A Cross-Lingual Perspective on Neural Machine Translation Difficulty

Esther Ploeger[1], **Johannes Bjerva**[1], **Jörg Tiedemann**[2], **Robert Östling**[3]
[1]Aalborg University [2]University of Helsinki [3]Stockholm University
{espl,jbjerva}@cs.aau.dk jorg.tiedemann@helsinki.fi robert@ling.su.se

Abstract

Intuitively, machine translation (MT) between closely related languages, such as Swedish and Danish, is easier than MT between more distant pairs, such as Finnish and Danish. Yet, the notions of 'closely related' languages and 'easier' translation have so far remained underspecified. Moreover, in the context of neural MT, this assumption was almost exclusively evaluated in scenarios where English was either the source or target language, leaving a broader cross-lingual view unexplored. In this work, we present a controlled study of language similarity and neural MT difficulty for 56 European translation directions. We test a range of language similarity metrics, some of which are reasonable predictors of MT difficulty. On a text-level, we reassess previously introduced indicators of MT difficulty, and find that they are not well-suited to our domain, or neural MT more generally. Ultimately, we hope that this work inspires further cross-lingual investigations of neural MT difficulty.

1 Introduction

In neural machine translation (NMT), the choice of the language pair(s) under study is often heavily influenced by data availability. But how does the choice of language pair influence the difficulty of the translation task? This question has been studied extensively for statistical MT (e.g., Koehn, 2005; Birch et al., 2008; Paul et al., 2009); these works are still frequently cited to explain linguistic disparity in MT (e.g., Rowe et al., 2025). But to date, similar studies on NMT have been scarce. A notable exception was presented by Bugliarello et al., 2020, who quantified translation difficulty for Transformer models (Vaswani et al., 2017). Yet, whereas the aforementioned studies in statistical MT focused on more than 100 translation directions, the NMT study is limited to English-centric translation scenarios: settings where English is either the source or the target language. As a result, little is known about cross-lingual NMT difficulty in a broader sense. The first goal of this work is to address this knowledge gap, by expanding the scope of analyzed language pairs (e.g., Figure 1).

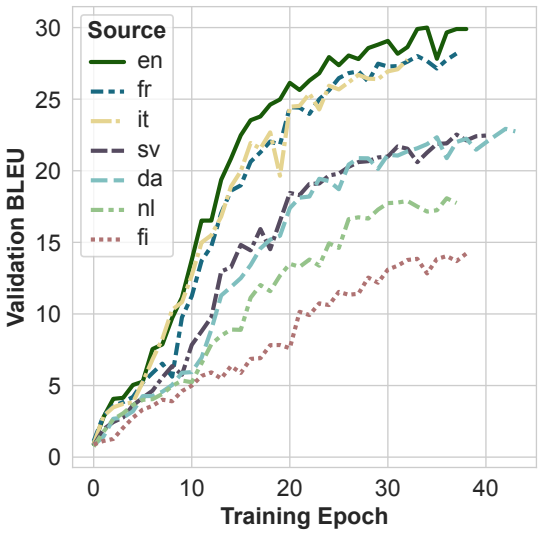


Figure 1: Training curves for MT into Portuguese; Italian (*it*) and French (*fr*) are 'easier' source languages to learn than e.g. Danish (*da*) and Swedish (*sv*).

Beyond advances in MT architectures, new approaches for quantifying language similarity have also emerged. The `lang2vec` (Littell et al., 2017) toolkit is one of the most popular typological similarity tools in natural language processing (NLP), despite major issues of data sparsity and irreproducibility (Khan et al., 2025). Typological research has resulted in newer databases, such as Grambank (Skirgård et al., 2023), which was developed with computational applications in mind (Haynie et al., 2023). This data was shown to be informative for NLP tasks, for example in the case of part-of-speech tagging (Rice et al., 2025), but there has not yet been a study on relating this new source of information to MT difficulty. The second objective of this work is therefore to systematically compare these typological databases in the context of MT.

Proceedings of the Tenth Workshop on Machine Translation (WMT), pages 340–354
November 8–9, 2025.

The third aim of this work is to bridge two perspectives on MT difficulty. Separately from language-level approaches to performance prediction (e.g., Birch et al., 2008; Bugliarello et al., 2020), MT difficulty has long been approached from the perspective of *text-level* difficulty (e.g., Underwood and Jongejan, 2001; Bernth and Gdaniec, 2001; O'Brien, 2004). It is not clear how these two perspectives influence each other. Are typical indicators of translation difficulty, such as sentence length, perhaps less problematic when the source and target language are more similar? In summary, this work aims to address the following three research questions:

- **RQ1:** Are there differences in NMT difficulty, beyond English-centric language pairs?
- **RQ2:** Which measures of language similarity are informative for predicting NMT performance?
- **RQ3:** How language pair-dependent are text-level translatability indicators?

Our study additionally aims to address a methodological issue in similar previous work. The inclusion of translated text in MT evaluation sets can artificially inflate results (Zhang and Toral, 2019; Graham et al., 2020), but this variable was not controlled for in relevant previous studies. To control as much as possible for variations beyond language similarity (data size, domain, topic, genre, proportion of translated text), we construct a new fully multi-parallel dataset. Because of these strict constraints and data availability, we limit our scope to 8 European languages. We train and analyze bilingual NMT models for each of the 56 resulting translation directions.

2 Related work

2.1 Machine Translation Difficulty

Estimating the difficulty of a translation task (*translatability*) from text has long been an important research direction in translation studies (Sun, 2015), for both manual and automatic translation (Bernth and Gdaniec, 2001). For MT, assessing the difficulty of translation tasks has had two prominent applications in the translation industry. The first is to distinguish samples that are suitable for MT from more difficult samples that likely require manual translation (Underwood and Jongejan, 2001). This line of research was especially popular when the performance of MT systems was much poorer than it is now. Still, more recently, Fernicola et al. (2023) showed that NMT suitability can also be predicted from source texts with reasonable accuracy. The second application has been to inform *controlled language* writing (e.g., Miyata et al., 2015). Given indicators of text characteristics that pose issues to MT, writers can choose to avoid these, to make their texts easier for MT systems to translate.

Text-level Approaches Certain text-level characteristics have been associated with MT difficulty. Examples include personal pronouns, post-modifying adjective phrases, ellipsis and very long or short sentences (Bernth and Gdaniec, 2001). Such indicators are also called *translatability indicators* (TIs). Underwood and Jongejan (2001) distinguish *general* TIs from *system-specific* ones. O'Brien (2004) writes that "it has been acknowledged that some TIs are more problematic for certain language pairs and directions than others". In the context of neural MT, however, the relevance of these TIs has not been investigated, and the influence of the language pair is also unclear. We explore a number of general TIs across multiple language pairs in Section 7.

Language-level Approaches The relationship between language similarity and MT difficulty has been a longstanding area of interest in MT research. Early work on statistical MT by Koehn (2005) and Birch et al. (2008) explored translation challenges across language pairs, using BLEU scores (Papineni et al., 2002) as a primary evaluation metric. These studies showed that translation performance tends to correlate with linguistic proximity; historically closely related languages generally yielded higher BLEU scores. Bugliarello et al. (2020) presented the first similar study on neural MT, noting that BLEU scores are only comparable on test sets in the same target language. They separated source- and target-language difficulty explicitly, through measuring cross-mutual information, with evaluations on 40 English-centric language pairs.

2.2 Language Similarity in MT

How to best measure language similarity to inform NLP research is an open question. Blaschke et al. (2025) conducted a large-scale study on distance measures for cross-lingual transfer in three NLP tasks: dependency parsing, part-of-speech tagging and topic classification. They found that the definition of linguistic similarity is an important factor

for cross-lingual transfer success, and the most effective similarity measure is dependent on the downstream task. Since MT was not investigated, it is not clear which kind of similarity measure is most informative for this task. Within MT research, language similarity has mostly been of interest for improving transfer learning performance (e.g., Lin et al., 2019; Oncevay et al., 2020; Fekete et al., 2025). Factors beyond language similarity can play a major role in such investigations, such as data availability and domain similarity (Khiu et al., 2024). By contrast, we are interested specifically in the difficulty of the translation task by itself, all other factors being as equal as possible.

3 Constructing a Multi-Parallel Corpus

Our aim is to investigate the difficulty of translation tasks, while controlling for factors other than language similarity. We first establish three criteria that should be met for comparable cross-lingual MT studies (Section 3.1). We then describe how these were accounted for in the creation of our dataset (Section 3.2), and how these restrictions impact the diversity of our language selection (Section 3.3).

3.1 Criteria

If MT models for different language pairs are trained on different datasets, then differences in results could be attributed to that instead of linguistic divergences. So, to rule out the effects of dataset size and domain differences, the dataset should be fully multi-parallel across all included languages. Secondly, within the multi-parallel corpus, the distribution of 'original text' should be equal across languages. A well-described problem in the evaluation of MT systems is that the presence of translated data in the evaluation set can inflate performance assessments (Zhang and Toral, 2019; Graham et al., 2020).[1] Ideally, test sets should contain text originally written in a language, to not exhibit 'translation artefacts'.[2] However, since such a dataset should also be completely multi-parallel, this is not possible for more than one language. We argue that a dataset should therefore ensure equal *proportions* of translated text in the multi-parallel test set. In addition to consistent proportions across languages, it should be the same across training and testing sets, to avoid training on one text type and evaluating another. Lastly, to provide a cross-lingual perspective beyond English, the dataset should contain translation pairs that do not include English as either source or target language. In summary, we establish three criteria:

1. Full multi-parallelism
2. Equal translated text distribution
3. Beyond English-centric MT

To the best of our knowledge, a ready-to-use dataset that satisfies these three criteria does not yet exist.

3.2 Dataset Creation

Multiple popular multi-parallel datasets with broad language coverage exist (e.g., Parallel Bible Corpus, Mayer and Cysouw, 2014; OpenSubtitles, Lison and Tiedemann, 2016). Yet, these datasets do not contain information on the original languages of the data. The CoStEP corpus (Graën et al., 2014) includes speaker turns from the European Parliament, cleaned and aligned across languages, with original-language annotations. In Figure 2, we illustrate the steps for creating a dataset that satisfies the criteria from Section 3.1. Starting from the CoStEP corpus (1) we first extract all bilingually aligned speaker turns (2). Since speaker turns vary in length (some are very long while others are very short), we split the turns into sentences (3) using the `sentence-splitter` toolkit.[3] Then, we re-align the bilingual parallel sentences (4) using `hunalign` (Varga et al., 2008). Note that we refrain from using embedding-based models like VecAlign (Thompson and Koehn, 2019), to avoid potential cross-lingual biases (English-centric pairs having higher-quality alignments, because English is very well-represented in pre-trained models). Next, given these aligned sentences per original language, we assess for which languages we have enough multi-parallel data for MT training and evaluation (5). We find that there are eight languages that have both high coverage and meta-information available in the typological database Grambank (Skirgård et al., 2023), which we use because of its suitability for computational applications (Haynie et al., 2023): Danish, Dutch, English, Finnish, French, Italian, Swedish and Portuguese. We select all samples for which we have translations in all languages and make a multi-parallel dataset with equal original-language proportions (6). Lastly, we divide the text

[1] In the MT literature, this is commonly described as the 'translationese effect', while this term is not uncontroversial (Jimenez-Crespo, 2023).

[2] We empirically verify this for our dataset in Appendix A.

[3] https://github.com/mediacloud/sentence-splitter

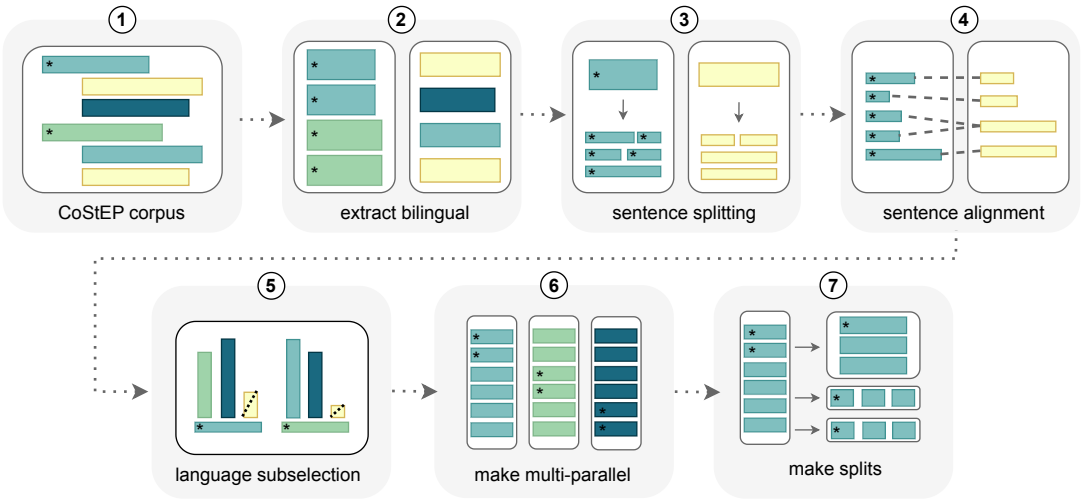


Figure 2: Schematic overview of the dataset creation process, starting from the CoStEP corpus, where we ensure an equal original text distribution (indicated schematically with an asterisk) among the included languages.

into splits for MT training, validation and testing such that the original text proportion remains equal among languages (7). We reserve 800 samples for validation (100 original per language) and 1,600 for testing (200 original per language). We use the remaining lines (77,941 per language) for training.

The number of space-separated tokens per language (as obtained by running the `wc -w` command) is in Table 1. Here, it stands out that Finnish contains a lower number of 'words' than the other languages, as it exhibits more morphologically complexity than the others.

Language	Train	Valid	Test
Danish (da)	1.9M	20.0K	39.3K
Dutch (nl)	2.1M	21.8K	43.0K
English (en)	2.1M	22.0K	42.6K
Finnish (fi)	1.4M	14.8K	28.9K
French (fr)	2.2M	23.0K	44.6K
Italian (it)	2.0M	21.2K	40.9K
Swedish (sv)	1.9M	22.2K	43.0K
Portuguese (pt)	2.1M	19.8K	38.8K

Table 1: Number of 'words' per language in the multi-parallel dataset.

Our strict filtering criteria result in a controlled, but small dataset which is not necessarily representative of state-of-the-art high-resource NMT more generally. These strict controls are necessary for our study, and we retrieve reliable results within our set-up, but we cannot make claims regarding the broad generalizability of these results.

3.3 Typological Diversity

Our language selection contains languages from three genera (Germanic: Danish, Swedish, Dutch, English; Romance: Italian, French, Portuguese; and Finnic: Finnish). To gain a more fine-grained image beyond genealogical similarities, we look into the typological similarity of these languages using the Grambank database (Skirgård et al., 2023).[4] Figure 3 shows a PCA plot of the Grambank feature vectors, showing our language selection is not representative of typological diversity generally. While this limited language diversity, also in terms of writing systems, means we cannot make any claims about what makes NMT difficult *in general*, it is appropriate for the objective of providing cross-lingual (but not universal) insights.

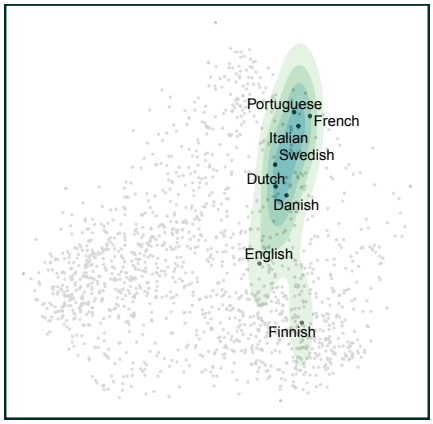


Figure 3: PCA plot from Grambank features, illustrating the (limited) typological diversity in this study.

[4]Typological diversity using *typdiv* (Ploeger et al., 2025, default settings): MPD: 0.51, FVO: 0.79, FVI: 0.77, H: 0.42.

T (→) / S (↓)	da		nl		en		fi		fr		it		sv		pt	
	chrF	*BS*	*chrF*	*BS*	*chrF*	*BS*	*chrF*	*BS*	*chrF*	*BS*	*chrF*	*BS*	*chrF*	*BS*	*chrF*	*BS*
da	–	–	69.9	**0.81**	74.3	0.85	67.3	0.78	72.2	0.82	70.1	0.80	**75.2**	**0.85**	71.4	0.82
nl	69.8	0.82	–	–	**69.8**	**0.82**	60.9	**0.72**	66.9	0.77	**62.8**	**0.76**	**67.4**	0.80	67.5	0.80
en	74.6	0.84	**71.3**	**0.81**	–	–	**70.0**	**0.80**	75.4	**0.84**	72.6	**0.83**	74.0	0.84	**75.6**	**0.85**
fi	**68.1**	**0.79**	**62.8**	**0.73**	**69.8**	0.83	–	–	**58.2**	**0.73**	68.1	0.77	68.0	0.80	**66.8**	**0.78**
fr	70.9	0.82	67.5	0.79	74.0	0.85	**60.5**	0.73	–	–	73.9	**0.83**	70.3	0.81	74.7	0.84
it	68.8	0.80	67.7	0.77	73.3	0.84	65.3	0.76	72.0	0.82	–	–	67.8	**0.79**	73.9	0.83
sv	**75.1**	**0.85**	70.8	0.80	75.0	**0.86**	68.5	0.79	72.3	0.82	71.2	0.80	–	–	71.3	0.82
pt	71.1	0.82	67.8	0.79	**76.3**	**0.86**	66.9	0.77	**75.8**	**0.84**	**74.8**	**0.83**	72.3	0.82	–	–

Table 2: chrF2 and BERTScore for each target language (columns), per source language (rows). **Highest** and **lowest** scores per metric are highlighted for each target language.

4 Machine Translation Models

For each language separately, we train a Unigram (Kudo, 2018) subword segmenter on the training corpus, with vocabulary size 8,000, to tokenize the text. We then train a separate bilingual MT model for each of the 56 language pairs in our dataset. We choose to train small models from scratch, rather than leveraging pre-trained NMT models or LLMs, to avoid cross-lingually unfair pre-training distributions. The goal is not to create optimally functioning systems, which would require intensive parameter tuning, but rather to compare systems that learn to translate between different language pairs under otherwise identical circumstances. We use the Fairseq toolkit (Ott et al., 2019) to train MT models with a standard Transformer (Vaswani et al., 2017) architecture, consisting of 6 encoder and 6 decoder layers. We use a learning rate of 0.0001 and a dropout rate of 0.2. We use a cross entropy with a label smoothing of 0.2. and a maximum of 4,000 tokens per training batch. As a best checkpoint metric we use cross-entropy, with patience 5. No model took more than 50 epochs to converge.

5 Cross-lingual Translatability (RQ1)

5.1 Measuring Translation Difficulty

We approach translation difficulty from two perspectives: MT performance estimates, and the computational resources required for training.

Translation Accuracy Inspired by works from human translatability, e.g., Vanroy et al. (2019) and Hale and Campbell (2002), as well as MT approaches (Koehn, 2005; Birch et al., 2008), we deem a translation task difficult if it triggers errors. The intuition is that language pairs that are more difficult to translate lead to lower translation accuracy. As a widely spread measure of translation quality, we report chrF2 (Popović, 2015) scores (Table 2), which were previously shown to be robust against varying degrees of morphological complexity (Popović, 2016). This is a surface-level metric, based on a reference translation. We use SacreBLEU (Post, 2018) to calculate this.[5] One shortcoming of these measures, is that wording differences in the human reference translations can influence the results. This is why we also ran an embedding-based evaluation. We do not use reference-free metrics such as COMET (Rei et al., 2020), since the unequal training data in language embeddings may introduce cross-lingual unfairness in evaluation. Instead, we compare (monolingually) the MT hypothesis against the reference using BERTScore (Zhang et al., 2019), which renders it comparable per target language.

Computational Resources Inspired by works on human translation and post-editing that measured translation difficulty through cognitive and temporal effort (e.g., Campbell, 1999; Beinborn, 2010), we deem a translation direction more difficult if it requires more resources. We examine "machine effort" through a proxy: training dynamics. Beyond the overall performance, we record the loss and BLEU on the validation set per epoch as measures of effort, and compare this across source languages for MT into the same target language.

5.2 Difficulty as Translation Accuracy

Overall Performance We list the chrF2 and BERTScore per target language in Table 2. We exclusively compare the source languages per target language, since direct comparison across different target languages' test sets may be unfair (Bugliarello et al., 2020). Higher scores indicate higher translation accuracy, and thereby suggest lower MT difficulty. While the absolute scores are

[5] signature: `nrefs:1|case:mixed|eff:yes|nc:2|nw:0|space:no|version:2.5.0"`

low compared to the state-of-the-art, we observe that, indeed, there are language-level translatability differences. In other words, it is not the case that the same source language is the easiest for all target languages. For example, the easiest source language for MT into Danish is Swedish, while the easiest for Portuguese is French. Finnish, as the only language outside of the Indo-European language family in this study, also stands out in terms of MT performance. For most target languages (Danish, Dutch, English, French, Portuguese), it is among the most difficult. Moreover, we see that Dutch is a difficult source language for many target languages. These results indicate that, also in scenarios beyond English-centric MT, translation difficulty varies translation per direction. We observe some intuitive patterns, such as Danish↔Swedish being a relatively easy translation direction. In Section 6, we assess the connection with language similarity more systematically.

5.3 Difficulty as Effort

Training Epochs for Validation BLEU Analogous to "the extent to which cognitive resources are consumed by a translation task for a translator to meet objective and subjective performance criteria." (Sun, 2015), we assess how many computational resources ("machine effort") are required for an objective performance criterion. Specifically, we assess how many training epochs are needed to reach 15 BLEU on the validation set. While this threshold is somewhat arbitrary, we note that the full training progressions are shown in Appendix C. While this metric cannot be compared across languages, due to its reliance on word boundaries, it provides an intuitive and easy to interpret measure of difficulty. Here, a lower number (of epochs) indicates that fewer computational resources are required, signaling that the translation direction is easier. Table 3 shows the results. Firstly, we observe that this analysis reveals different nuances than when approaching difficulty as accuracy. Namely, English is an 'easy source language' for more languages here. As such, generalizing English-centric translation results may lead to overestimating NMT performance. This is especially noteworthy, as most previous studies (and MT in general) center research claims around this language. An explanation for this, relative to our dataset, could be that the European Parliament sometimes uses relay translations, i.e. manual translations are produced through English as a pivot language. Unfortunately, it is not possible to control for this variable.

Furthermore, it is interesting to analyze the translation directions which never reach 15 validation BLEU, indicated by the ∅ symbol. Especially Finnish stands out here, as it only reaches the threshold from English. Additionally, none of the Romance target languages reach 15 BLEU when Finnish is the source language. These are patterns that follow intuitive language similarity ideas, which are further explored in Section 6.

Loss Slopes While intuitive, the 15 BLEU threshold has shortcomings. For one, it is only comparable per target language, due to its reliance on word boundaries. Here, we analyze a more cross-lingually comparable metric: the loss slope between two consistent, pre-defined points in the training process (5^{th} epoch vs. 25^{th} epoch). Starting from epoch 5 gives the decoders a chance to learn a very basic language model, ahead of actually learning to translate. In Appendix D, the full slopes are visualized. Higher numbers indicate steeper slopes, which indicates faster learning, and

S (↓)	da	nl	en	fi	fr	it	sv	pt
da	–	19	12	∅	15	29	**12**	19
nl	19	–	16	∅	32	∅	29	27
en	**12**	**16**	–	**32**	**11**	16	14	**12**
fi	31	∅	18	–	∅	∅	34	∅
fr	19	25	12	∅	–	15	20	14
it	30	∅	12	∅	14	–	35	13
sv	13	22	12	∅	17	26	–	18
pt	18	29	**9**	∅	12	**14**	24	–

Table 3: Number of training epochs needed to reach at least 15 BLEU on the validation set per translation direction, lowest number per target language **highlighted**.

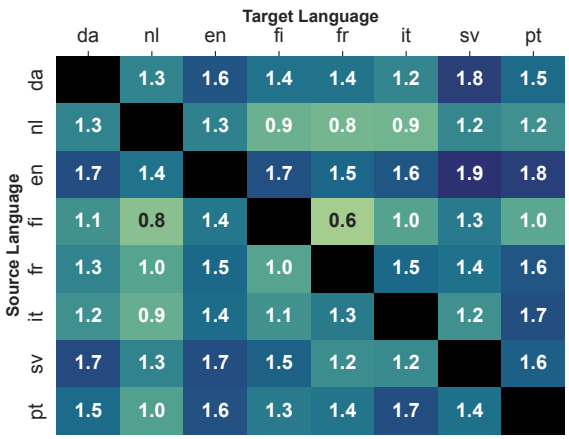
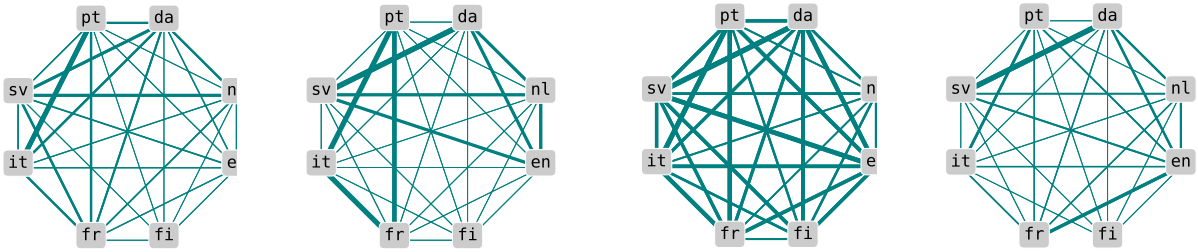


Figure 4: Δ losses (5^{th} vs. 25^{th} epoch) in MT training.



(a) Grambank similarity (b) Genealogical similarity (c) Word order similarity (d) Subword overlap

Figure 5: Visual comparison of language-based (a, b) and data-based (c, d) language similarity, where line thickness illustrates similarity. For example, subword overlap between Danish and Swedish is particularly high (d).

thus suggests lower translation difficulty. In Figure 4, the steepest slopes (darker colors) are mostly found for English. This is in line with our findings from using the validation BLEU threshold, where English also was commonly 'easy'. Additionally, it stands out that Dutch is relatively difficult. One possible explanation for this is found in Figure 5c, where Dutch generally has the most dissimilar word order to the other languages. We verify this relationship in Section 6.

All in all, we conclude that there are indeed language pair-level differences in NMT translation difficulty. This indicates that the vague notion of 'easier translation' can be operationalized more systematically. Results on translation difficulty, both from the angles of accuracy and effort, show intuitive patterns with regard to language similarity patterns, which are further examined in the next section.

6 Language Similarity Analysis (RQ2)

6.1 Measuring Language Similarity

Any effort to reduce languages and their similarities to single floating point numbers risks being simplistic. Yet, to enable systematic comparison, we need to compare the different aspects of similarity on the same scale. We compare two categories of metrics: those derived on a language-level, and dataset-specific measures.

Language-Level Metrics A first approach is to determine language similarity based on expert annotations, such as phylogenies and typological features. The popular `lang2vec` toolkit provides six categories of language distances (geographical, genealogical, syntactic, phonological, featural and inventory-based), based on the URIEL database (Littell et al., 2017). Given these distances, d, the resulting language similarity is then defined as $1 - d$. We compare this with a Grambank-based (Skirgård et al., 2023) measure, as proposed in Ploeger et al. (2025): the Euclidean distance between Grambank's morphosyntactic feature vectors, accounting for missing values. We again subtract this from 1, to compute a similarity score.

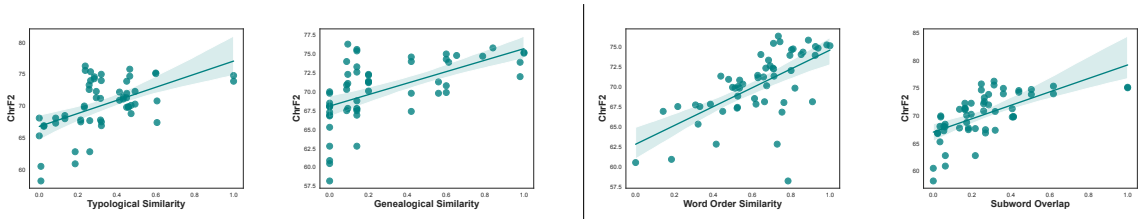
Text-Driven Metrics As a more data-specific measure of *syntactic* similarity, we use the relative amount of word reordering between two languages, calculated over the test set using Eflomal (Östling and Tiedemann, 2016). Since reordering scores are directional, we take the average over both directions to retrieve a single similarity score, in line with the language-level metrics. To go from reordering to word order similarity, we again subtract it from 1. As a *lexical* measure of data-driven similarity, we take the proportion of overlapping subwords from the tokenized texts, relative to the sum of the number of subwords of both languages in the test set. Finally, we apply MinMax scaling.

6.2 Results and Analysis

Table 4 shows the Pearson correlation coefficients for each of the language similarity metrics with

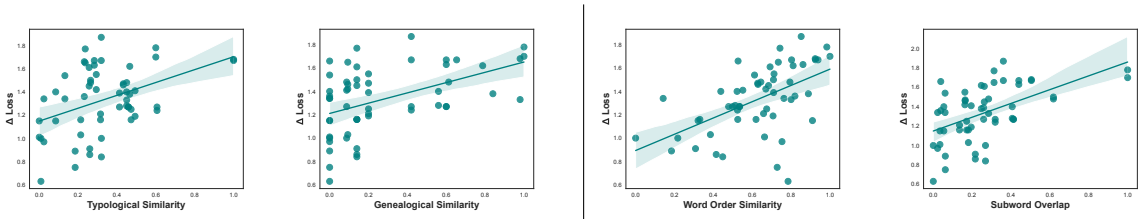
Similarity Measure	chrF2	Δ Loss
12v (geographic)	0.41*	0.26
12v (genetic)	**0.56***	**0.47***
12v (syntactic)	0.46*	0.43*
12v (featural)	-0.08	0.10
12v (inventory)	-0.18	-0.22
12v (phonological)	-0.04	-0.04
Grambank	**0.52***	**0.41***
Word reordering	**0.63***	**0.54***
Subword overlap	**0.63***	0.53*

Table 4: Pearson correlation coefficients between similarity metric and performance. Statistically significant values ($p < 0.005$) are indicated with an asterisk.



(a) Typological similarity (gb) ($r \approx \mathbf{0.52}, p < 0.005$)
(b) Genealogical similarity ($r \approx \mathbf{0.56}, p < 0.005$)
(c) Word order similarity ($r \approx \mathbf{0.63}, p < 0.005$)
(d) Subword overlap ($r \approx \mathbf{0.63}, p < 0.005$)

Figure 6: Correlation between linguistic similarity measures and chrF2 scores.



(a) Typological similarity (gb) ($r \approx \mathbf{0.41}, p < 0.005$)
(b) Genealogical similarity ($r \approx \mathbf{0.47}, p < 0.005$)
(c) Word order similarity ($r \approx \mathbf{0.54}, p < 0.005$)
(d) Subword overlap ($r \approx \mathbf{0.53}, p < 0.005$)

Figure 7: Correlation between linguistic similarity measures and Δ training loss.

chrF2 and Δ loss. From the table, it follows that morphosyntactic distance based on Grambank is a better predictor of MT difficulty than `lang2vec`'s syntactic distance. Still, `lang2vec`'s genetic distance and the data-driven measures yield higher correlations. We now examine these four metrics with the strongest correlations (Table 4, bold font) in more detail. These language similarity metrics are correlated positively with chrF2 and the Δ loss. That is, the more similar the languages in the translation direction, the 'easier' it is to translate between them. All correlations are statistically significant, with $p < 0.005$. Interestingly, the correlation coefficients of the data-driven metrics (c, d) are higher than those of the language-level metrics (a, b). This indicates that tailoring language similarity measures to the dataset under study may be beneficial for retrieving accurate difficulty predictions.

To gain an insight into why these correlations differ, we qualitatively assess the language similarities. Figure 5 illustrates the pairwise distances per metric. Absolute numbers are given in Appendix B. In the Grambank-based measure (a), the link between Italian and Portuguese is especially pronounced, while with the genealogical metric (b), we for example see a strong similarity between Portuguese, Italian and French. In terms of the data-driven measures, what stands out in (c) is that Dutch has relatively low word order similarity scores with all other languages, and that the subword overlap (d) between Danish and Swedish is especially prominent. This influences the correlation coefficients. As visualized in Figure 6 (for difficulty as accuracy; chrF2) and Figure 7 (for difficulty as effort; Δ loss), outliers –Portuguese and Italian, Danish and Swedish– influence the correlation coefficients, notably in Figure 6/7a and 6/7d respectively.

7 Translatability Indicators (RQ3)

As mentioned in Section 2, approaching translation difficulty from text on a data-level has been an important research direction in the past, mostly in statistical MT (Bernth and Gdaniec, 2001). Which textual characteristics make MT difficult? We revisit source-text indicators of translation difficulty that were proposed in previous work (Underwood and Jongejan, 2001), and contribute an investigation of these indicators in the context of neural MT.

7.1 Identifying TIs

We reassess 'general indicators' of machine translatability from Underwood and Jongejan (2001), listed in Table 5. For this case study, we take a closer look at translation with English as a source, since these TIs were formulated for that language specifically. We obtain the TIs through POS tags and dependency relations, retrieved through Trankit (Nguyen et al., 2021), with the default XLM-RoBERTa as the underlying model. We follow the heuristics defined in Underwood and Jongejan

Translatability Indicator	#	da	nl	fi	fr	it	sv	pt
No verb	11	+13.85	+12.71	+14.79	+19.30	+12.67	+13.13	+13.11
No finite verb	13	+12.92	+10.04	+12.38	+14.84	+11.43	+12.31	+13.28
Long (> 25 words)	713	**-0.21**	**-0.14**	**-0.36**	+0.52	+0.43	+0.01	+0.30
Short (< 3 words)	4	+19.7	+24.98	+27.71	+32.75	+0.68	+18.15	+25.45
$\geq$ 1 nominal compound	520	**-0.09**	**-0.13**	+0.06	+0.55	+0.56	+0.30	+0.38
Multiple coordination	759	**-0.18**	**-0.30**	+0.14	**-0.20**	+0.48	+0.43	**-0.35**

Table 5: Delta between the average chrF2 score for the TI lines, and the average non-TI chrF2. Negative values (bold font) imply increased translation difficulty. # Indicates how often the TI appears in our English test set.

(2001) for detecting the translatability indicators. Our implementation is as follows:

Missing Verbs A line does not have a verb if it contains no token with an `AUX` or `VERB` tag. An example from our dataset is: *"All well and good."* A line does not contain a finite verb if there is no verb with `VerbForm=Fin` present. An example: *"But what about all the other protective considerations listed in Article 13 of the Treaty?"*.

Sequence Length Long and short sentences are determined through the number of words: long sentences contain more than 25 words, while short sentences contain less than 3. While long sequences are ubiquitous in our dataset, an example of a (much rarer) short line is: *"No one!"*.

$\geq$1 Nominal Compound A sentence contains a nominal compound if it contains a `NOUN` that has the dependency relation of `compound`, for example: *"But we also need better regulation and principles for future EU legislation when it comes to motor vehicles."*.

Multiple Coordination Lastly, we detect multiple coordination if the sentence contains more than one `SCONJ` and/or `CCONJ`. An example: (e.g. *"That is why, of course, donor cards should be voluntary, and the same applies to the European donor card, which we intend to introduce in our action plan."*).

We first identify which samples in the English test set contain these markers. Then, we calculate the chrF2 score per sample in the test set, and compare the average score of the TIs with the average score of the non-TIs. We expect that if a certain group of sequences (TIs) is more difficult to translate, it yields a lower chrF2 score than the average non-TI chrF2 score per sample. Table 5 shows the difference between the average TI score, minus the average non-TI score. If a table cell contains a negative number (bold), this indicates that the translation of the TI samples obtained a lower chrF2 score than the others, indicating potential translatability issues.

7.2 Results and Analysis

We find no consistent patterns indicating that the defined TIs are more difficult for MT than non-TIs. For some TIs, there is too little evidence to base any robust conclusion on (e.g. only 4 lines contain $<$ 3 words in our dataset) and verb-less lines are rare. This may differ per studied domain. Secondly, it could be that common indicators of translation difficulties do not apply to our systems, because the training data is domain-specific. For example, nominal compounds occur relatively frequently in our dataset, as a result of the domain (e.g. *"member states"*, *"employment plans"*, *"terrorist list"*, *"labour taxes"*). If a model is trained on many of these, this may result in better capabilities to deal with such features. Another possible reason is that the paradigm shift to neural MT has weakened the impact of formerly informative TIs. For example, Transformers' (Vaswani et al., 2017) cross-attention implies that long-range dependencies have become less problematic.

Beyond general challenge sets, future work could be dedicated to defining TIs specifically for neural MT (cf. Bisazza et al., 2021): are there common translatability issues for state-of-the-art architectures, and consistently across datasets? Special care could be taken to make these more cross-lingually comparable: for example, taking the number of words as length indicator (Underwood and Jongejan, 2001) is dependent on the morphological complexity of a language.

8 Conclusions

In this work, we aimed to operationalize language similarity and translation difficulty in the context of neural MT. We control for confounding factors from previous work, and use the resulting dataset to answer three research questions about how language similarity affects translatability. In summary, we find the following. Firstly, there are language pair-level differences in NMT difficulty in our experiments, beyond English-centric scenarios (RQ1). Moreover, NMT difficulty can be predicted from language pair similarity with reasonable success, with syntactic and genetic measures of similarity. Text-driven metrics, tailored to the dataset, are even more informative (RQ2). Lastly, we found that text-level indicators of MT difficulty from previous work were not suitable for our dataset or evaluation set-up (RQ3).

Our models achieve limited performance (approx. 60-70 chrF2). This is because we train models on relatively small datasets. This performance is far below the state-of-the-art of these European language pairs. Still, it provides interesting insights. We show that some language pairs are reliably and consistently more difficult (e.g. much lower chrF2 scores) than others, under the same, controlled circumstances. These results are stable, as demonstrated by the steady train loss decrease in Appendix D. Furthermore, the performance differences between translation directions are substantial and predictable ($p<0.005$, Fig. 6, 7), showing that even in a limited setting, consistent results emerge. Although these strict controls are necessary for our controlled study and we retrieve reliable results in our case study, we cannot make claims regarding the broad generalizability of these results. Generalizability is a challenging topic in MT more generally, as it is unclear whether scaling approaches will always solve previously encountered issues.

These findings open up various possible applications and future research directions. For example, future work could investigate to what extent MT difficulty is influenced by tokenization strategies. In our experiments, we kept the tokenizer and vocabulary size consistent across languages, but variations could yield different results. In downstream scenarios, systematic notions of language similarity could be used to select pivot languages, especially in data-scarce scenarios. On a text-level, we showed that identifying new TIs, relevant to neural MT, may be an interesting research direction.

Limitations

Several limitations of this work should be noted. Firstly, due to the strict constraints (full multi-parallelism, equal distribution of original text), the typological diversity of the languages in our study is limited. Our findings may not generalize to other translation directions. For the same reason, we only trained and evaluated in one domain; this cannot be assumed to generalize directly either. Note that previous studies on translation difficulty (Koehn, 2005; Birch et al., 2008; Bugliarello et al., 2020) were also limited to this domain. Despite these strict constraints, 100 percent clean data is not guaranteed. For one, automatic alignments may still include noise. Also, the European Parliament sometimes includes relay translations, meaning that certain translated texts may have been translated 'through English', which could impact the results. Yet, as this information is not available in CoStEP or elsewhere, we cannot control for this. While the Transformer architecture is representative for neural MT, our study only uses this one neural architecture. For broader insights into neural MT as a whole, more approaches could be investigated. Furthermore, the language similarity measures that we evaluate are not exhaustive; for example, semantic language similarity (for example in the form of colexification) was not taken into account. Lastly, no human evaluation was included. This choice was made to ensure cross-lingually consistent comparisons, but it could still have yielded important novel insights, for example comparing human translatability and NMT features (cf. Lim et al., 2024).

Acknowledgements

We thank the AAU-NLP group, in particular Mike Zhang, for proofreading earlier versions of this article. EP and JB are funded by the Carlsberg Foundation, under the *Semper Ardens: Accelerate* Programme (project nr. CF21-0454). EP was further supported by a travel grant from the Otto Mønsteds Fond.

References

Lisa Beinborn. 2010. Post-editing of statistical machine translation: A crosslinguistic analysis of the temporal, technical and cognitive effort. Master's thesis, Saarland University.

Arendse Bernth and Claudia Gdaniec. 2001. Mtranslatability. *Machine translation*, 16:175–218.

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

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

Verena Blaschke, Masha Fedzechkina, and Maartje Ter Hoeve. 2025. Analyzing the effect of linguistic similarity on cross-lingual transfer: Tasks and experimental setups matter. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8653–8684, Vienna, Austria. Association for Computational Linguistics.

Emanuele Bugliarello, Sabrina J. Mielke, Antonios Anastasopoulos, Ryan Cotterell, and Naoaki Okazaki. 2020. It's easier to translate out of English than into it: Measuring neural translation difficulty by cross-mutual information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1640–1649, Online. Association for Computational Linguistics.

Stuart Campbell. 1999. A cognitive approach to source text difficulty in translation. *Target. International Journal of Translation Studies*, 11(1):33–63.

Marcell Fekete, Nathaniel Romney Robinson, Ernests Lavrinovics, Djeride Jean-Baptiste, Raj Dabre, Johannes Bjerva, and Heather Lent. 2025. Limited-resource adapters are regularizers, not linguists. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 222–237, Vienna, Austria. Association for Computational Linguistics.

Francesco Fernicola, Silvia Bernardini, Federico Garcea, Adriano Ferraresi, and Alberto Barrón-Cedeño. 2023. Return to the source: Assessing machine translation suitability. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 79–89, Tampere, Finland. European Association for Machine Translation.

Johannes Graën, Dolores Batinić, and Martin Volk. 2014. Cleaning the europarl corpus for linguistic applications. In *Konvens 2014*. Stiftung Universität Hildesheim.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.

Sandra Hale and Stuart Campbell. 2002. The interaction between text difficulty and translation accuracy. *Babel*, 48(1):14–33.

Hannah J. Haynie, Damián Blasi, Hedvig Skirgård, Simon J. Greenhill, Quentin D. Atkinson, and Russell D. Gray. 2023. Grambank's typological advances support computational research on diverse languages. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 147–149, Dubrovnik, Croatia. Association for Computational Linguistics.

Miguel A. Jimenez-Crespo. 2023. "translationese" (and "post-editese"?) no more: on importing fuzzy conceptual tools from translation studies in MT research. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 261–268, Tampere, Finland. European Association for Machine Translation.

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

Eric Khiu, Hasti Toossi, David Anugraha, Jinyu Liu, Jiaxu Li, Juan Flores, Leandro Roman, A. Seza Doğruöz, and En-Shiun Lee. 2024. Predicting machine translation performance on low-resource languages: The role of domain similarity. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1474–1486, St. Julian's, Malta. Association for Computational Linguistics.

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

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Zheng Wei Lim, Ekaterina Vylomova, Charles Kemp, and Trevor Cohn. 2024. Predicting human translation difficulty with neural machine translation. *Transactions of the Association for Computational Linguistics*, 12:1479–1496.

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

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from

movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

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

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

Rei Miyata, Anthony Hartley, Cécile Paris, Midori Tatsumi, and Kyo Kageura. 2015. Japanese controlled language rules to improve machine translatability of municipal documents. In *Proceedings of Machine Translation Summit XV: Papers*, Miami, USA.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*.

Sharon O'Brien. 2004. Machine translatability and post-editing effort: How do they relate. In *Proceedings of Translating and the Computer 26*, London, UK. Aslib.

Arturo Oncevay, Barry Haddow, and Alexandra Birch. 2020. Bridging linguistic typology and multilingual machine translation with multi-view language representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2391–2406, Online. Association for Computational Linguistics.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with markov chain monte carlo. *The Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Michael Paul, Hirofumi Yamamoto, Eiichiro Sumita, and Satoshi Nakamura. 2009. On the importance of pivot language selection for statistical machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 221–224, Boulder, Colorado. Association for Computational Linguistics.

Esther Ploeger, Wessel Poelman, Andreas Holck Høeg-Petersen, Anders Schlichtkrull, Miryam de Lhoneux, and Johannes Bjerva. 2025. A principled framework for evaluating on typologically diverse languages. *To appear in Computational Linguistics*.

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

Maja Popović. 2016. chrF deconstructed: beta parameters and n-gram weights. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 499–504, Berlin, Germany. Association for Computational Linguistics.

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

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

Enora Rice, Ali Marashian, Hannah Haynie, Katharina Wense, and Alexis Palmer. 2025. Untangling the influence of typology, data, and model architecture on ranking transfer languages for cross-lingual POS tagging. In *Proceedings of the 1st Workshop on Language Models for Underserved Communities (LM4UC 2025)*, pages 22–31, Albuquerque, New Mexico. Association for Computational Linguistics.

Jacqueline Rowe, Edward Gow-Smith, and Mark Hepple. 2025. Limitations of religious data and the importance of the target domain: Towards machine translation for Guinea-Bissau creole. In *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)*, pages 183–200, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Hedvig Skirgård, Hannah J Haynie, Harald Hammarström, Damián E Blasi, Jeremy Collins, Jay

Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Michael Dunn, and 1 others. 2023. Grambank v1.0.

Sanjun Sun. 2015. Measuring translation difficulty: Theoretical and methodological considerations. *Across languages and cultures*, 16(1):29–54.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. 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)*, pages 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Nancy Underwood and Bart Jongejan. 2001. Translatability checker: a tool to help decide whether to use MT. In *Proceedings of Machine Translation Summit VIII*, Santiago de Compostela, Spain.

Bram Vanroy, Orphée De Clercq, and Lieve Macken. 2019. Correlating process and product data to get an insight into translation difficulty. *Perspectives*, 27(6):924–941.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2008. Parallel corpora for medium density languages. In *Recent advances in natural language processing IV: selected papers from RANLP 2005*, pages 247–258. John Benjamins Publishing Company.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Translation Performance for Source-original and Target-original Test Lines

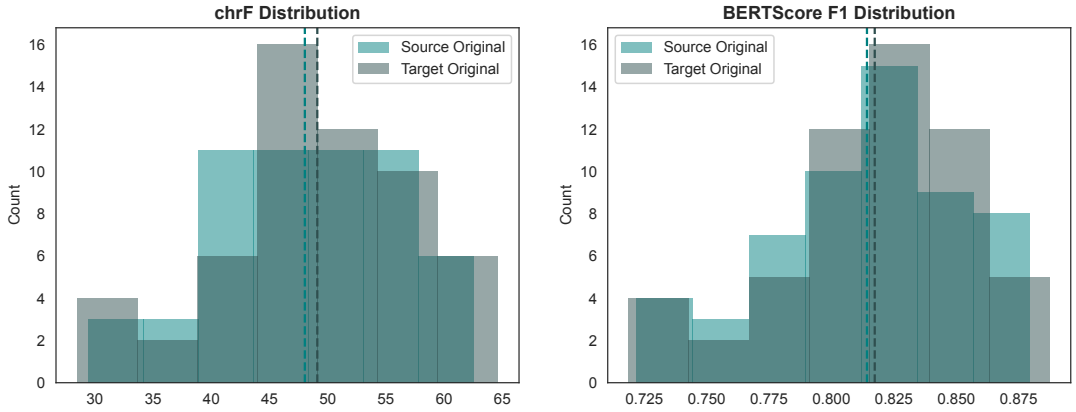


Figure 8: For each of the 56 translation directions in our study, we compute the average chrF2 and BERTScore (F1), for only those test samples that were originally spoken in the source language, with those that were originally spoken in the target language. Comparing these distributions, we observe that, in line with previous work (Zhang and Toral, 2019), scores for the target-original portion are on average higher, implying that "translationese" in the source text can inflate MT performance, albeit slightly. For this reason, we control for the proportion of translated text in our experiments.

B Pairwise Language Similarity Estimates

	Typological Similarity (Grambank)									**Genealogical Similarity (lang2vec)**							
	da	**nl**	**en**	**fi**	**fr**	**it**	**sv**	**pt**		**da**	**nl**	**en**	**fi**	**fr**	**it**	**sv**	**pt**
da	–	0.46	0.28	0.08	0.41	0.47	0.60	0.43	da	–	0.60	0.60	0.00	0.20	0.20	1.00	0.20
nl	0.46	–	0.45	0.18	0.32	0.26	0.61	0.21	nl	0.42	–	0.56	0.00	0.14	0.14	0.42	0.14
en	0.28	0.45	–	0.23	0.26	0.26	0.32	0.24	en	0.42	0.56	–	0.00	0.14	0.14	0.42	0.14
fi	0.08	0.18	0.23	–	0.01	0.00	0.13	0.02	fi	0.00	0.00	0.00	–	0.00	0.00	0.00	0.00
fr	0.41	0.32	0.26	0.01	–	0.45	0.49	0.47	fr	0.09	0.09	0.09	0.00	–	0.61	0.09	0.79
it	0.47	0.26	0.26	0.00	0.45	–	0.32	1.00	it	0.14	0.14	0.14	0.00	0.98	–	0.14	0.98
sv	0.60	0.61	0.32	0.13	0.49	0.32	–	0.29	sv	1.00	0.60	0.60	0.00	0.20	0.20	–	0.20
pt	0.43	0.21	0.24	0.02	0.47	1.00	0.29	–	pt	0.09	0.09	0.09	0.00	0.84	0.65	0.09	–

	Word Order Similarity									**Subword Overlap**							
	da	**nl**	**en**	**fi**	**fr**	**it**	**sv**	**pt**		**da**	**nl**	**en**	**fi**	**fr**	**it**	**sv**	**pt**
da	–	0.50	0.86	0.52	0.71	0.67	0.98	0.63	da	–	0.41	0.44	0.05	0.26	0.17	1.0	0.16
nl	0.53	–	0.51	0.18	0.45	0.41	0.52	0.33	nl	0.41	–	0.41	0.06	0.27	0.22	0.32	0.18
en	0.80	0.44	–	0.53	0.71	0.69	0.85	0.74	en	0.44	0.41	–	0.04	0.62	0.26	0.36	0.31
fi	0.91	0.73	0.82	–	0.78	0.66	0.76	0.75	fi	0.05	0.06	0.04	–	0.00	0.03	0.06	0.02
fr	0.48	0.22	0.64	0.00	–	0.79	0.55	0.81	fr	0.26	0.27	0.62	0.00	–	0.28	0.20	0.25
it	0.53	0.31	0.68	0.32	0.8	–	0.62	0.92	it	0.17	0.22	0.26	0.03	0.28	–	0.13	0.50
sv	1.00	0.54	0.93	0.61	0.71	0.66	–	0.71	sv	1.00	0.32	0.36	0.06	0.20	0.13	–	0.16
pt	0.63	0.38	0.73	0.14	0.89	0.94	0.67	–	pt	0.16	0.18	0.31	0.02	0.25	0.50	0.16	–

Table 6: Pairwise normalized language similarities for all non-`lang2vec` measures in our study.

C BLEU on Validation Set per Training Epoch

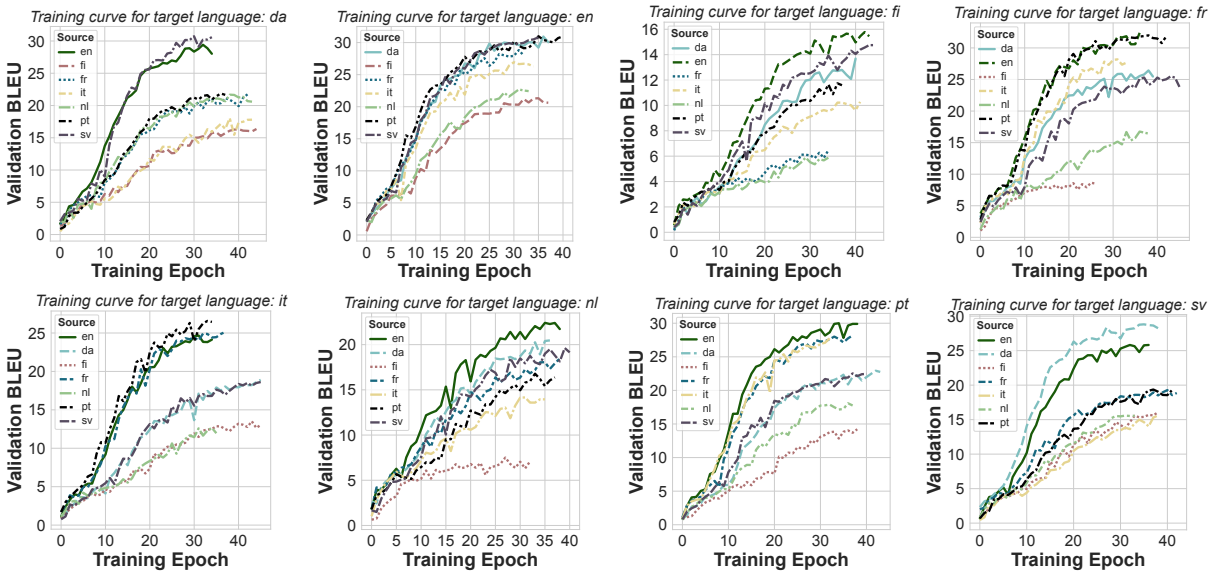


Figure 9: BLEU on validation set per training epoch per target language.

D Loss per Training Epoch per Target Language

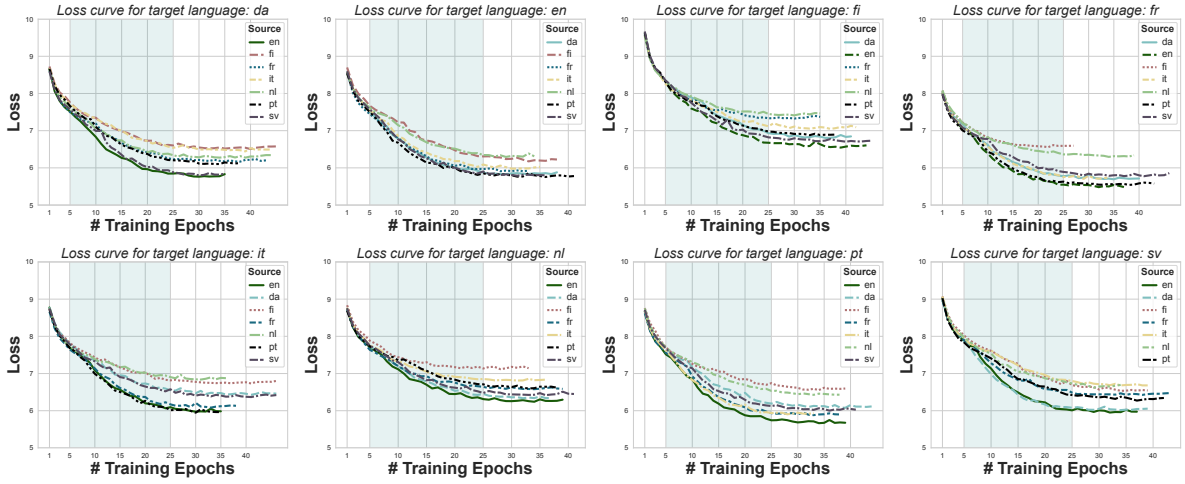


Figure 10: Loss curves per target language in training, with marked the slope that we calculate.