Training Models on Dialects of Translationese Shows How Lexical Diversity and Source–Target Syntactic Similarity Shape Learning

Anonymous ACL submission

Abstract

Machine-translated data is widely used in multilingual NLP, particularly where native text is scarce. However, translated text differs systematically from native text. This phenomenon is known as *translationese*, and it reflects both traces of the source language and characteristic properties of translation itself. In this paper, we study how training on machine-translated data affects small English language models, focusing on how translationese from different source languages shapes linguistic acceptability judgments and language modeling for different domains. We train models on English text translated from 24 typologically and resource-diverse source languages, enabling a systematic analysis of how source language and corpus properties influence what models learn. Our results show that the source language has a clear impact on model behavior: general perplexity is more driven by the lexical diversity of the translated corpus, but grammatical performance is strongly correlated to typological similarity to English if trained on enough data.[1]

1 Introduction

Translated text is widely used in multilingual NLP, particularly for low-resource languages. However, translated text differs systematically from native text, even when done by professional human translators. These differences, known as *translationese*, include traces of the source language and characteristic stylistic and structural patterns (Gellerstam, 1986; Baker, 1993). Even human translations resemble non-native language more than native production (Rabinovich et al., 2016), and machine translation is known to reduce lexical and morphological richness (Vanmassenhove et al., 2021).

Translated data is nonetheless very useful, and including it in pre-training can improve model performance for lesser-resourced languages. Doshi et al. (2024) show that adding data translated from English helps for Indic languages. However, in control experiments in the other translation direction, they also find that training on text translated to English degrades linguistic acceptability compared to native English, and that the source language of the translation can make a difference. We build on these findings, studying how translationese from 24 diverse source languages shapes the linguistic knowledge of small language models, using language modeling on different domains and intrinsic linguistic acceptability preferences for different syntactic and morphological phenomena.

We train models on English text translated from 24 typologically diverse languages using the FineTranslations dataset (Penedo et al., 2026). This setup reverses the typical translation direction (English → other language) but enables evaluation on consistent, high-quality English benchmarks. It likely provides an upper bound for the performance of models trained on translated data, as machine translations into English are much easier than translations into smaller languages due to the models' proficiency in English. Following the Goldfish setup (Chang et al., 2024), we train small models with varying data scale. Our main findings are:

- Translationese has a flattening effect on the resulting models, increasing perplexity on native text and reducing linguistic acceptability scores.
- Surface lexical diversity in the translated corpus correlates with lower perplexity in low-data settings, but is not strongly correlated with grammatical performance.
- Typological similarity between the source language and English significantly improves grammatical performance given enough data.
- Models trained on typologically similar source languages show lower mutual perplexity on each other's pretraining data, indicating that their varieties of translationese are also similar.

[1] The models are available at: https://huggingface.co/collections/---anonymized---

2 Related Work

The concept of translationese was introduced by Gellerstam (1986) to describe systematic properties of translated text. Early work in translation studies shows that translated texts tend to be lexically and syntactically simpler, more explicit, and more conventional in style (Baker, 1993). Translations also carry traces of the source language, similar to cross-linguistic influence in multilingual speakers (Jarvis and Pavlenko, 2008). Although first studied in human translation, translationese also arises in machine translation (MT). Traditional MT systems often produced literal translations that closely followed source-language structure, amplifying these effects. Machine-translated text has been shown to be less structurally diverse than human translations (Luo et al., 2024) and to lose lexical and morphological richness compared to native text (Vanmassenhove et al., 2021).

Translationese is easy to detect automatically. Simple classifiers can reliably distinguish translated from native text (Koppel and Ordan, 2011), and even unsupervised clustering separates them within the same domain (Rabinovich and Wintner, 2015). Classifiers outperform professional translators at this task, with function words, morphosyntactic categories, and pronouns being particularly informative features (Baroni and Bernardini, 2005).

Several approaches aim to reduce translationese in MT output, including models trained to produce more natural target-language text (Riley et al., 2020). Large language models have also been shown to produce more fluent, less literal translations than previous MT systems (Raunak et al., 2023). Nevertheless, even high-quality MT output differs detectably from native text, and the trade-off between fluency and accuracy, with accuracy typically prioritized (Lim et al., 2024), may further reduce naturalness. The use of translated data in NLP raises concerns for both evaluation and training. Translated test sets tend to be easier than native ones, potentially inflating performance estimates on multilingual benchmarks (Chen et al., 2024b). For training, translated data is widely used particularly in instruction tuning, since few languages have large native corpora (Holmström and Doostmohammadi, 2023; Muennighoff et al., 2023; Chen et al., 2024a). However, models trained on translated data may develop weaker preferences for natural and idiomatic language (Kunz, 2026).

Building on this literature, we systematically study how translationese from different source languages affects the syntactic and morphological knowledge acquired by small language models.

3 Experimental Setup

3.1 Models

To train our models, we follow the Goldfish setup (Chang et al., 2024), which is a suite of small monolingual GPT-style Transformer models trained on controlled data budgets, using their code and hyperparameters, which means that all models used in this paper have 125M parameters. To ensure comparability of perplexity scores, we use the tokenizer of the original English Goldfish model for all experiments, with a vocabulary of 50,000 tokens.

3.2 Languages

We include 24 typologically and genealogically diverse languages. From the Indo-European language family, we include Germanic languages (Swedish, Icelandic, Faroese), Slavic languages (Ukrainian, Czech, Upper Sorbian) and Indo-Iranian languages (Persian, Urdu, Northern Kurdish). We also include languages from the Uralic (Hungarian, Estonian, Northern Sámi), Semitic (Standard Arabic, Hebrew, Maltese), Dravidian (Tamil, Telugu, Kannada), Austronesian (Indonesian, Standard Malay, Javanese), and Niger-Congo/Bantu (Swahili, Kinyarwanda, Zulu) families. Within each group, we include higher- and lower-resource languages, but with a minimum of about 100MB of available data To analyze syntactic similarity to English and to each other, we compute cosine similarity over WALS syntactic features using lang2vec (Littell et al., 2017) where these features exist in WALS. This allows us to relate model behavior to typological distance. Results are given in Figure 1.

3.3 Data

We train on parts of the FineTranslations dataset(Penedo et al., 2026), which provides machine translations of the FineWeb-2 corpus (Penedo et al., 2025) into English. We adopt the Goldfish paper setups that train on 100MB and 1000MB of data.[2] As native English baselines, we train English models on FineWeb (Penedo et al., 2024) using matched data sizes.

[2]While in this paper we only use the 100MB and 1000MB models, we also train and release the smaller variants 5MB and 10MB for completeness and future research.

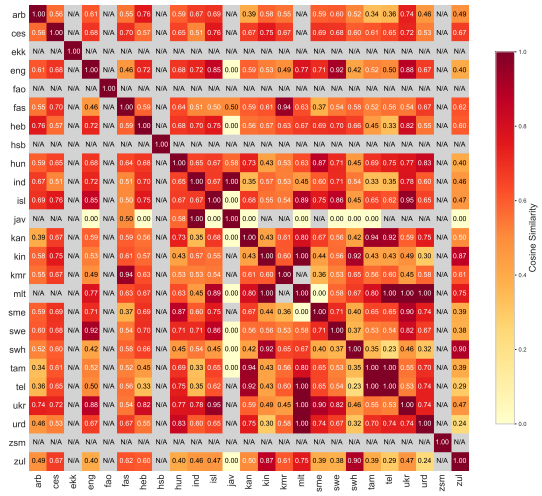

Figure 1: WALS (Dryer and Haspelmath, 2013) syntax feature cosine similarity, generated with lang2vec (Littell et al., 2017) for all language pairs. Some languages are unavailable or do not have common features (N/A).

Statistics In Table 1, we analyze the translated 100MB corpora using surface-level statistics: size of the base corpus (in number of words and disk space), vocabulary size (whitespace-tokenized), type–token ratio (TTR), bigram TTR, and average sentence length.[3] The English baseline shows the highest diversity, measured in vocabulary size, TTR and bigram TTR. Among translated corpora, vocabulary size ranges from 350k to 610k types. Hungarian and Javanese exhibit relatively high TTR and bigram TTR, while Kinyarwanda shows the lowest diversity in all metrics. Average sentence length also varies substantially (8.8–20.9 tokens). Shorter sentences appear in smaller languages such as Northern Sámi, Faroese, and Upper Sorbian. Arabic and Kinyarwanda show the longest sentences, followed by Persian, Urdu, and Hebrew.

We further explore how predictive base corpus size is for diversity in the 100MB translated corpus. For the 100MB datasets, correlations between base corpus size and TTR are weak and not statistically significant[4] (Pearson $r = 0.238$, $p = 0.2636$; Spearman $\rho = +0.177$, $p = 0.4070$) Bigram TTR showed a moderate positive correlation (Pearson $r = 0.410$, $p = 0.0468$; Spearman $\rho = +0.445$, $p = 0.02925$). For the 1000MB datasets, correlations are stronger for TTR (Pearson $r = 0.439$, $p = 0.0603$; Spearman $\rho = 0.439$, $p = 0.0603$) and clearly significant for bigram TTR (Pearson $r = 0.579$, $p = 0.0094$; Spearman $\rho = 0.681$, $p = 0.0013$). This suggests that in larger base corpora, the diversity of translated text increases.

3.4 Evaluation

We evaluate the models in three ways:

Perplexity on native English text We compute perplexity on the English source side of FLORES+ (Costa-jussà et al., 2024), which includes *Wikinews*, *Wikitravel*, and *Wikibooks* domains, and on a 10MB subset of native English Fineweb text.

Linguistic Acceptability To target syntactic and morphological phenomena more directly, we evaluate on BLiMP (Warstadt et al., 2020), which focuses on minimal pairs of correct sentences and sentences containing grammatical errors. As in the original setup, we count a sample as correct if the correct sentence is assigned a lower mean perplexity than the wrong one. For more fine-grained analyses of lingustic phenomena, we group them into the same categories as Warstadt et al. (2020).

Cross-translation perplexity We measure perplexity on translated corpora originating from other source languages (using the 10MB datasets). This allows us to analyze similarities between models and data trained on different translation sources.

4 Results and Discussion

4.1 Perplexity

We first evaluate general language modeling performance using perplexity on FLORES domains (English Wikibooks, Wikivoyage, Wikinews) and the English FineWeb subset. Results are in Table 2.

4.1.1 Performance and gap to baseline.

No model trained on translated data matches the native English baseline across any domain or model size. For the 100MB models, the gap to the English baseline is smaller on FLORES (1–2.5 points) but very large on FineWeb (+19.55 points). Note however that even the baseline models' training data is a part of FineWeb, so distributions of train and test data are more similar than for the other domains. At 1000MB, perplexity decreases substantially for all models. The gap on FLORES narrows to under 1 point on Wikinews, while the FineWeb gap remains large (+9.33 points). So, scaling helps overall, but does not recover native-like performance.

[3] In Appendix B, we also analyze experimental measures of morphological complexity of source and target corpora.

[4] All reported p-values in this paper are two-tailed. Statistical significance was assessed at $\alpha = 0.05$.

Source	Language	Base Corpus		Translated Corpus			
		Words	UTF-8 Bytes	VocabSize	TTR (%)	BigrTTR (%)	AvgSenLen
eng	English	18,500,000M	–	835517	4.77	35.42	13.69
arb	Standard Arabic	32,813M	293.59 GB	516177	3.02	25.88	20.87
ces	Czech	35,479M	206.33 GB	590228	3.42	29.19	13.63
ekk	Estonian	6,564M	40.82 GB	553751	3.25	28.36	13.32
fao	Faroese	101M	564.10 MB	448266	2.55	21.25	10.76
fas	Persian	39,706M	304.62 GB	435917	2.56	24.11	18.30
heb	Hebrew	8,463M	68.71 GB	470477	2.68	24.85	17.09
hsb	Upper Sorbian	15M	92.13 MB	397972	2.60	19.76	11.89
hun	Hungarian	30,919M	199.69 GB	561913	3.33	29.63	14.42
ind	Indonesian	60,264M	348.65 GB	550569	3.22	28.53	14.53
isl	Icelandic	1,696M	10.27 GB	481243	2.77	24.95	13.84
jav	Javanese	140M	783.32 MB	611377	3.58	28.93	12.33
kan	Kannada	748M	12.91 GB	494342	2.90	25.94	13.47
kin	Kinyarwanda	128M	793.30 MB	350870	1.98	18.69	20.89
kmr	Northern Kurdish	221M	1.16 GB	495600	2.85	23.83	15.57
mlt	Maltese	287M	1.53 GB	480824	2.76	24.28	14.24
sme	Northern Sámi	25M	177.92 MB	419365	2.52	20.30	8.75
swe	Swedish	35,745M	202.96 GB	509665	2.88	26.75	14.27
swh	Swahili	570M	3.08 GB	492842	2.83	24.23	16.32
tam	Tamil	1,937M	36.97 GB	532375	3.11	26.29	12.74
tel	Telugu	891M	14.42 GB	552762	3.21	26.96	12.09
ukr	Ukrainian	25,586M	254.86 GB	594748	3.54	27.25	13.67
urd	Urdu	2,733M	19.93 GB	417644	2.43	23.81	18.21
zsm	Standard Malay	5,648M	31.94 GB	572369	3.28	27.87	13.20
zul	Zulu	71.7M	487.38 MB	526206	3.00	27.10	14.83

Table 1: Base corpus sizes (FineWeb-2) and translated corpus (FineTranslations) surface statistics.

Variation across source languages. Performance varies strongly across source languages. For the 100MB models, perplexity ranges from 20.09–28.21 on Wikibooks, 25.91–39.59 on Wikivoyage, 19.44–28.93 on Wikinews, and 38.22–79.37 on FineWeb. The 1000MB models reduce this variance but do not eliminate it. Among FLORES domains, Wikivoyage shows the largest variation while Wikinews is the most stable after scaling.

Typological distance vs. performance. Typological similarity to English does not consistently predict lower perplexity. Typologically distant languages such as Estonian, Hungarian, Maltese, and Zulu often perform as well as or better than Germanic languages on FLORES. For instance, Estonian and Hungarian match or slightly outperform Swedish on Wikibooks at 1000MB, and Maltese yields the best Wikivoyage scores in both data regimes. On FineWeb, Swedish consistently yields the lowest perplexity among translated models, though this may reflect domain overlap between Swedish and English web corpora rather than, or in addition to, typological proximity.

4.1.2 Correlation Analysis

To better understand performance differences, we analyze correlations between perplexity and three predictors: base corpus size, lexical diversity (TTR) of the translated corpus, and syntactic similarity to English. Results are in Table 3. For the 100MB models ($n = 24$), TTR shows the largest and most consistent correlations. It correlates negatively with perplexity on both FLORES (Pearson $r = -0.56$ for the full dataset) and FineWeb ($r = -0.51$). The exception is Wikinews, where the correlation is not significant. Corpus size shows weaker but directionally consistent negative correlations; Pearson r values are not significant, but Spearman ρ reaches significance on Wikinews and FineWeb, consistent with the skewed distribution of corpus sizes, where rank-based correlations like Spearman may be more robust. Syntactic similarity (lang2vec) shows no significant correlation with perplexity at this scale. For the 1000MB models ($n = 19$), the picture shifts. The TTR association is no longer observed (Pearson $r = +0.06$ on FLORES, $r = -0.01$ on FineWeb), suggesting that lexical diversity no longer reliably predicts performance once sufficient data is available, or possibly reflecting the exclusion of low-resource languages that do not have 1000MB corpora. Corpus size remains negatively correlated but without reaching significance. But most notably, syntactic similarity becomes considerably more predictive: lang2vec

	100MB				1000MB			
Source	Wikibooks	Wikivoyage	Wikinews	FineWeb	Wikibooks	Wikivoyage	Wikinews	FineWeb
eng	18.77	24.30	17.01	18.67	10.49	14.84	12.31	11.63
arb	21.94	30.47	19.79	46.47	14.94	19.58	13.53	23.53
ces	20.09	26.05	19.85	40.13	14.39	17.62	13.71	22.26
ekk	20.47	26.98	20.56	41.98	14.24	17.69	14.25	22.33
fao	23.79	30.11	23.99	51.56	19.46	24.68	19.44	34.73
fas	21.71	31.76	21.42	48.51	14.97	20.99	14.42	24.84
heb	21.95	29.61	22.29	46.80	14.49	18.90	14.61	22.67
hsb	27.88	36.37	28.93	79.37	26.89	34.24	27.59	73.30
hun	20.70	27.48	19.86	44.63	14.41	18.18	13.66	23.06
ind	21.15	28.36	20.44	39.40	14.46	18.79	13.63	21.61
isl	20.81	26.63	20.38	40.02	14.67	17.96	14.13	22.01
jav	21.25	28.18	23.00	46.95	15.53	19.08	15.40	24.68
kan	23.10	32.52	21.59	48.96	16.05	21.49	14.65	25.89
kin	28.21	39.59	23.81	63.50	20.25	27.97	17.60	34.86
kmr	21.15	31.07	21.37	47.53	15.08	19.98	14.87	25.14
mlt	20.66	25.91	19.99	39.32	14.37	16.87	13.44	21.34
sme	27.42	33.78	28.38	65.10	24.56	31.43	25.77	55.32
swe	21.39	26.40	21.05	38.22	14.40	17.10	14.21	20.96
swh	21.57	28.38	19.44	42.84	15.04	18.71	13.28	22.62
tam	22.61	32.17	20.60	46.80	15.90	21.43	14.09	25.19
tel	23.24	32.28	21.21	46.32	15.86	21.27	14.65	24.51
ukr	22.22	30.46	22.07	47.93	15.61	20.65	15.22	25.95
urd	22.40	31.38	19.66	48.49	15.41	20.07	13.28	25.08
zsm	22.26	30.02	21.09	44.14	15.35	19.19	14.01	22.47
zul	20.70	26.37	20.41	40.87	16.12	20.09	15.38	25.67

Table 2: Perplexity on FLORES domains and English FineWeb (10MB eval set) for 100MB and 1000MB models. Results in gray indicate that the translated corpora for these low-resource languages were smaller than 1000MB.

Model	Predictor	Pearson r					Spearman ρ				
		Books	Vovage	News	All	FineWeb	Books	Voyage	News	All	FineWeb
100MB (n=24)	Corpus Size	-0.35	-0.29	-0.31	-0.34	-0.34	-0.40	-0.31	-0.45^{*}	-0.39	-0.47^{*}
	TTR	-0.60^{**}	-0.56^{**}	-0.38	-0.56^{**}	-0.51^{*}	-0.48^{*}	-0.42^{*}	-0.25	-0.44^{*}	-0.45^{*}
	lang2vec	-0.02	-0.17	-0.07	-0.11	-0.17	-0.11	-0.29	-0.03	-0.22	-0.34
1000MB (n=19)	Corpus Size	-0.44	-0.17	-0.21	-0.28	-0.29	-0.42	-0.15	-0.21	-0.27	-0.27
	TTR	$+0.07$	-0.07	$+0.33$	$+0.06$	-0.01	$+0.04$	-0.09	$+0.22$	$+0.02$	-0.01
	lang2vec	-0.45	-0.35	-0.31	-0.42	-0.43	-0.53^{*}	-0.50^{*}	-0.14	-0.52^{*}	-0.45

Table 3: Correlations between predictor variables and perplexity (FLORES subsets and FineWeb) across model sizes. Significance levels: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. Note: 1000MB excludes fao, hsb, kin, sme, zul (small corpora). lang2vec analyses have reduced sample sizes due to missing data (100MB n=20, 1000MB n=17).

WALS similarity now shows moderate negative correlations on FLORES (average Pearson $r = -0.42$; Spearman $\rho = -0.52^{*}$, the latter significant except on Wikinews), i.e., languages syntactically closer to English are associated with lower perplexity at this scale. The effect is weaker and non-significant on FineWeb, but directionally consistent.

4.2 Linguistic Acceptability

Overall BLiMP accuracy scores are in Table 4; per-category results are in Table 6 in Appendix A.

4.2.1 Performance and gap to baseline.

For the 100MB models, the English baseline reaches 66.20%. Translated-source models perform very similarly, ranging from 63.29% (Upper Sorbian) to 67.67% (Swedish); a spread of only 4.38 points, noticeably smaller than what we observed for perplexity. Swedish, the typologically closest language to English, achieves the highest translated score, just *above* the baseline. However, several typologically distant languages perform nearly as well: Tamil (67.54%), Northern Sámi (67.26%), Zulu (66.96%), and Kannada (66.83%), showing that typological proximity alone does not determine performance at this scale. In Table 6 in Appendix A, we see that performance differs across phenomena. Across phenomena, agreement categories (Anaphor Agreement, Irregular Forms, Determiner–Noun Agreement) are generally high

Language	100MB	1000MB
eng	66.20	76.01
arb	65.14	70.89
ces	66.08	73.33
ekk	63.98	72.31
fao	65.66	70.61
fas	63.38	71.26
heb	65.08	73.11
hsb	63.29	62.82
hun	65.11	72.90
ind	66.39	72.48
isl	65.68	73.37
jav	65.30	68.95
kan	66.83	72.28
kin	65.18	71.64
kmr	66.24	72.68
mlt	65.06	71.84
sme	67.26	66.89
swe	67.67	74.37
swh	66.00	72.77
tam	67.54	73.49
tel	66.94	73.36
ukr	64.53	71.48
urd	65.36	73.14
zsm	65.96	72.43
zul	66.96	70.46

Table 4: Overall BLiMP accuracy (%) by language for 100MB and 1000MB models. Gray indicates training corpora smaller than 1000MB. Table 6 in Appendix A reports accuracy for 12 categories of phenomena.

(75–90%), suggesting that local morphosyntactic constraints are learned reliably even from limited translated data. Island Effects are consistently the most difficult (English: 43.67%; most translations: 40–49%), followed by NPI Licensing and Filler–Gap dependencies (often 45–60%), reflecting difficulty with long-distance dependencies. Some distant languages show unexpected peaks: Zulu reaches 93.10% on Anaphor Agreement, above the English score of 91.30%, maybe reflecting translation patterns that emphasize certain constructions. Scaling to 1000MB improves performance across models. The baseline increases from 66.20% to 76.01%, and translated-source models also improve substantially but are for this scale *consistently below* the baseline. Swedish again performs best (74.37%), followed by Tamil (73.49%), while Javanese scores lowest among fully trained models (68.95%). Gains are largest for structurally complex phenomena: Island Effects improve by over 13 points for English (43.67% → 56.76%), with similar gains across languages, and Control/Raising and Binding improve by 6–10 points. Agreement phenomena approach ceiling (often above 95%). NPI Licensing and Quantifiers remain challenging

	Corpus Size	TTR	lang2vec
100MB	-0.09 (-0.09)	$+0.11$ ($+0.08$)	$+0.06$ (-0.10)
1000MB	$+0.01$ ($+0.08$)	-0.32 (-0.13)	$+0.61^{**}$ ($+0.32$)

Table 5: Correlations with overall BLiMP accuracy (Pearson r, Spearman ρ in parentheses). n=24 for 100MB, n=19 for 1000MB models. For per-category results, see Table 7 in Appendix A.

even at this scale, typically staying in the mid-50s to low-60s and mid-70s respectively.

4.2.2 Correlation Analysis

We report correlations between BLiMP accuracy and the three predictors corpus size, TTR, and syntactic similarity to English in Table 5. For the 100MB models ($n = 24$), none of the predictors shows significant correlations with overall BLiMP accuracy, and effect sizes are small (($|r| \leq 0.11$)). This is consistent with the narrow accuracy range at this scale: in low-data settings, BLiMP scores appear insensitive to corpus properties. For the 1000MB models ($n = 19$), the pattern shifts. While corpus size remains unrelated to BLiMP accuracy (Pearson $r = +0.01$; Spearman $\rho = +0.08$), and TTR shows a weak, non-significant negative correlation ($r = -0.32$), lang2vec similarity to English shows a strong and significant positive correlation with BLiMP accuracy ($r = +0.61$). This suggests that once sufficient training data is available, typological similarity to English is strongly associated with BLiMP scores. Correlations for the different categories can be found in Table 7 in Appendix A. At 100MB, correlations between predictors and per-category BLiMP scores are weak and inconsistent across phenomena. At 1000MB, syntactic similarity shows the largest and most consistent correlations, with particularly strong effects for agreement phenomena (Anaphor and Subject–Verb Agreement: $r = +0.74^{***}$) and long-distance dependencies (Filler–Gap: $r = +0.66^{**}$; Ellipsis: $r = +0.61^{**}$; Island Effects: $r = +0.53^{*}$). TTR correlations are mostly negative and not significant.

4.3 Cross-Evaluation Perplexity

We next ask whether models trained on translations from language A generalize better to translations from language B when the two are typologically similar. Figure 2 shows cross-evaluation perplexity matrices; Figure 3 reports the correlation between syntactic similarity and cross-evaluation perplexity.

Across both training dataset scales, we find a

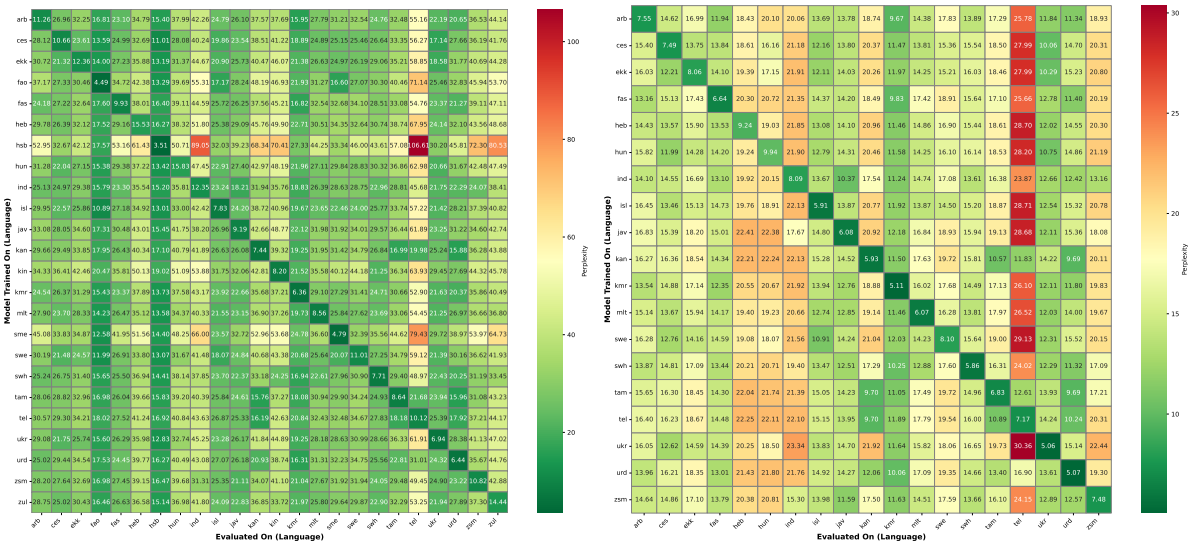

Figure 2: Cross-evaluation perplexity matrices (heatmaps) for models trained on translations from language A, evaluated on translations from language B. Left: 100MB models; right: 1000MB models.

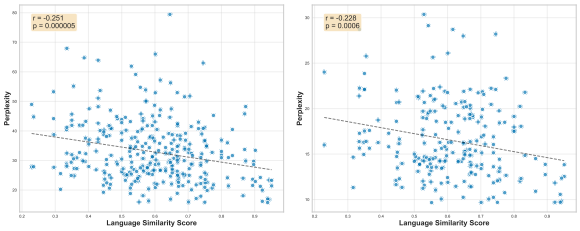

Figure 3: Correlations between language similarity and cross-evaluation perplexity across language pairs. Left: 100MB models; right: 1000MB models.

consistent and statistically significant negative correlation between typological similarity and cross-evaluation perplexity: the more syntactically similar two source languages are, the lower the perplexity when a model trained on the translation from one is evaluated on the translation from other. For the 100MB models (322 source language pairs), the correlation is moderate and significant (Pearson $r = -0.251$, $p < 0.001$; Spearman $\rho = -0.244$, $p < 0.001$),[5] and a similar pattern holds for the 1000MB models (226 pairs; Pearson $r = -0.228$, $p < 0.001$; Spearman $\rho = -0.188$, $p < 0.01$). The effect appears slightly stronger at 100MB, suggesting that smaller models are more sensitive to typological proximity. At 1000MB, the effect remains robust but somewhat weaker. As expected, the diagonal of the heatmaps in Figure 3 shows the lowest perplexities, with models performing best on the source language corpus they were trained on. Beyond the diagonal, related languages tend to show relatively low mutual perplexity, providing a complementary view to previous analyses: typological similarity matters not only from the source language to English, but also between translations from different source languages, indicating that similar languages result in similar translations into English, i.e., in similar dialects of *translationese*.

The two lowest-resource languages stand out in the 100MB setting. Models trained on Upper Sorbian and Northern Sámi yield high perplexities on nearly all evaluation languages, reflecting that their corpus is likely less diverse and possibly generally of lower quality. Interestingly, Upper Sorbian as an *evaluation* target proves relatively easy to model across many training languages, a pattern that however does not hold for Northern Sámi. Even the Telugu dataset stands out in Figure 3 as it gets unusually high perplexities by most models. A brief inspection of this dataset makes us assume that this likely reflects the high density of named entities in the corpus translated from Telugu. These terms, such as personal names, place names, and culturally specific terms, are long and tokenize into many subword units. Models trained on translations from typologically and geographically closer languages such as Kannada, Tamil, and to a lesser extent Urdu, perform better on Telugu, possibly due to shared named entities and overlapping cultural vocabulary.

4.4 Takeaways

A first takeaway is that the choice of source language matters when building translated training

[5]We note that the language pairs are not statistically independent, and p-values should be interpreted cautiously here.

corpora, but how to best select it depends on the goal. For general language modeling (perplexity), typological similarity to English is not a reliable guide: some typologically distant languages yield competitive results, suggesting that corpus characteristics such as domain coverage and lexical distribution also play an important role. That said, at sufficient data scale and excluding very small languages, similarity *does* have a consistent effect; it is simply one factor among several.

For grammatical competence, typological similarity is more important. For the 1000MB training data regimes, syntactic similarity strongly and significantly predicts performance on BLiMP, particularly for agreement and long-distance dependencies. When translated corpora are used to support grammatical generalization, choosing a structurally similar source language is beneficial.

More generally, we find that using translated corpora has an impoverishing effect, leading to lower language modelling scores on native English corpora and to lower grammatical performance in the (more realistic) setup with more training data.

These findings align with prior work on cross-lingual transfer in other setups. Rice et al. (2025) find that typology, lexical overlap and dataset properties together predict zero-shot POS tagging performance. Lin et al. (2019) similarly show that phylogenetic similarity, typological features, lexical overlap, and data size all contribute to transfer in tagging and parsing. Blaschke et al. (2025) further show that syntactic similarity is most predictive for word-level tasks, while surface overlap matters more for topic classification. Work on bilingual pre-training (Malkin et al., 2022) shows that language structure, not just data size, shapes transfer, and that structurally central languages can act as strong donors. For translated pre-training data specifically, Doshi et al. (2024) also find that the source language matters and that translationese negatively affects linguistic acceptability judgments.

Overall, our results suggest that cross-lingual generalization from translated data is shaped by both structural and corpus-level factors, with structural similarity becoming especially important for grammatical competence as data scale increases.

4.5 Future Work

Our findings indicate some promising directions for future work. It would be useful to disentangle structural similarity, lexical overlap and base corpus domain coverage more directly, for example using controlled synthetic corpora, and to identify which specific typological features (word order, agreement systems, case marking) drive the strongest transfer effects. It also remains to be seen whether the observed patterns hold for larger models and datasets, and whether adding a small amount of native English data or combining multiple source languages may already improve grammatical competence and language modeling performance.

5 Conclusion

This study examined how the source language of machine-translated training data affects English language modeling and grammatical knowledge in small language models. General perplexity results showed that no translated corpus matched the English baseline and that source language choice clearly matters. At the larger data scale, syntactic closeness to English was correlated with lower perplexity, but some distant languages performed as well as or better than closely related ones, indicating that general language modeling quality depends on more than syntactic similarity alone. Surface-level lexical diversity in the corpus was a stronger predictor in low-data settings. Linguistic acceptability preferences showed a different pattern. At the larger data scale, typological similarity to English was strongly correlated with grammatical accuracy, particularly for agreement phenomena and long-distance dependencies, while corpus size had little impact and lexical similarity showed unstable effects. Effects were generally weaker in the smaller-data models, suggesting that structural similarity matters more for learning syntax as training data increases. Cross-evaluation experiments showed that models trained on translations from one language assign lower perplexity to translations from a different language when the two languages are syntactically closer; a pattern that held consistently across both model sizes. It appears that data translated from more similar languages results in similar dialects of *translationese*.

Taken together, these results suggest that the source language of translated training data shapes learned linguistic knowledge in systematic ways. For grammatical competence specifically, typological proximity to the target language becomes the dominant factor at scale. Future work should further disentangle corpus effects from linguistic similarity, and examine whether these patterns hold for larger models and more target languages.

Limitations

Our setup translates into English rather than into low-resource languages, which is the more common use case in multilingual NLP. Translations into a high-resource language like English tend to be more fluent and well-formed, since MT systems are generally stronger in this direction. Our results therefore represent something closer to an upper bound on translated data quality; effects may be more pronounced when translating into lower-resource target languages where MT quality is weaker and translationese artifacts are stronger.

We only consider the FineTranslations corpus that was translated with one LLM. Other translation systems may introduce different patterns.

We do not formally disentangle the effect of the source language's typological properties from the properties of the corpus produced by translating from it, which is relevant as the corpora in our study differ in many properties such as the lexical distribution, domain coverage, noise level, and available data volume. Isolating the contribution of typological structure would at least require a more carefully matched data collection.

An important limitation of translated corpora that we do not address in this study is that translated corpora inherit the content of their source documents, which does not reflect the cultural knowledge, named entities, or pragmatic conventions of the target language community. Native English corpora contain culturally grounded references, idioms, and discourse patterns that translated text tends to flatten or omit. Models trained on translated data may therefore lack not only linguistic naturalness but also the world knowledge that native speakers of the language commonly have.

Lastly, our experiments use small language models (following the Goldfish setup) trained on at most 1000MB of text. It is not clear whether the patterns we observe, particularly the growing importance of typological similarity at larger data scales, generalize to the much larger models and datasets used in current practice. Larger models may be more robust to translationese effects, or may amplify them in ways our setup cannot detect.

Ethical Considerations

We release models trained on translated web data. Although the underlying data has been filtered by its creators, it may still contain content that is inappropriate, biased, or otherwise problematic. As a result, the models may reproduce such patterns. However, the models are intended solely for research purposes, and as they are small and trained on limited data, we consider the likelihood that anyone would deploy them in ways that cause significant real-world harm to be low.

References

Mona Baker. 1993. Corpus Linguistics and Translation Studies: Implications and Applications. In *Text and Technology*. John Benjamins.

Marco Baroni and Silvia Bernardini. 2005. A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text. *Literary and Linguistic Computing*, 21(3):259–274.

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.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2024. Goldfish: Monolingual Language Models for 350 Languages. *Preprint*, arXiv:2408.10441.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. 2024a. Monolingual or Multilingual Instruction Tuning: Which Makes a Better Alpaca. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1347–1356, St. Julian's, Malta. Association for Computational Linguistics.

Pinzhen Chen, Simon Yu, Zhicheng Guo, and Barry Haddow. 2024b. Is It Good Data for Multilingual Instruction Tuning or Just Bad Multilingual Evaluation for Large Language Models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9706–9726, Miami, Florida, USA. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Meet Doshi, Raj Dabre, and Pushpak Bhattacharyya. 2024. Pretraining Language Models Using Translationese. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*,

pages 5843–5862, Miami, Florida, USA. Association for Computational Linguistics.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.4)*. Zenodo.

Haodi Feng, Kang Chen, Xiaotie Deng, and Weimin Zheng. 2004. Accessor variety criteria for chinese word extraction. *Computational Linguistics*, 30(1):75–93.

Martin Gellerstam. 1986. Translationese in Swedish novels translated from English. In L. Wollin and H. Lindquist, editors, *Translation studies in Scandinavia: Poceedings from the Scandinavian Symposium on Translation Theory (SSOTT) II*, number 75 in Lund Studies in English, page 88–95. CWK Gleerup, Lund.

Zellig S. Harris. 1955. From phoneme to morpheme. *Language*, 31(2):190–222.

Oskar Holmström and Ehsan Doostmohammadi. 2023. Making Instruction Finetuning Accessible to Non-English Languages: A Case Study on Swedish Models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 634–642, Tórshavn, Faroe Islands. University of Tartu Library.

Scott Jarvis and Aneta Pavlenko. 2008. *Crosslinguistic influence in language and cognition*. Routledge.

Moshe Koppel and Noam Ordan. 2011. Translationese and Its Dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics.

Jenny Kunz. 2026. Preferences for Idiomatic Language are Acquired Slowly – and Forgotten Quickly: A Case Study on Swedish. *Preprint*, arXiv:2602.03484.

Zheng Wei Lim, Ekaterina Vylomova, Trevor Cohn, and Charles Kemp. 2024. Simpson's Paradox and the Accuracy-Fluency Tradeoff in Translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–103, Bangkok, Thailand. Association for Computational Linguistics.

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

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

Jiaming Luo, Colin Cherry, and George Foster. 2024. To Diverge or Not to Diverge: A Morphosyntactic Perspective on Machine Translation vs Human Translation. *Transactions of the Association for Computational Linguistics*, 12:355–371.

Dan Malkin, Tomasz Limisiewicz, and Gabriel Stanovsky. 2022. A Balanced Data Approach for Evaluating Cross-Lingual Transfer: Mapping the Linguistic Blood Bank. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4903–4915, Seattle, United States. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual Generalization through Multitask Finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Guilherme Penedo, Hynek Kydlíček, Amir Hossein Kargaran, and Leandro von Werra. 2026. FineTranslations. `https://huggingface.co/datasets/HuggingFaceFW/finetranslations`.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. FineWeb2: One Pipeline to Scale Them All – Adapting Pre-Training Data Processing to Every Language. *Preprint*, arXiv:2506.20920.

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

Ella Rabinovich, Sergiu Nisioi, Noam Ordan, and Shuly Wintner. 2016. On the Similarities Between Native,

Non-native and Translated Texts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1881, Berlin, Germany. Association for Computational Linguistics.

Ella Rabinovich and Shuly Wintner. 2015. Unsupervised Identification of Translationese. *Transactions of the Association for Computational Linguistics*, 3:419–432.

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan. 2023. Do GPTs Produce Less Literal Translations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1041–1050, Toronto, Canada. Association for Computational Linguistics.

Enora Rice, Ali Marashian, Hannah Haynie, Katharina von der 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.

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a language in "multilingual" NMT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7737–7746, Online. Association for Computational Linguistics.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

A BLiMP Category Scores

We report the full accuracy scores for all BLiMP subcategories in Table 6. Table 7 presents the correlation of subcategory scores with base corpus size, type-token ratio, and syntactic similarity measures.

We use the following abbreviations in Table 6. For detailed descriptions of each category, we refer the reader to Warstadt et al. (2020).

- AnaAgr – Anaphor Agreement
- IrrFrm – Irregular Forms
- DetAgr – Determiner–Noun Agreement
- SVAgr – Subject–Verb Agreement
- ArgStr – Argument Structure
- Ellips – Ellipsis
- Bindng – Binding
- CtlRai – Control and Raising
- Quantf – Quantifiers
- FilGap – Filler–Gap Dependencies
- NPILic – NPI Licensing
- IslEff – Island Effects

B Approximated Morphological Complexity in the Corpora

To approximate morphological complexity, we use the bigram-based measures proposed by Poelman et al. (2025), which aim to capture surface-form diversity (potentially) arising from morphological variation. *Accessor Variety* (Harris, 1955; Feng et al., 2004) is defined as the number of distinct tokens that appear adjacent to a given token in the corpus. *Entropic Efficiency* is the Shannon entropy of the token's contextual distribution, normalized by its accessor variety. It measures how uniformly the observed contexts of a token are distributed. The intuition behind these measures is that morphologically richer languages tend to produce more diverse and less predictable token sequences. Poelman et al. (2025) show that such statistics can serve as proxies for morphological complexity without requiring explicit morphological annotation. It is however important to keep in mind that accessor variety and entropic efficiency capture surface-form diversity in token sequences rather than morphological structure directly. Both metrics are sensitive not only to morphological processes but also to lexical diversity, syntactic variability, and domain effects, which all are expected to vary substantially across the corpora in our study. High accessor variety, for example, may arise from productive morphology, but also from topical heterogeneity or larger vocabularies. Entropic efficiency similarly reflects how evenly contexts are distributed around tokens, which can be influenced by factors such as discourse structure and stylistic variation rather than morphological productivity alone.

We calculate accessor variety and entropic efficiency using the Goldfish tokenizer of each source language for the source corpora; for the translated

corpora, we always use the same English Goldfish tokenizer that we used for our models. We use the repository by Poelman et al. (2025) for calculating the measures. Results are given in Table 8.

B.1 Source versus Translated Corpora

Source and target corpus have a positive but consistently non-significant correlation. For accessor variety, Pearson $r = 0.3324$ ($p = 0.1125$) and Spearman $\rho = 0.3487$ ($p = 0.0949$). For entropic efficiency, Pearson $r = 0.3821$ (0.0654) and Spearman $\rho = 0.3000$ (0.1544).

As discussed in Section 2, translation tends to systematically reduce morphological variation, as rich inflectional systems are often mapped onto simpler target-language constructions. Table 8 confirms this: entropic efficiency is higher in the original English FineWeb than in all FineTranslations corpora, while accessor variety results are mixed.

B.2 Correlation with Performance

We report all correlation scores with BLiMP and perplexity scores in Table 9.

BLiMP accuracy. Neither source-corpus measure predicts BLiMP accuracy at either model size. Source accessor variety and entropic efficiency are non-significant across 100MB and 1000MB models alike, with inconsistent directions between Pearson and Spearman for entropic efficiency (100MB: $r = 0.137$, $p = 0.522$, $\rho = 0.188$, $p = 0.379$; 1000MB: $r = 0.273$, $p = 0.258$, $\rho = 0.358$, $p = 0.132$). For the translated corpus, accessor variety shows a significant positive relationship with BLiMP accuracy at 100MB ($r = 0.417$, $p = 0.043$, $\rho = 0.427$, $p = 0.037$), suggesting that more contextually diverse (and thus potentially more morphologically complex) translated data achieve slightly higher grammatical accuracy. This effect however does not replicate at 1000MB ($r = -0.236$, $p = 0.330$, $\rho = -0.138$, $p = 0.574$), where the sign reverses, indicating it may be an artifact. The correlation of the entropic efficiency of the translated corpus with BLiMP scores is positive but non-significant at both scales.

Examining the individual BLiMP categories (Table 10; translated corpus only), we observe that entropic efficiency generally yields more significant correlations than accessor variety. Accessor variety produces significant correlations for the 100MB models in two categories—Argument Structure and NPI Licensing—but none for the 1000MB models. In contrast, entropic efficiency produces several positive and significant correlations. For the 100MB models, significant results appear in three categories: Anaphora Agreement (significant only under Pearson correlation), Irregular Forms, and Binding. For the 1000MB models, positive and significant correlations are also found in three categories: Anaphora Agreement, Ellipsis, and Control/Raising. However, for NPI Licensing, the correlations are significantly *negative* for both sizes.

FLORES and FineWeb perplexity. Source-corpus measures show no consistent relationship with perplexity for any domain or model size. Accessor variety and entropic efficiency from the source corpus are non-significant across all eight conditions (four domains $\times$ two model sizes).

For the translated corpus, accessor variety is generally non-significant, with a single exception: at 1000MB on Wikibooks, $r = 0.534$ ($p = 0.018$), indicating that higher contextual diversity in the translated training data is associated with *higher* (i.e. worse) perplexity in this domain at larger scale.

Entropic efficiency of the translated corpus is the most consistent predictor. At 100MB, significant negative correlations emerge for Wikibooks ($\rho = -0.576$, $p = 0.003$), Wikivoyage ($r = -0.460$, $p = 0.024$; $\rho = -0.576$, $p = 0.003$), and FineWeb ($\rho = -0.542$, $p = 0.006$); Wikinews is the sole exception ($r = -0.153$, $p = 0.474$, $\rho = -0.289$, $p = 0.171$). The same pattern holds at 1000MB: significant effects for Wikibooks ($r = -0.616$, $p = 0.005$; $\rho = -0.656$, $p = 0.002$), Wikivoyage ($r = -0.576$, $p = 0.010$; $\rho = -0.635$, $p = 0.003$), and FineWeb ($r = -0.506$, $p = 0.027$; $\rho = -0.526$, $p = 0.021$), with Wikinews again non-significant. This means that across both scales, languages whose translated corpus has a more uniform contextual distribution tend to achieve lower perplexity on English text.

	AnaAgr	IrrFrm	DetAgr	SVAgr	ArgStr	Ellips	Bindng	CtlRai	Quantf	FilGap	NPILic	IslEff	Overall
eng	91.30	79.67	82.05	73.11	70.97	69.54	70.05	64.90	63.71	60.50	51.10	43.67	66.20
arb	76.65	68.05	79.35	83.53	71.21	74.03	66.15	61.08	62.99	58.45	52.23	40.15	65.14
ces	89.75	86.90	76.90	63.90	72.07	72.73	72.62	63.70	65.11	65.40	44.60	49.08	66.08
ekk	86.70	76.70	75.05	66.60	71.91	68.13	68.03	61.72	58.51	66.35	50.57	44.50	63.98
fao	84.00	81.60	76.00	72.83	69.61	68.87	70.95	66.66	63.74	60.65	59.29	39.83	65.66
fas	74.55	67.75	78.83	67.95	70.83	69.01	64.78	60.66	59.77	58.60	50.79	44.49	63.38
heb	90.40	79.20	73.71	71.90	70.04	72.59	69.17	63.08	60.66	54.55	46.19	52.27	65.08
hsb	80.15	71.85	67.66	66.58	69.22	65.10	67.67	61.64	62.20	57.05	58.23	47.36	63.29
hun	90.35	84.55	77.58	60.02	71.58	68.53	70.27	63.78	60.90	58.90	50.17	49.01	65.11
ind	86.45	73.75	80.60	73.95	71.93	69.04	69.12	63.06	56.31	63.55	65.91	43.00	66.39
isl	86.05	84.05	79.65	76.33	70.90	70.64	71.08	65.08	63.31	71.75	45.01	41.48	65.68
jav	91.80	72.65	73.91	81.05	71.77	67.43	68.33	60.44	60.84	61.00	61.11	40.65	65.30
kan	87.50	73.60	80.97	66.42	73.58	73.27	72.05	63.40	60.87	65.75	59.27	43.16	66.83
kin	89.15	86.55	78.33	68.70	71.50	68.96	70.45	58.64	61.23	55.35	55.01	43.48	65.18
kmr	87.85	79.05	83.92	76.08	71.88	69.24	69.45	61.92	64.57	57.30	54.70	40.19	66.24
mlt	87.10	79.50	81.33	63.45	70.17	70.53	69.88	65.30	63.96	55.90	49.37	43.14	65.06
sme	83.85	75.50	80.81	89.05	69.38	65.69	67.68	62.40	64.99	65.45	65.79	42.04	67.26
swe	87.40	86.85	80.53	76.60	71.57	69.43	74.40	65.50	64.79	66.35	56.27	43.88	67.67
swh	77.05	84.35	80.64	80.90	72.92	70.69	68.12	59.72	61.81	59.80	54.37	42.41	66.00
tam	82.95	79.75	81.24	69.42	73.29	71.67	71.57	65.28	59.53	57.20	62.23	48.54	67.54
tel	80.00	77.45	81.46	71.42	71.60	71.63	71.68	67.10	62.00	62.10	58.20	44.42	66.94
ukr	89.95	83.00	72.28	71.92	72.09	68.86	69.73	61.38	63.24	70.65	48.87	41.21	64.53
urd	77.00	76.45	75.25	72.38	70.81	70.20	68.45	63.40	62.59	56.10	58.94	45.14	65.36
zsm	84.20	72.15	80.62	69.50	72.19	68.61	67.68	63.08	58.84	60.85	63.31	44.42	65.96
zul	93.10	81.35	79.44	81.00	71.27	71.47	72.63	61.16	60.60	57.40	59.57	42.36	66.96

	AnaAgr	IrrFrm	DetAgr	SVAgr	ArgStr	Ellips	Bindng	CtlRai	Quantf	FilGap	NPILic	IslEff	Overall
eng	98.65	86.15	90.70	85.18	80.26	80.95	76.01	75.98	73.07	73.77	61.07	56.76	76.01
arb	89.45	89.50	88.12	81.10	77.27	72.45	77.03	68.82	64.17	66.26	54.79	46.52	70.89
ces	97.55	93.70	90.10	85.40	79.11	80.60	75.09	71.26	61.45	65.79	53.94	57.30	73.33
ekk	95.60	92.60	86.95	81.82	78.32	74.00	77.37	70.54	68.33	64.90	52.70	54.76	72.31
fao	90.80	87.25	82.51	81.13	74.34	74.60	68.99	69.12	74.95	66.70	63.97	45.85	70.61
fas	84.20	84.25	88.75	81.58	77.03	73.75	77.56	68.28	59.65	62.64	58.86	52.98	71.26
heb	95.50	90.55	88.30	84.37	79.30	80.95	74.77	66.22	62.62	67.81	56.50	57.83	73.11
hsb	73.20	82.40	65.88	68.17	68.46	50.40	66.47	60.12	68.15	62.00	53.99	49.32	62.82
hun	96.50	96.70	86.83	84.28	79.36	76.45	76.06	71.04	62.27	65.70	56.96	54.38	72.90
ind	94.55	91.25	86.75	83.92	78.10	78.80	72.73	69.48	61.83	63.36	69.19	49.35	72.48
isl	97.55	85.20	87.85	84.82	76.48	76.60	71.81	72.98	72.25	66.31	63.04	54.36	73.37
jav	85.15	86.65	83.41	80.20	77.21	73.90	69.33	66.18	61.55	60.70	65.17	42.70	68.95
kan	93.50	89.10	86.51	83.58	79.03	74.30	72.10	71.66	61.52	64.40	61.53	54.15	72.28
kin	96.25	91.05	87.65	82.27	75.76	67.25	73.40	63.50	74.38	64.69	64.43	47.69	71.64
kmr	91.50	88.95	90.00	84.80	76.61	73.65	75.13	65.40	72.30	67.54	65.59	46.10	72.68
mlt	93.60	88.10	88.44	84.35	77.86	74.15	75.40	69.20	62.15	67.93	52.51	52.73	71.84
sme	83.05	88.30	78.59	69.82	70.26	66.20	66.74	60.34	88.67	61.96	62.83	41.16	66.89
swe	96.80	92.30	89.91	86.95	77.87	82.95	73.84	74.26	68.97	70.16	60.33	52.45	74.37
swh	95.30	90.05	89.84	83.33	77.57	69.65	72.49	64.34	80.03	67.40	58.23	52.56	72.77
tam	92.35	93.45	89.03	82.87	79.56	74.20	75.43	69.52	62.02	66.29	64.47	54.92	73.49
tel	92.40	94.35	89.15	83.52	78.82	71.65	73.20	71.54	62.52	64.83	64.50	56.15	73.36
ukr	95.80	91.75	86.24	83.83	77.89	79.75	75.13	63.90	55.50	66.47	56.10	54.39	71.48
urd	95.55	90.20	88.34	83.70	77.30	75.31	72.75	68.06	71.85	65.97	56.26	58.51	73.14
zsm	95.50	89.65	88.00	81.62	78.48	74.50	74.44	67.58	63.67	64.46	66.33	50.51	72.43
zul	92.45	84.65	86.44	82.43	76.28	72.30	75.37	63.84	64.20	63.14	64.03	44.44	70.46

Table 6: BLIMP accuracy (%) by linguistic category and language. Above: 100MB; Below: 1000MB models. Results in gray indicate that the training corpora for these low-resource languages were smaller than 1000MB.

BLiMP Category	100MB (n=24) Corpus	100MB (n=24) TTR	100MB (n=24) lang2vec	1000MB (n=19) Corpus	1000MB (n=19) TTR	1000MB (n=19) lang2vec
Overall	-0.09 (-0.09)	$+0.11$ ($+0.08$)	$+0.06$ (-0.10)	$+0.01$ ($+0.08$)	-0.32 (-0.13)	$+0.61^{**}$ ($+0.32$)
Anaphora Agr.	-0.04 (-0.06)	$+0.33$ ($+0.39$)	$+0.04$ ($+0.05$)	-0.00 ($+0.31$)	-0.01 ($+0.14$)	$+0.74^{***}$ ($+0.73$)
Irregular Forms	-0.08 ($+0.03$)	-0.04 ($+0.04$)	$+0.36$ ($+0.32$)	$+0.18$ ($+0.36$)	$+0.45$ ($+0.48$)	$+0.25$ ($+0.26$)
Det-Noun Agr.	$+0.03$ (-0.10)	-0.04 (-0.05)	$+0.01$ (-0.15)	$+0.02$ ($+0.00$)	-0.46^{*} (-0.32)	$+0.34$ (-0.13)
Subj-Verb Agr.	-0.13 (-0.18)	-0.09 (-0.09)	-0.23 (-0.14)	$+0.16$ ($+0.13$)	-0.24 (-0.15)	$+0.74^{***}$ ($+0.71$)
Arg. Structure	$+0.13$ ($+0.32$)	$+0.50^{*}$ ($+0.68$)	-0.26 (-0.24)	$+0.05$ ($+0.22$)	$+0.31$ ($+0.33$)	$+0.13$ ($+0.21$)
Ellipsis	$+0.10$ ($+0.26$)	$+0.12$ ($+0.02$)	$+0.09$ (-0.05)	$+0.53^{*}$ ($+0.48$)	$+0.21$ ($+0.23$)	$+0.61^{**}$ ($+0.75$)
Binding	-0.08 (-0.03)	$+0.13$ ($+0.15$)	$+0.27$ ($+0.19$)	$+0.28$ ($+0.38$)	-0.23 (-0.17)	$+0.36$ ($+0.02$)
Control/Rais.	-0.02 ($+0.14$)	$+0.13$ ($+0.03$)	$+0.47^{*}$ ($+0.51$)	$+0.20$ ($+0.24$)	-0.06 (-0.02)	$+0.39$ ($+0.35$)
Quantifiers	-0.21 (-0.20)	-0.22 (-0.27)	$+0.33$ ($+0.45$)	-0.37 (-0.29)	-0.45 (-0.39)	-0.04 (-0.04)
Filler-Gap	$+0.23$ ($+0.27$)	$+0.41^{*}$ ($+0.47$)	$+0.41$ ($+0.35$)	-0.09 (-0.12)	-0.39 (-0.35)	$+0.66^{**}$ ($+0.49$)
NPI Licensing	-0.16 (-0.35)	-0.05 (-0.05)	-0.36 (-0.37)	-0.03 (-0.18)	$+0.22$ ($+0.18$)	-0.35 (-0.38)
Island Effects	$+0.13$ ($+0.34$)	$+0.04$ ($+0.04$)	$+0.20$ ($+0.12$)	-0.07 ($+0.15$)	-0.35 (-0.18)	$+0.53^{*}$ ($+0.26$)

Table 7: Correlations for all BLiMP categories across model sizes, including overall BLiMP accuracy (first row). Values are Pearson r with Spearman ρ in parentheses. Significance levels: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$ (based on Pearson p-values). Note: 100MB n=24; 1000MB n=19 (excludes fao, hsb, kin, sme, zul).

	Source Corpus		Translated Corpus	
	Accessor Variety	Entropic Efficiency	Accessor Variety	Entropic Efficiency
English	6.83	0.412	–	–
Arabic	17.42	0.468	6.74	0.337
Czech	13.90	0.485	6.53	0.373
Estonian	11.84	0.548	6.68	0.362
Faroese	8.65	0.535	5.01	0.330
Persian	11.69	0.494	5.32	0.309
Hebrew	12.13	0.501	5.11	0.312
Upper Sorbian	8.48	0.438	5.74	0.330
Hungarian	17.47	0.527	6.40	0.369
Indonesian	8.89	0.347	6.59	0.342
Icelandic	11.18	0.536	5.63	0.355
Javanese	12.46	0.428	8.10	0.295
Kannada	13.83	0.575	7.03	0.323
Kinyarwanda	11.08	0.490	6.08	0.326
Kurmanji Kurdish	8.93	0.432	6.41	0.334
Maltese	14.22	0.456	5.34	0.350
Northern Sami	11.62	0.471	7.00	0.329
Swedish	9.21	0.513	5.84	0.375
Swahili	11.09	0.470	5.98	0.315
Tamil	15.22	0.540	7.65	0.314
Telugu	14.13	0.517	8.44	0.325
Ukrainian	15.80	0.449	6.15	0.349
Urdu	5.30	0.408	5.11	0.277
Malay	7.17	0.329	6.75	0.298
Zulu	18.57	0.534	6.13	0.353

Table 8: Accessor variety and entropic efficiency, as suggested by Poelman et al. (2025) as approximators of morphological complexity. Higher values indicate greater surface-form diversity.

Evaluation	n	Source corpus AV	Source corpus EE	Target corpus AV	Target corpus EE
BLiMP (100MB)	24	0.088 / −0.007 (.682) / (.974)	0.137 / 0.188 (.522) / (.379)	0.417* / 0.427* (.043) / (.037)	0.091 / 0.047 (.672) / (.826)
BLiMP (1000MB)	19	−0.241 / −0.191 (.320) / (.433)	0.273 / 0.358 (.258) / (.132)	−0.236 / −0.138 (.330) / (.574)	0.321 / 0.260 (.181) / (.283)
Wikibooks (100MB)	24	−0.260 / −0.340 (.220) / (.104)	−0.036 / −0.040 (.866) / (.853)	0.014 / 0.043 (.948) / (.842)	−0.283 / −0.576** (.180) / (.003)
Wikivoyage (100MB)	24	−0.231 / −0.253 (.278) / (.233)	−0.067 / −0.062 (.755) / (.774)	0.095 / 0.188 (.659) / (.379)	−0.460* / −0.576** (.024) / (.003)
Wikinews (100MB)	24	−0.258 / −0.242 (.223) / (.255)	−0.061 / 0.017 (.778) / (.939)	0.002 / 0.046 (.994) / (.832)	−0.153 / −0.289 (.474) / (.171)
FineWeb (100MB)	24	−0.236 / −0.226 (.267) / (.288)	−0.068 / −0.059 (.752) / (.785)	−0.040 / −0.029 (.853) / (.893)	−0.286 / −0.542** (.175) / (.006)
Wikibooks (1000MB)	19	0.058 / 0.042 (.814) / (.864)	0.013 / −0.053 (.958) / (.831)	0.534* / 0.431 (.018) / (.066)	−0.616** / −0.656** (.005) / (.002)
Wikivoyage (1000MB)	19	0.108 / 0.098 (.660) / (.689)	0.073 / 0.030 (.767) / (.904)	0.376 / 0.357 (.112) / (.133)	−0.576** / −0.635** (.010) / (.003)
Wikinews (1000MB)	19	0.119 / 0.124 (.629) / (.614)	0.151 / 0.198 (.537) / (.418)	0.379 / 0.308 (.110) / (.200)	−0.106 / −0.101 (.667) / (.681)
FineWeb (1000MB)	19	0.173 / 0.253 (.480) / (.297)	0.121 / 0.042 (.623) / (.864)	0.318 / 0.240 (.185) / (.321)	−0.506* / −0.526* (.027) / (.021)

Table 9: Pearson r and Spearman ρ correlations between Accessor Variety (AV) and Entropic Efficiency (EE) from source and target corpora and model evaluation metrics. Higher BLiMP accuracy and lower FLORES/FineWeb perplexity indicate better performance. Superscripts: $^{*}p < .05$, $^{**}p < .01$.

Category	100MB ($n = 24$) AV	100MB ($n = 24$) EE	1000MB ($n = 19$) AV	1000MB ($n = 19$) EE
AnaAgr	0.108 / 0.067 (.617) / (.754)	0.448* / 0.349 (.028) / (.094)	−0.063 / −0.138 (.799) / (.574)	0.671** / 0.691** (.002) / (.001)
IrrFrm	−0.210 / −0.264 (.324) / (.213)	0.519** / 0.533** (.009) / (.007)	0.144 / 0.146 (.555) / (.552)	0.424 / 0.411 (.070) / (.081)
DetAgr	0.264 / 0.376 (.213) / (.070)	0.080 / −0.019 (.710) / (.929)	−0.386 / −0.258 (.102) / (.286)	−0.080 / −0.165 (.746) / (.500)
SVAgr	0.165 / 0.087 (.441) / (.686)	−0.190 / −0.110 (.374) / (.607)	−0.236 / −0.341 (.331) / (.153)	0.406 / 0.411 (.085) / (.081)
ArgStr	0.505* / 0.598** (.012) / (.002)	0.003 / −0.014 (.987) / (.949)	0.194 / 0.211 (.426) / (.385)	0.334 / 0.356 (.163) / (.135)
Ellips	0.048 / 0.029 (.822) / (.892)	0.058 / 0.016 (.788) / (.941)	−0.152 / −0.050 (.534) / (.839)	0.479* / 0.461* (.038) / (.047)
Bindng	0.109 / 0.029 (.612) / (.892)	0.491* / 0.421* (.015) / (.040)	−0.352 / −0.291 (.140) / (.226)	0.205 / 0.123 (.400) / (.616)
CtlRai	0.035 / −0.051 (.870) / (.813)	0.211 / 0.252 (.321) / (.234)	0.106 / 0.211 (.667) / (.385)	0.567* / 0.604** (.011) / (.006)
Quantf	−0.216 / −0.238 (.312) / (.262)	0.274 / 0.376 (.195) / (.070)	−0.158 / −0.160 (.518) / (.514)	0.140 / 0.170 (.568) / (.486)
FilGap	0.202 / 0.358 (.344) / (.086)	0.451* / 0.395 (.027) / (.056)	−0.249 / −0.267 (.303) / (.270)	0.339 / 0.281 (.155) / (.244)
NPILic	0.409* / 0.405* (.047) / (.050)	−0.440* / −0.427* (.031) / (.037)	0.244 / 0.252 (.315) / (.298)	−0.457* / −0.460* (.049) / (.048)
IslEff	−0.084 / −0.073 (.696) / (.734)	0.022 / −0.070 (.917) / (.745)	−0.298 / −0.091 (.216) / (.710)	0.291 / 0.230 (.226) / (.344)
Overall	0.417* / 0.427* (.043) / (.037)	0.091 / 0.047 (.672) / (.826)	−0.236 / −0.138 (.330) / (.574)	0.321 / 0.260 (.181) / (.283)

Table 10: Pearson r and Spearman ρ correlations between target-corpus Accessor Variety (AV) and Entropic Efficiency (EE) and BLiMP accuracy by category. Superscripts: $^{*}p < .05$, $^{**}p < .01$.