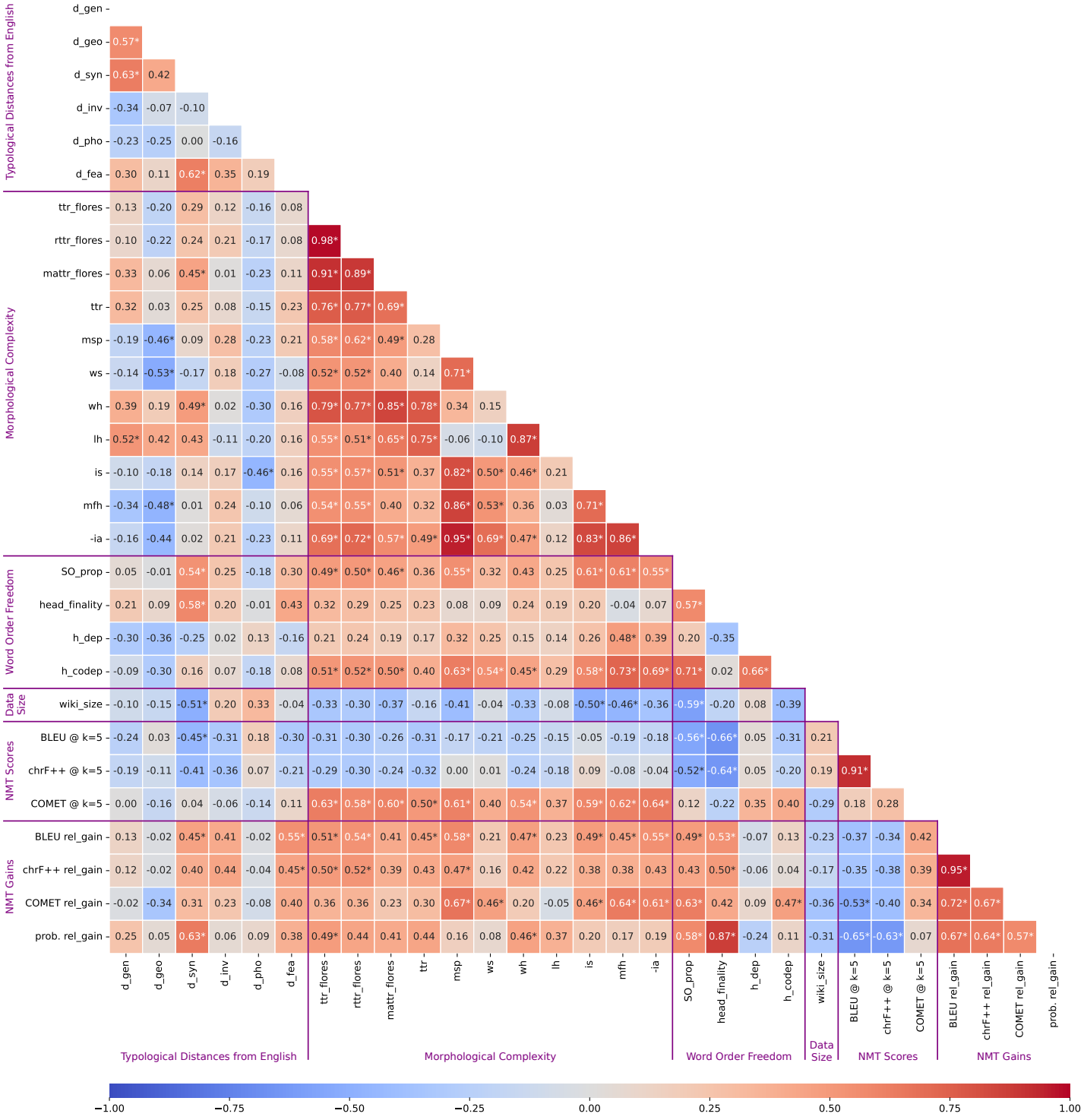

Figure 19: Spearman correlations between continuous language properties, translation quality scores at beam size $k = 5$ (BLEU, chrF++, COMET), and relative gains (BLEU, chrF++, COMET, generation probability). Correlations are calculated on the same language sample of size 20.