Natural Language Processing (2025), **31**, pp. 965–981
doi:10.1017/nlp.2025.1

CAMBRIDGE
UNIVERSITY PRESS

EMERGING TRENDS

Emerging trends: translationese

Kenneth Church[1], Boyang Li[2], Peter Vickers[1], Shiran Dudy[1] and Richard Yue[1]

[1]Northeastern University, Boston, MA, USA and [2]Nanyang Technological University, Singapore, Singapore
Corresponding author: Kenneth Church; Email: k.church@northeastern.edu

(Received 11 March 2025; accepted 12 March 2025)

Abstract

Audits of multilingual resources are reporting shockingly poor quality: “less than 50% . . . acceptable quality.” There is too much translationese in too many of our multilingual resources, e.g., Wikipedia, XNLI, FLORES, WordNet. We view translationese as a form of noise that makes it hard to generalize from a benchmark based on translation to a real task of interest that does not involve translation. Worse, too much of this translationese is in the “wrong” direction. Directionality matters. Professional translators translate from their weaker language into their stronger language. Unfortunately, many of our resources translate in the other direction, from a stronger (higher-resource) language into a weaker (lower-resource) language. In Wikipedia, for example, there is more translation *out of* English than *into* English. We recommend more investments in high-quality data, and less in translation, especially in the “wrong” direction.

Keywords: translation; translation direction; large language models; linguistic resources; low-resource languages

1. Introduction and recommendations

There is too much translationese in too many of our multilingual resources. The problem appears to be worse for low-resource languages. There is a considerable literature establishing that LLMs (large language models) are more effective for high-resource languages (English, Chinese, and Arabic) than for low-resource languages (Ahuja *et al.* 2023; Koehn and Knowles, 2017; Limisiewicz *et al.* 2024; Nie *et al.* 2024; Wu and Dredze 2020).

Much has been written about risks and LLMs. These risks may be worse in low-resource languages. Guardrails appear to be more effective in high-resource languages than low-resource languages; Yong, Menghini, and Bach (2023) reported that guardrails can be circumvented by asking inappropriate questions in low-resource languages where guardrails are less effective. There are similar challenges with toxicity (Church *et al.* 2023).[a] In Nigeria, for example, tweets are more toxic in Hausa than in English because moderation processes are more effective in English.

Given the widespread use of translation from English, Anglo-centric biases are likely. Much has been written about biases in LLMs. For example, Mihalcea *et al.* (2024) is titled, “Why AI Is WEIRD. . .,” where WEIRD is an acronym for Western, Educated, Industrialized, Rich and Democratic. Ojo *et al.* (2023) asked “*How good are Large Language Models for African Languages?*” and found that LLMs for African languages are not as good as LLMs for English. We believe the root causes involve relatively poor resources, both in terms of quantity as well as quality.

One of the quality issues involves the use of translation in popular resources for training LLMs. Much of the widespread use of translation in our multilingual resources can (and should) be avoided. But when it is necessary to translate, as much as possible, we should translate in the

[a]https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html

"right" direction. As will be discussed in Section 8 and Table 8, many of our resources translate in the "wrong" direction.

What do we mean by the "right" and "wrong" direction? Figure 3 of Xu *et al.* (2023) reports higher BLEU scores for *en* → *xx* than *xx* → *en*. Assuming that machines are stronger in high-resource languages than low-resource languages, when possible, machines should translate into their stronger language and avoid translating from their stronger language. This advice follows standard practice among professional translators. Professional translators prefer to translate into their stronger language and avoid translating from their stronger language.[b] There are good reasons for this standard practice. We should do likewise when we develop LLMs.

In short, when collecting resources for training and testing LLMs, we recommend the following:

1. Invest more in quality control (by native speakers),
2. Avoid translation when possible, and
3. When translation is necessary, it is better to translate in the "right" direction:
   (a) "right" direction: weaker (low-resource) → stronger (high-resource)
   (b) "wrong" direction: vice versa.

This advice is perhaps more important for testing LLMs than training LLMs, given suggestions about the effectiveness of powerful training methods in machine learning. Such methods may be effective even in the presence of noise: Arpit *et al.* 2017; Chaudhury and Yamasaki 2021. Even so, training with less noise is likely to be more effective than training with more noise, and therefore, we should do what we can to reduce dependencies on translationese.[c]

2. Quality

There are *shocking* issues with quality in many of our resources:

> *a significant fraction contains less than 50% sentences of acceptable quality... these issues are easy to detect even for non-proficient speakers* (Caswell *et al.* 2021)

Their conclusion is based on an audit of five popular multilingual corpora:

1. CCAligned (El-Kishky *et al.* 2020),
2. ParaCrawl (Bañón *et al.* 2020; Esplà-Gomis *et al.* 2019),
3. WikiMatrix (Schwenk *et al.* 2021),
4. OSCAR (Suárez, Sagot, and Romary 2019; Ortiz Suárez, Romary, and Sagot 2020) and
5. mC4 (Xue *et al.* 2021)

Many details for many languages are reported in the appendices of Caswell *et al.* (2021), a few of which are shown in Table 1. Caswell *et al.* (2021) also concluded that quality is worse for low-resource languages, though that conclusion is based on more languages than those in Table 1.

Quality needs to be taken seriously, especially in low-resource languages. In ArtELingo-28 (Mohamed *et al.* 2024), we created a dataset with annotations of WikiArt in 28 languages including

[b] https://cbltranslations.com/blog/legal-translators-native-language/

[c] It has become standard practice to pivot via English, both in the research community and elsewhere. For example, one might have thought that the translation services in the European Union would need to support n^2 language pairs, but in practice, the bulk of their work pivots via English, reducing demand to $2n$ pairs. While this approach saves costs, there are obvious trade-offs in terms of quality. From an academic perspective, English is probably not an ideal choice as a pivot language (Anastasopoulos and Neubig 2020).

Table 1. Poor quality (based on Tables 11–14 of Caswell *et al.* (2021))

Corpus	Table	Language	Correct	Incorrect Translation	Wrong Language	Not Language	Porn
mC4	14	ha	81%		14%	5%	2%
mC4	14	hi	80%		20%	0%	3%
WikiMatrix	12	en-hi	36%	60%	1%	3%	0%
CCAligned	11	en-vi	31%	54%	1%	14%	6%
CCAligned	11	en-ha	30%	49%	9%	12%	1%

Table 2. Source text is more predictable (fluent) than target text (for these models)

Language	Model	Source	Target
French	flaubert/flaubert_base_uncased	**50.4%**	34.6%
Russian	ai-forever/ruBert-base	**49.1%**	44.7%
Spanish	dccuchile/bert-base-spanish-wwm-uncased	**38.5%**	34.6%

Hausa. This project invested 6.3K hours for annotators plus an additional 2.5K hours for coordinators (quality assurance). Most projects do not invest as much in quality control, but we hope this paper will help persuade the community to invest more in quality.

3. Translationese is more or less predictable

There have been several studies of translationese in Corpus-based Lexicography. Baker (2004) and Xiao (2009) report predictable differences in word frequencies between translationese and natural language, as one would expect. This section describes an experiment with LLMs that shows similar differences in fluency.

As will be discussed in Section 4, there is quite a bit of translation in Wikipedia. Wikipedia is porting much of their content to more languages using a combination of machine translation, post-editing, human translation, and various quality control steps.[d] Although these processes are so effective that native speakers may not be aware of the use of translation, we are able to show in Table 2 that source text is more predictable (fluent) than target text.

Table 2 is based on intermediate files published by Wikipedia. For each language pair, there is a file containing sentences in the source and target language.[e] The six scores in Table 2 are based on six samples of 10k sentences from these intermediate files, covering three languages in both source and target conditions. Table 2 shows a difference in fluency between:

1. Source: text originally written in language x, and
2. Target (Translationese): translations from some other language into x

To estimate fluency, we remove a token randomly from each test sentence and estimate how often the model correctly predicts the missing token. The scores in Table 2 report accuracies averaged over 10k sentences.

[d] https://en.wikipedia.org/wiki/Wikipedia:Content_translation_tool
[e] https://dumps.wikimedia.org/other/contenttranslation

https://doi.org/10.1017/nlp.2025.1 Published online by Cambridge University Press

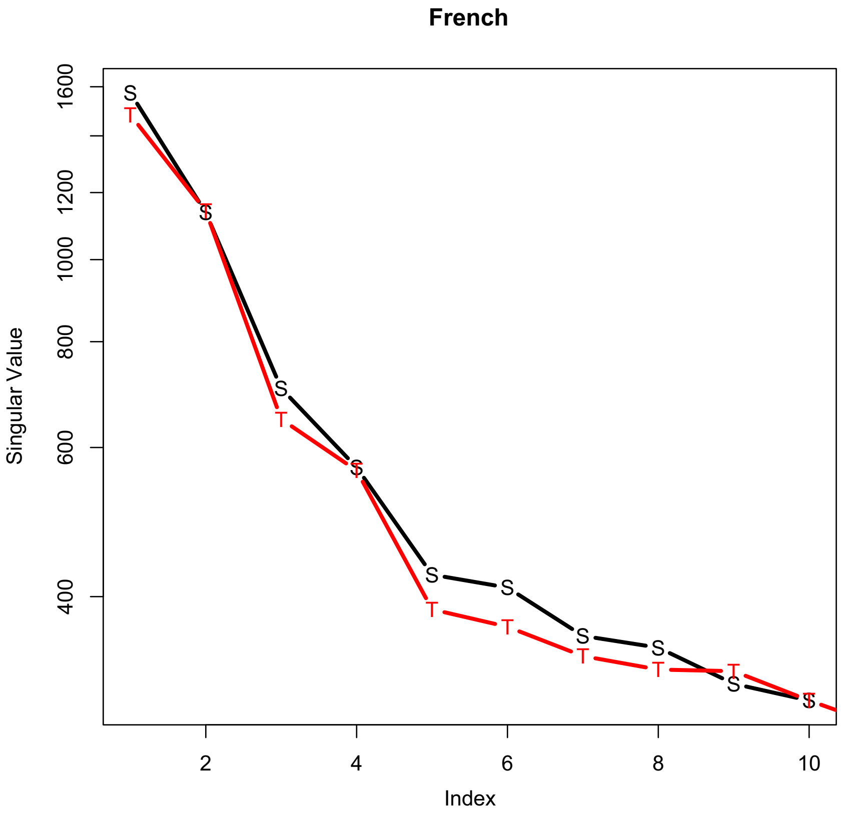

Figure 1. For the first 10 values in *D*, source (black S) is usually above target (red T).

3.1 Trace: a proposed intrinsic metric

We propose trace as an intrinsic metric to make a similar point to Table 2. This metric can be used to compare embeddings over models, languages, and different sources of text. Reliability of the proposed metric will be established by showing that scores are relatively stable over multiple batches. The machine learning literature tends to focus on extrinsic downstream evaluations such as Table 2, though intrinsic metrics may be easier to compute than extrinsic metrics.

Traces are computed from embeddings. For each model and each test set of $n = 10$k sentences in Table 2, we compute an embedding, $Z \in \mathbb{R}^{n \times d}$. After normalizing Z,[f] we use SVD to approximate $Z \approx U\, DV^T$. Trace is simply: sum(D).

Figure 1 shows the first 10 values of D for two cases: the source case (S) and the target case (T). Note that S is usually above T. Both S and T use the French model in Table 2; the difference is whether the test set contains French *with* translation (T) or *without* (S).

Figure 2 shows the pattern in Figure 1 holds over most of the 768 values in D for the French row in Table 2. The other two panels of Figure 2 are like the top panel, but for different languages, using different models and different test sets. In all three panels, the ratios are usually above the red baseline of 1.

We have found trace to be a useful figure of merit for comparing D from different models, languages, and corpora, though there are a number of alternatives:

[f] We have also experimented with Z with and without centering (removing means). Normalizing is important, but the results below do not depend too much on whether or not we remove the means, at least for the embeddings that we have looked at.

https://doi.org/10.1017/nlp.2025.1 Published online by Cambridge University Press

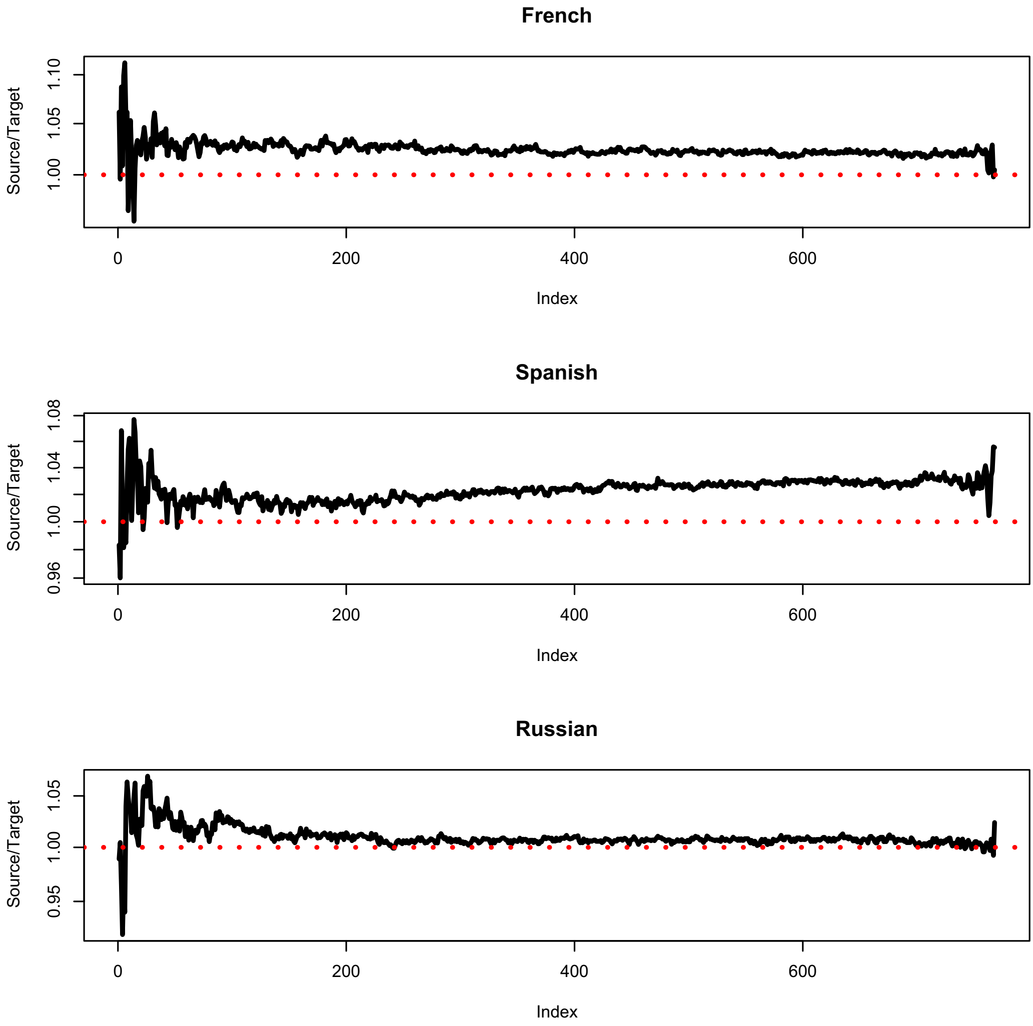

Figure 2. The pattern in Figure 1 holds over D. Ratio (S/D) is usually above 1 (red).

1. Spectral radius: $\max(D)$
2. Condition number: $\max(D)/\min(D)$
3. Trace: $\mathrm{sum}(D)$
4. Determinant: $\mathrm{prod}(D)$

We have applied many models to many batches of text in many languages. Figure 3 summarizes the trace over many experiments using data from Wikipedia with (a) two models, BGE-ME and NLLB, (b) ten languages, and (c) source *versus* translationese. All three factors are significant, but an analysis of variance (ANOVA) shows that (a) accounts for more variance than (b), and (b) accounts for more variance than (c). To see the translationese effect (c), we need to condition on the other two factors, as in Figure 4. Figure 4 shows the translationese factor for a single model (NLLB) and nine languages. The translationese effect is robust, as indicated by the relatively thin bars. The bars summarize batches. We have 60 batches for common cases (English and Spanish, with and without translation). The bars are somewhat thinner for more unusual cases where we have fewer batches, but in all cases, the translationese bar is different from the alternative.

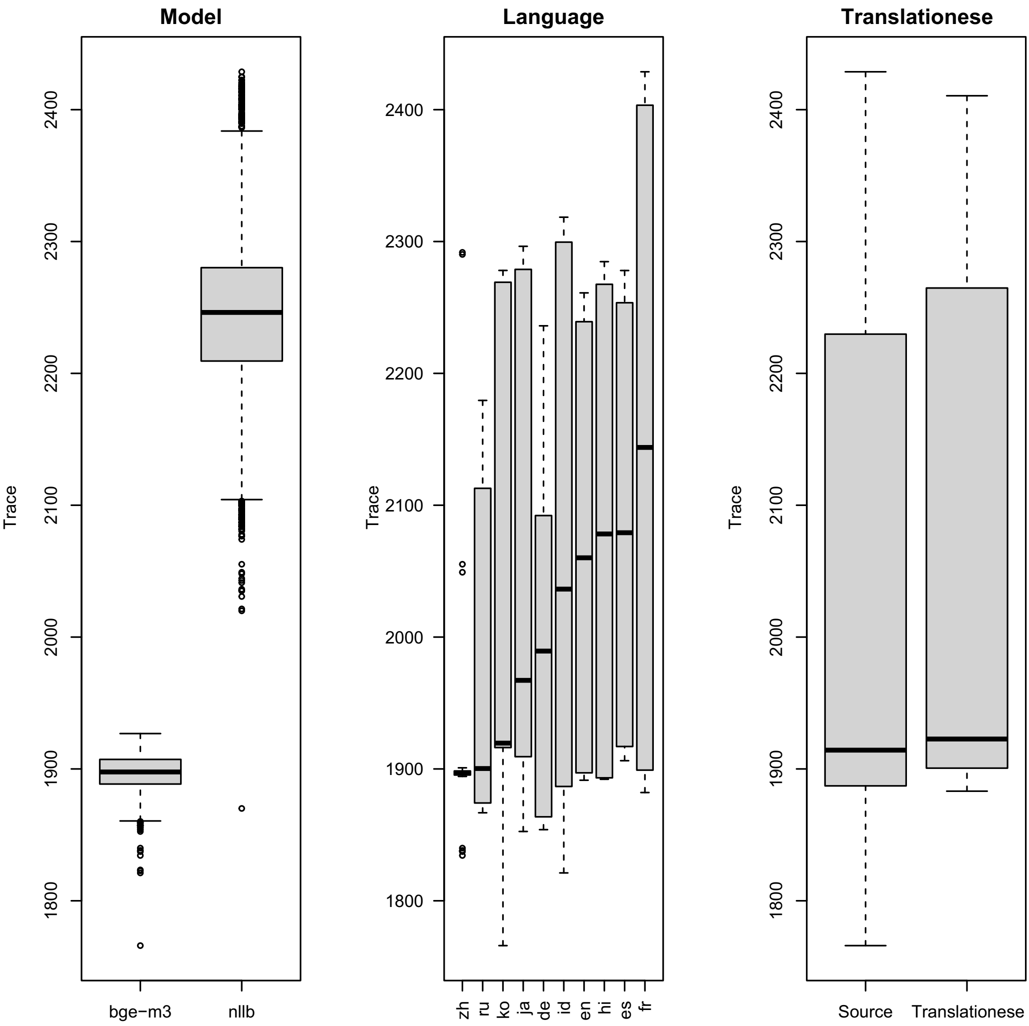

Figure 3. The trace depends on at least three factors: (a) model, (b) language, and (c) translationese. An analysis of variance (ANOVA) shows that (a) accounts for more of the variance than (b), and (b) accounts for more than (c).

That said, the directionality of the translationese effect is different from Table 2. It appears that NLLB favors translationese (except for French), unlike models in Table 2. We suspect these preferences may reflect differences in training data. In any case, since the translationese bars are different from the other bars, we conclude that translationese is not representative of natural language.

In addition to that conclusion, this section introduced a novel intrinsic metric based on traces to compare embeddings over models, languages, etc. Reliability of the proposed metric was established by showing that traces are relatively stable over batches.

4. Wikipedia is full of translationese

Although translation has played a major role in helping the community to make considerable progress, the community is also well-aware of its limitations. Unfortunately, the community may not be aware of how much translationese there is in popular corpora. Clark *et al.* (2020), for

https://doi.org/10.1017/nlp.2025.1 Published online by Cambridge University Press

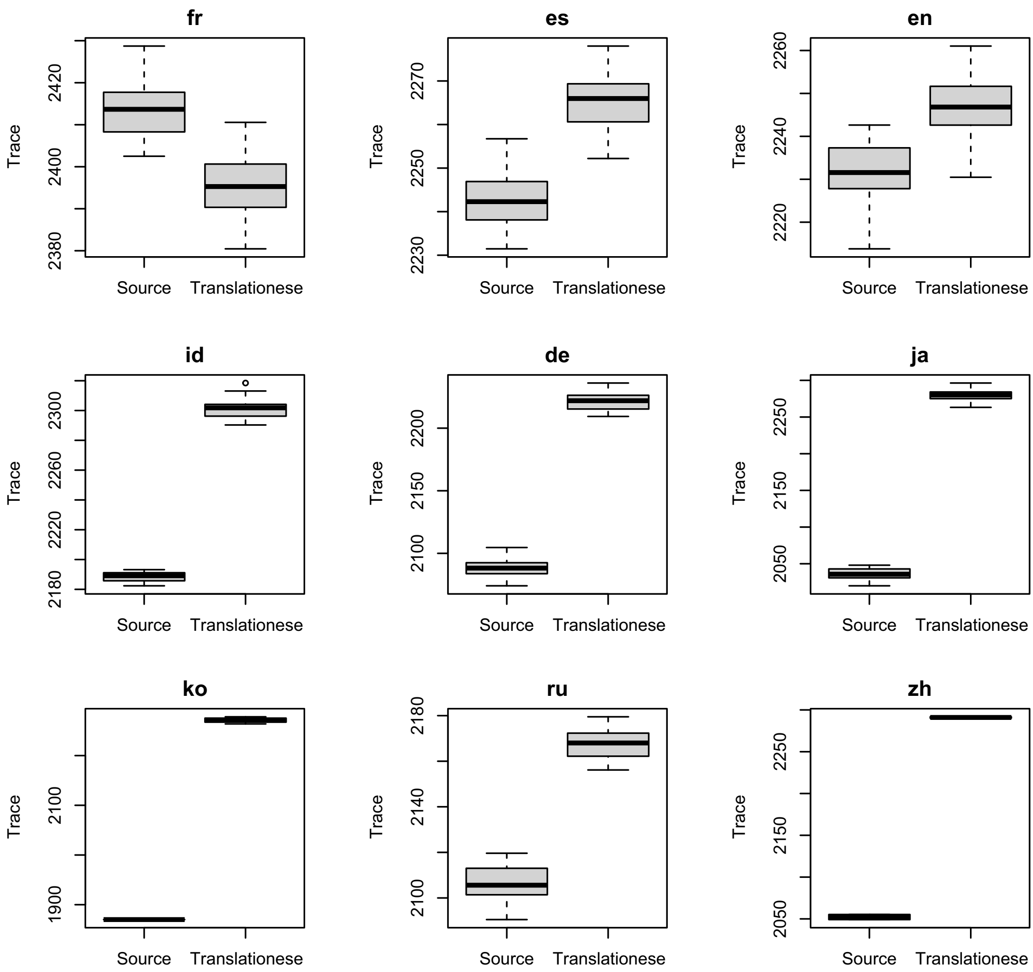

Figure 4. There are small but significant differences in traces between texts in the source language and translationese. These differences are shown for many batches over nine languages using the NLLB model.

example, start with a criticism of translationese and then recommend Wikipedia as a translation-free alternative, though in fact, much of Wikipedia is also based on translation, especially for low-resource languages such as Hausa, as will be discussed in Table 8.

We were surprised to discover that 90% of Hausa pages such as this[g] have links to similar pages in other languages such as this.[h] That is, a crawl of Hausa pages found links to the same content in other languages on 51,019 of 56,492 pages.

5. Translationese in test sets

As mentioned in the recommendations above, we are particularly concerned about the use of translation in test sets. Unfortunately, translation is common in many test sets, as illustrated in Table 3. It can be difficult to generalize performance on translationese to tasks of interest.

[g] https://ha.wikipedia.org/wiki/Harshen_Kx%CA%BCa

[h] https://en.wikipedia.org/wiki/Kx%CA%BCa_languages

Table 3. Three test sets that were translated from English to other languages

Test set in English	Translations to other languages
MultiNLI (Bowman *et al.* 2015)	XNLI (Conneau *et al.* 2018)
StoryCloze (Mostafazadeh *et al.* 2016)	XStoryCloze (Lin *et al.* 2022)
COPA (Roemmele, Bejan, and Gordon 2011)	XCOPA (Ponti *et al.* 2020)

Table 4. Examples of poorly translated Chinese in XNLI, followed by back-translations to English to demonstrate the poor quality of the Chinese

Sentence 1	Sentence 2	Label
从概念上看，奶油收入有两个基本方面 产品和地理。	产品和地理是什么是奶油抹霜工作。	neutral
Conceptually, cream income has two basic aspects, products and geography.	Product and geography is what is cream applying work.	
看看杰克逊的调查结果)	石板对杰克逊的调查结果有意见	entailment
Look at the survey result from Jackson)	The stone slate has a negative opinion on Jackson’s survey result	

Poncelas *et al.* (2020) discuss the impact of translationese on sentiment analysis, and conclude: “sentiment classifiers do not classify translated data as well as original sentences.”

Translation often makes the task harder than tasks of interest, but sometimes translationese is easier. In Table 9, we will discuss a case where translationese makes the task too easy. XNLI is an example where translationese makes the task too hard. XNLI hinges on connecting the dots between two text strings (premise and hypothesis). Translation is likely to make it harder to connect the dots, even with excellent translations.

Unfortunately, the translations in XNLI are far from excellent. For example, *cream skimming* (“to prioritize easy cases”) is translated to French as *la crème de la crème* (“the best of the best”). There are many more examples of dubious translations in XNLI in other languages such as Chinese, as illustrated in Table 4. It appears that the Wikipedia quality control processes described in footnote d are more effective than the quality control processes for translating test sets for machine learning experiments.

In summary, we view translationese as a form of noise that makes it hard to generalize from a task based on translation to a task of interest that does not involve translation. Machine learning assumes the training set is representative of (and identically distributed to) the population of interest. In general, one would expect poor translations and inadequate processes for quality control to increase noise, decreasing confidence in generalizations from the test set to tasks of interest.

Another example of translationese in test sets is FLORES Goyal *et al.* (2022), the test set for NLLB (No Language Left Behind) (Costa-jussà *et al.* 2022; NLLB Team 2024). Given the concerns about quality in section 2, one might not be surprised by quality concerns in FLORES. According to Abdulmumin *et al.* (2024), much of the Hausa in FLORES appears to be machine translation output (with little if any post-editing), and some of the “Hausa” is not even grammatical. We suspect quality is worse for translationese than for natural language. In addition, we conjecture the gap to be larger for low-resource languages than for high-resource languages.

The FLORES dev set translates a tiny set of 997 sentences from English into 200 languages. The 997 English sentences were selected from three sources: en.wikinews.org, en.wikivoyage.org,

https://doi.org/10.1017/nlp.2025.1 Published online by Cambridge University Press

and en.wikibooks.org. This sample is too small (and too specialized) to be representative of many genres of English, let alone hundreds of other languages (and many genres of interest).

6. Balanced Corpora and representative sampling

Translationese is not representative of the target language. Representative sampling and stratified sampling are important concepts in Statistics and Machine Learning. The term, *balanced corpora*, comes from lexicography. The Brown Corpus (Francis and Kučera, 1982), for example, was designed to be representative of contemporary American English when it was collected in the 1960s at Brown University. Similarly, the British National Corpus (BNC) (Aston and Burnard 1998) was designed to be representative of British English in the 1990s. Both of these corpora can be viewed as a stratified sample over genres. Fiction is different from non-fiction, and local news is different from international news. For example, the word "said" is as frequent as a function word in the news, but rare in technical abstracts. Users of these corpora are expected to scale statistics based on these strata appropriately, depending on the populations of interest (in tasks of interest).

7. Translationese in lexical resources

Translationese is also common in lexical resources such as WordNet (Miller 1995) and NRC-VAD[i] (Mohammad 2018). Both WordNet and NRC-VAD started out as lexicons for English, but later, global versions became available. For WordNet, there is a convenient NLTK interface[j] that provides translations of synsets to a few dozen languages. This approach assumes the English ontology (is-a links) is universal, which seems unlikely.

NRC-VAD is similar. NRC-VAD started out as a list of 20k English words with VAD (Valence, Arousal, and Dominance) values. Some examples are shown in Table 5. VAD can be viewed as an embedding of words into a vector space using a framework from the 1950s (Osgood, Suci, and Tannenbaum 1957). The multilingual version used machine translation (Google) to translate 20k English words to more than 100 languages, as illustrated in Table 6. This use of translation assumes the VAD values for English are universal, which seems unlikely.

8. Too much translation in the "Wrong" direction

There is often an asymmetry in directionality. Of course, many pages in Wikipedia are not translated, but when a page is translated from one language to another, it is more likely for the translation to start with a page in a high-resource language such as English than a page in a low-resource language such as Hausa. It may be a sensible policy for Wikipedia to focus on this direction since they have more content in high-resource languages, but it is not ideal for training LLMs.

As mentioned above, professional translators typically translate from their weaker language into their stronger language, and not vice versa. By this reasoning, we expect that machine translation should have better fluency when translating from a low-resource language to a high-resource language, but many resources translate in the more challenging direction: out of English as opposed to into English.

Figure 5 shows that Wikipedia has more content in high-resource languages and that most of the source for translation in Wikipedia comes from high-resource languages. The data in Figure 5 are borrowed from Tables 7–8. Table 7 estimates Wikipedia's content for some widely spoken languages. The columns in Table 7 are based on the following sources:

[i] https://saifmohammad.com/WebPages/nrc-vad.html

[j] https://www.nltk.org/howto/wordnet.html

https://doi.org/10.1017/nlp.2025.1 Published online by Cambridge University Press

Table 5. Words with extreme VAD (Valance, Arousal, Dominance) values. Very large and very small values are highlighted in bold

Word	V	A	D	Word	V	A	D
love	**1.000**	0.519	0.673	toxic	**0.008**	0.885	0.492
happy	**1.000**	0.735	0.772	nightmare	**0.005**	0.810	0.436
happily	**1.000**	0.690	0.674	shit	**0.000**	0.678	0.294
abduction	0.062	**0.990**	0.673	mellow	0.633	**0.069**	0.265
exorcism	0.163	**0.980**	0.557	siesta	0.740	**0.046**	0.295
homicide	0.010	**0.973**	0.518	napping	0.765	**0.046**	0.306
powerful	0.865	0.830	**0.991**	empty	0.188	0.183	**0.081**
leadership	0.870	0.690	**0.983**	frail	0.255	0.333	**0.069**
success	0.959	0.880	**0.981**	weak	0.180	0.241	**0.045**

Table 6. French version of Table 5 (assumes VAD values are universal)

Word	V	A	D	Word	V	A	D
aimer	**1.000**	0.519	0.673	toxique	**0.008**	0.885	0.492
heureux	**1.000**	0.735	0.772	cauchemar	**0.005**	0.810	0.436
Heureusement	**1.000**	0.690	0.674	merde	**0.000**	0.678	0.294
enlèvement	0.062	**0.990**	0.673	moelleux	0.633	**0.069**	0.265
exorcisme	0.163	**0.980**	0.557	sieste	0.740	**0.046**	0.295
homicide	0.010	**0.973**	0.518	faire la sieste	0.765	**0.046**	0.306
puissant	0.865	0.830	**0.991**	vide	0.188	0.183	**0.081**
leadership	0.870	0.690	**0.983**	frêle	0.255	0.333	**0.069**
Succès	0.959	0.880	**0.981**	faible	0.180	0.241	**0.045**

1. Joshi classification of availability of resources[k] (Joshi *et al.* 2020)
2. Speakers[l]
3. Wikipedia pages and active users[m]
4. Wikipedia views per month (averaged over 5 years)[n]
5. GDP growth over 10 years[o]

Table 8 estimates directionality in Wikipedia. Directionality is based on footnote e, a list of files for many language pairs. If we sum file sizes for files that translate out of English and files that

[k] https://microsoft.github.io/linguisticdiversity/assets/lang2tax.txt

[l] https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers

[m] https://en.wikipedia.org/wiki/List_of_Wikipedias

[n] https://wikimedia.org/api/rest_v1/metrics/pageviews/aggregate/en.wikipedia.org/all-access/all-agents/monthly/2020010100/2025010100

[o] https://en.wikipedia.org/wiki/List_of_countries_by_real_GDP_growth_rate

Table 7. Some resources for languages are in Table 8. Semantic Scholar is a promising opportunity with less translationese than Wikipedia

			Wikipedia			Semantic	GDP
Language	Joshi	Speakers	Pages	Users	Views	Scholar	Growth
English (en)	5	1515M	6,954,347	127,724	10,137M	87,761,551	2.3
Spanish (es)	5	560M	2,009,228	14,593	1204M	2,788,284	1.6
Russian (ru)	4	255M	2,028,621	9,694	1071M	501,337	1.2
French (fr)	5	312M	2,664,644	18,931	1046M	2,782,124	1.1
Chinese (zh)	5	1140M	1,463,391	7,037	724M	3,030,550	6.1
Japanese (ja)	5	123M	1,449,173	13,124	1250M	369,345	0.7
German (de)	5	134M	2,988,033	19,020	1103M	1,210,474	1.1
Vietnamese (vi)	4	86M	1,293,172	1,361	125M	64,610	6.0
Tamil (ta)	3	87M	171,718	337	19M	282	6.0
Indonesian (id)	3	199M	719,461	2,843	186M	2,082,924	4.3
Malay (ms)	3	34M	404,668	659	33M	233,470	4.1
Korean (ko)	4	81M	697,786	1,739	117M	773,283	2.5
Hausa (ha)	2	88M	54,291	180	2M	2454	2.3
Hindi (hi)	4	609M	164,829	817	69M	33,169	6.0

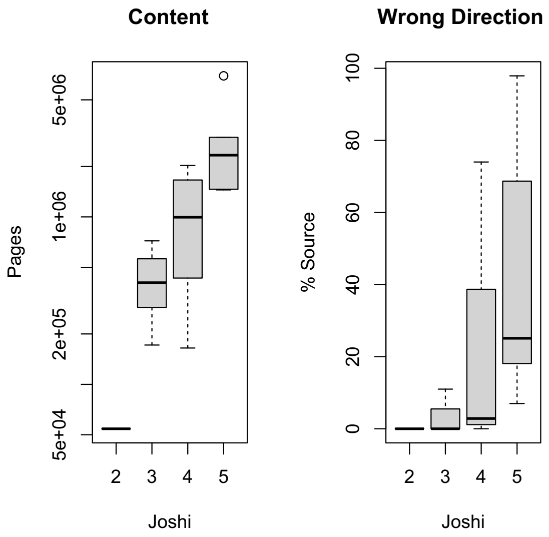

Figure 5. High-resource languages have more pages in Wikipedia (left) and are more likely to be translated to other languages (right). Data is borrowed from Tables 7–8.

Table 8. Much of Wikipedia translates in the "wrong" direction from high-resource languages near the top of the table into low-resource languages near the bottom

Language	Size (GB)	% Source	% Target	Joshi
English (en)	29.496	97.9	2.1	5
Spanish (es)	4.300	28.0	72.0	5
Russian (ru)	3.519	74.0	26.0	4
French (fr)	3.152	22.2	77.8	5
Chinese (zh)	1.017	7.0	93.0	5
Japanese (ja)	0.964	18.1	81.9	5
German (de)	0.835	68.7	31.3	5
Vietnamese (vi)	0.830	0.0	100.0	4
Tamil (ta)	0.517	0.0	100.0	3
Indonesian (id)	0.427	11.0	89.0	3
Malay (ms)	0.388	0.0	100.0	3
Korean (ko)	0.320	3.4	96.6	4
Hausa (ha)	0.312	0.0	100.0	2
Hindi (hi)	0.215	2.3	97.7	4

translate into English, we find 29.5 GB. The % Source and % Target columns estimate how much of this translation goes in one direction as opposed to the other. In general, the % Source column is larger for high-resource languages than low-resource languages. English, for example, is usually the source (97.9% by file sizes), and rarely the target. In contrast, Hausa has less (0%) source.

If we were to use translation to produce parallel corpora, it would be better to translate in the "right" direction (from Hausa to English), but when we use Wikipedia to create parallel corpora, we are translating in the "wrong" direction. It makes sense for Wikipedia to translate in the direction they did, because they are trying to port their content to more languages. It happens that they have more content in high-resource language than low-resource languages. But when we use this data for our purposes, we should be aware that this directionality is not ideal for training LLMs for low-resource languages.

Table 7 shows that Semantic Scholar (S2) is a promising untapped alternative to Wikipedia with less translationese. The challenge is how to take advantage of this opportunity. The two suggestions below involve a number of steps such as translation (in the "right" direction), LID (language identification), cross-language information retrieval, and fine-tuning LLMs.

1. create parallel corpora by translating in the "right" direction, or
2. create comparable corpora by finding comparable text in high-resource languages

9. Scaling from standard benchmarks to practical use cases

Many of the steps above go beyond the scope of this paper, but while starting to investigate opportunities for taking advantage of S2, we discovered some challenges with standard benchmarks, even on a relatively simple task: LID. In general, both for LID and many other tasks, performance tends to be better on benchmarks such as FLORES than on use cases of interest such as S2 because:

https://doi.org/10.1017/nlp.2025.1 Published online by Cambridge University Press

Table 9. LID (cld3 Botha *et al.* (2017)) performs (too) well on FLORES (Goyal *et al.* 2022) dev set (997 rows per language)

Reference	Accuracy	Correct hypothesis	Incorrect hypotheses
Vietnamese (vi)	100%	997 vi	
Hindi (hi)	100%	996 hi	1 ne
Hausa (ha)	100%	996 ha	1 su
Tamil (ta)	98%	982 ta	11 hi, 2 ne, 2 mr
Indonesian (id)	79%	785 id	208 ms, 2 eu, 1 jv, 1 it

1. Over-fitting: Off-the-shelf tools may have been trained on popular test sets.
2. Priors: Output labels (languages) are equally likely in many test sets (e.g., FLORES), but not in use cases of interest. In S2, for example, 83% is English, 2% is Indonesian, and 0.02% is Hausa.
3. Dirty data, closed-world assumptions and reject modeling: Most off-the-shelf LID tools assume the input is reasonable, and output one of n languages, but S2 is not like that. A common failure mode in Table 1 is "not language."
4. Translation: Translation is more common in popular resources than use cases of interest. There is considerable translation in FLORES, but not on S2. Most of the translation in FLORES is in the "wrong" direction. Evaluations based on translationese may not generalize to S2.

Consider LID, a relatively simple case. Table 9 shows an evaluation of an off-the-shelf LID tool, cld3 Botha *et al.* (2017), on five languages, using the dev split from FLORES (Goyal *et al.* 2022). As mentioned above, the dev split contains 997 English sentences, translated to 100 languages.

We are concerned that "your mileage may vary." That is, performance on an evaluation such as Table 9 may be misleading, if we are interested in performance on S2. Note that performance in Table 9 is (too) good, except for a confusion between Indonesian and Malay, a difficult minimal pair. Table 9 shows that cld3 correctly labeled all 977 rows in Vietnamese and all but one row in Hindi and Hausa.

Given the impressive results in Table 9, we were disappointed by the precision of cld3 on abstracts in Semantic Scholar (S2). To estimate precision on S2, we made a sample of 50 abstracts labeled Hausa and found that only 12% were Hausa.

Most of the "abstracts" labeled Hausa were not only not Hausa, but they were not even abstracts. Dirty data are a reality. A small percentage of the "abstracts" in S2 are not abstracts. Unfortunately, it appears that cld3 tends to assign many of these non-abstracts to low-resource languages such as Hausa. While performing this pilot experiment, we discovered effective prompts for rejecting non-abstracts. These prompts may also be effective for addressing "not language" defects in Table 1.

The larger take-away lesson is that there is too much translationese in too many of our resources. The consequences are bad for training LLMs, but even worse for testing. Evaluations such as Table 9 can be seriously misleading. It turns out that it is harder to take advantage of S2 than it might appear because too much of the literature is based on misleading evaluations that suggest the field is doing better than it is.

Based on evaluations such as Table 9, one might have thought that LID was a solved problem and that it would be relatively straightforward to find Hausa articles in S2. LID is a relatively simple task. Much of the literature on other steps such as cross-language information retrieval is also based on Wikipedia (and translationese), for example (Sun and Duh, 2020), though the community may not be aware of how much translationese there is in Wikipedia. Fortunately, the quality control processes on translation are more effective for Wikipedia than for many of the test sets in our field such as those in Table 3. Evaluations on those test sets may be even more problematic than evaluations based on Wikipedia, given how much Wikipedia is investing in quality control processes.

10. Conclusions and recommendations

There is good news and bad news. The bad news: there is too much translationese in too many of our resources. Audits are producing shocking conclusions: *less than 50% . . . acceptable quality* (Caswell *et al.* 2021). Worse, the community may not be aware of the poor quality. There is too much (poor) translation in popular datasets. Poor test sets produce misleading results.

As mentioned above, we should invest more in quality control. When possible, we should avoid translation. When translation is necessary, it is better to translate in the "right" direction (*into* high-resource languages) than the "wrong" direction (*out of* high-resource languages). The literature from Corpus-based Lexicography as well as experiments with LLMs in Section 3 suggest that translationese is different from texts that were originally written in the source language.

That said, ANOVA analyses mentioned above showed that other factors (models and language) account for more of the variance than translationese. Thus, translationese is, perhaps, a third-order effect. While the field is struggling with first and second-order effects, it may be appropriate to move third-order effects to the parking lot, but even so, we should be thinking about how to address these third-order effects in the future.

Too many of our resources are catch-as-catch-can. For example, since most of the content in Wikipedia happens to be in high-resource languages, when they want to port their content to more markets, it makes sense for them to translate in the direction they do. But unfortunately, this is the "wrong" direction for our purposes (developing LLMs).

The good news is that there is plenty of text in source languages of interest. For example, there are millions of articles in Semantic Scholar in Indonesian. There are opportunities to create parallel corpora by translating these articles in the "right" direction, as well as opportunities to create comparable corpora by finding comparable articles in other languages.

Competing interests. None

Ethical standards. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Author contributions. All authors approved the final submitted draft.

References

Abdulmumin I., Mkhwanazi S., Mbooi M.S., Muhammad S.H., Ahmad I.S., Putini N., Mathebula M., Shingange M., Gwadabe T.R. and Marivate V. (2024). Correcting FLORES evaluation dataset for four african languages. In Proceedings of the Ninth Conference on Machine Translation, edited by Haddow B., Kocmi T., Koehn P. and Monz C., 570–578. Miami, FL: Association for Computational Linguistics. https://aclanthology.org/2024.wmt-1.44.

Ahuja K., Diddee H., Hada R., Ochieng M., Ramesh K., Jain P., Nambi A., Ganu T., Segal S., Ahmed M., Bali K. and Sitaram S. (2023). MEGA: multilingual evaluation of generative AI. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, edited by Bouamor H., Pino J. and Bali K., 4232–4267. Singapore: Association for Computational Linguistics. https://aclanthology.org/2023.emnlp-main.258.

Anastasopoulos A. and Neubig G. (2020). Should all cross-lingual embeddings speak English? In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, edited by Jurafsky D., Chai J., Schluter N. and Tetreault J., 8658–8679. Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.766.

Arpit D., Jastrzebski S., Ballas N., Krueger D., Bengio E., Kanwal M.S., Maharaj T. (2017). A closer look at memorization in deep networks. In International Conference on Machine Learning, 233–242. PMLR. https://arxiv.org/abs/1706.05394.

Aston G. and Burnard L. (1998). *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Baker M. (2004). A corpus-based view of similarity and difference in translation. *International Journal of Corpus Linguistics* **9**(2), 167–193.

Bañón M., Chen P., Haddow B., Heafield K., Hoang H., Esplà-Gomis M., Forcada M.L., Kamran A., Kirefu F., Koehn P., Ortiz Rojas S., Pla Sempere L., Ramírez-Sánchez G., Sarrías E., Strelec M., Thompson B., Waites W., Wiggins D. and Zaragoza J. (2020). ParaCrawl: web-scale acquisition of parallel corpora. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, edited by Jurafsky D., Chai J., Schluter N. and Tetreault J., 4555–4567. Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.417.

Botha J.A., Pitler E., Ma J., Bakalov A., Salcianu A., Weiss D., McDonald R. and Petrov S. (2017). Natural language processing with small feed-forward networks. Proceedings of the 2017 conference on empirical methods in natural language processing, edited by Palmer M., Hwa R. and Riedel S., 2879–2885. Copenhagen, Denmark: Association for Computational Linguistics.

Bowman S.R., Angeli G., Potts C. and Manning C.D. (2015). A large annotated corpus for learning natural language inference. Proceedings of the 2015 conference on empirical methods in natural language processing, edited by Màrquez L., Callison-Burch C. and Su J., 632–642. Lisbon, Portugal: Association for Computational Linguistics.

Caswell I., Kreutzer J., Wang L., Wahab A., Esch D., Ulzii-Orshikh N., Tapo A., Subramani N., Sokolov A., Sikasote C., Setyawan M., Sarin S., Samb S., Sagot B., Rivera C., Rios Gonzales A., Papadimitriou I., Osei S., Suarez P.O., … Adeyemi M. (2021). Quality at a glance: an audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics* **10**, 50–72.

Chaudhury S. and Yamasaki T. (2021). Robustness of adaptive neural network optimization under training noise. *IEEE Access* **9**, 37039–37053.

Church K., Schoene A., Ortega J.E., Chandrasekar R. and Kordoni V. (2023). Emerging trends: unfair, biased, addictive, dangerous, deadly, and insanely profitable. *Natural Language Engineering* **29**(2), 483–508. doi:10.1017/S1351324922000481.

Clark J.H., Choi E., Collins M., Garrette D., Kwiatkowski T., Nikolaev V. and Palomaki J. (2020). TyDi QA: a benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics* **8**, 454–470. doi:10.1162/tacl_a_00317.

Conneau A., Rinott R., Lample G., Williams A., Bowman S., Schwenk H. and Stoyanov V. (2018). XNLI: evaluating cross-lingual sentence representations. In Proceedings of the 2018 conference on empirical methods in natural language processing, edited by Riloff E., Chiang D., Hockenmaier J. and Tsujii J., 2475–2485. Brussels, Belgium: Association for Computational Linguistics.

Costa-jussà M.R., Cross J., Çelebi O., Elbayad M., Heafield K., Heffernan K., Kalbassi E., Lam J., Licht D., Maillard J., (2022). No language left behind: scaling human-centered machine translation. arXiv preprint arXiv:2207.04672.

Esplà-Gomis M., Forcada M.L., Ramırez-Sánchez G. and Hoang H. (2019). ParaCrawl: web-scale parallel corpora for the languages of the EU. In Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks, edited by Forcada M., Way A., Tinsley J., Shterionov D., Rico C. and Gaspari F., 118–119. Dublin, Ireland: European Association for Machine Translation. https://aclanthology.org/W19-6721.

Francis W.N. and Kucera H. (1982). *Frequency Analysis of English usage: Lexicon and Grammar*. Boston: Houghton Mifflin.

Goyal N., Gao C., Chaudhary V., Chen P.-J., Wenzek G., Ju D., Krishnan S., Ranzato M.-A., Guzmán F. and Fan A. (2022). The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics* **10**, 522–538. doi:10.1162/tacl_a_00474.

Joshi P., Santy S., Budhiraja A., Bali K. and Choudhury M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, edited by Jurafsky D., Chai J., Schluter N. and Tetreault J., 6282–6293. Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.560.

El-Kishky A., Chaudhary V., Guzmán F. and Koehn P. (2020). CCAligned: a massive collection of cross-lingual web-document pairs.. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), edited by Webber B., Cohn T., He Y. and Liu Y., 5960–5969. Association for Computational Linguistics. https://aclanthology.org/2020.emnlp-main.480.

Koehn P. and Knowles R. (2017). Six challenges for neural machine translation. In Proceedings of the First Workshop on Neural Machine Translation, edited by Luong T., Birch A., Neubig G. and Finch A., 28–39. Vancouver: Association for Computational Linguistics. https://aclanthology.org/W17-3204.

Limisiewicz T., Blevins T., Gonen H., Ahia O. and Zettlemoyer L. (2024). MYTE: morphology-driven byte encoding for better and fairer multilingual language modeling. In Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: long papers), edited by Ku L.-W., Martins A. and Srikumar V., 15059–15076. Bangkok, Thailand: Association for Computational Linguistics.

Lin X. V., Mihaylov T., Artetxe M., Wang T., Chen S., Simig D., Ott M., Goyal N., Bhosale S., Du J., Pasunuru R., Shleifer S., Koura P.S., Chaudhary V., O{'}Horo B., Wang J., Zettlemoyer L., Kozareva Z., Diab M., Stoyanov V. and Li X. (2022). Few-shot learning with multilingual generative language models. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, edited by Goldberg Y., Kozareva Z. and Zhang Y., 9019–9052. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. https://aclanthology.org/2022.emnlp-main.616.

Mihalcea R., Ignat O., Bai L., Borah A., Chiruzzo L., Jin Z., Kwizera C., Nwatu J., Poria S. and Solorio T. (2024). Why AI is WEIRD and should not be this way: towards AI for everyone, with everyone, by everyone. arXiv preprint arXiv:2410.16315.

Miller G.A. (1995). WordNet: a lexical database for english. *Communications of the ACM* **38**(11), 39–41.

Mohamed Y., Li R., Ahmad I. S., Haydarov K., Torr P., Church K. and Elhoseiny M. (2024). No culture left behind: ArtELingo-28, a benchmark of WikiArt with captions in 28 languages. In Proceedings of the 2024 conference on empirical methods in natural language processing, edited by Al-Onaizan Y., Bansal M. and Chen Y.-N., 20939–20962. Miami, Florida, USA: Association for Computational Linguistics.

Mohammad S. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers), 174–184. Melbourne, Australia: Association for Computational Linguistics.

Mostafazadeh N., Chambers N., He X., Parikh D., Batra D., Vanderwende L., Kohli P. and Allen J. (2016). A corpus and cloze evaluation for deeper understanding of commonsense stories. In Proceedings of the 2016 conference of the north American chapter of the association for computational linguistics: human language technologies, edited by Knight K., Nenkova A. and Rambow O., 839–849. San Diego, California: Association for Computational Linguistics.

Nie Z., Feng Z., Li M., Zhang C., Zhang Y., Long D. and Zhang R. (2024). When text embedding meets large language model: a comprehensive survey. arXiv preprint arXiv:2412.09165.

NLLB Team (2024). Scaling neural machine translation to 200 languages. *Nature* **630**, 8018–8841.

Ojo J., Ogueji K., Stenetorp P. and Adelani D.I. (2023). How good are large language models on African languages? ArXiv abs/2311.07978.

Ortiz Suárez P.J., Romary L. and Sagot B. (2020). A monolingual approach to contextualized word embeddings for mid-resource languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, edited by Jurafsky D., Chai J., Schluter N. and Tetreault J., 1703–1714. Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.156.

Osgood C.E., Suci G.J. and Tannenbaum P.H. (1957). *The Measurement of Meaning*. University of Illinois Press. https://api.semanticscholar.org/CorpusID:144731157.

Poncelas A., Lohar P., Hadley J. and Way A. (2020). The impact of indirect machine translation on sentiment classification. In Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track), edited by Denkowski M. and Federmann C., 78–88. Virtual: Association for Machine Translation in the Americas. https://aclanthology.org/2020.amta-research.7.

Ponti E. M., Glavaš G., Majewska O., Liu Q., Vulic I. and Korhonen A. (2020). XCOPA: a multilingual dataset for causal commonsense reasoning. arXiv preprint arXiv:2005.00333.

Roemmele M., Bejan C.A. and Gordon A.S. (2011). Choice of plausible alternatives: an evaluation of commonsense causal reasoning. In 2011 AAAI spring symposium series. https://people.ict.usc.edu/~gordon/publications/AAAI-SPRING11A.PDF.

Schwenk H., Chaudhary V., Sun S., Gong H. and Guzmán F. (2021). WikiMatrix: mining 135M parallel sentences in 1620 language pairs from Wikipedia. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, edited by Merlo P., Tiedemann J. and Tsarfaty R., 1351–1361. Association for Computational Linguistics. https://aclanthology.org/2021.eacl-main.115.

Suárez P.J.O., Sagot B. and Romary L. (2019). Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In 7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7). Leibniz-Institut für Deutsche Sprache.

Sun S. and Duh K. (2020). CLIRMatrix: a massively large collection of bilingual and multilingual datasets for cross-lingual information retrieval. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), edited by Webber B., Cohn T., He Y. and Liu Y., 4160–4170. Association for Computational Linguistics. https://aclanthology.org/2020.emnlp-main.340.

Wu S., Dredze M., Gella J.W., Rei M., Petroni F., Lewis P., Strubell E., Seo M. and Hajishirzi H. (2020). Are all languages created equal in multilingual BERT? In Proceedings of the 5th Workshop on Representation Learning for NLP, edited by Gella S., J. Welbl, Rei M., Petroni F., Lewis P., Strubell E., Seo M. and Hajishirzi H., 120–130. Association for Computational Linguistics. https://aclanthology.org/2020.repl4nlp-1.16.

Xiao R. (2009). How different is translated Chinese from native Chinese. *International Journal of Corpus Linguistics* **15**(1), 5–35.

Xu H., Kim Y.J., Sharaf A. and Awadalla H. (2023). A paradigm shift in machine translation: boosting translation performance of large language models. ArXiv abs/2309.11674.

https://doi.org/10.1017/nlp.2025.1 Published online by Cambridge University Press

Xue L., Constant N., Roberts A., Kale M., Al-Rfou R., Siddhant A., Barua A. and Raffel C. (2021). MT5: a massively multilingual pre-trained text-to-text transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, edited by Toutanova K., Rumshisky A., Zettlemoyer L., Hakkani-Tur D., Beltagy I., Bethard S., Cotterell R., Chakraborty T. and Zhou Y., 483–498. Association for Computational Linguistics. https://aclanthology.org/2021.naacl-main.41.

Yong Z.-X., Menghini C. and Bach S.H. (2023). Low-resource languages jailbreak gpt-4. In NeurIPS Workshop on Socially Responsible Language Modelling Research (SoLaR). https://arxiv.org/abs/2310.02446.

Cite this article: Church K, Li B, Vickers P, Dudy S and Yue R (2025). Emerging trends: translationese. *Natural Language Processing* **31**, 965–981. https://doi.org/10.1017/nlp.2025.1