A Dataset for Probing Translationese Preferences in English-to-Swedish Translation

Jenny Kunz **Anja Jarochenko** **Marcel Bollmann**
Department of Computer and Information Science
Linköping University
jenny.kunz@liu.se anjja777@student.liu.se marcel.bollmann@liu.se

Abstract

Translations often carry traces of the source language, a phenomenon known as *translationese*. We introduce the first freely available English-to-Swedish dataset contrasting translationese sentences with idiomatic alternatives, designed to probe intrinsic preferences of language models. It includes error tags and descriptions of the problems in the original translations. In experiments evaluating smaller Swedish and multilingual LLMs with our dataset, we find that they often favor the translationese phrasing. Human alternatives are chosen more often when the English source sentence is omitted, indicating that exposure to the source biases models toward literal translations, although even without context models often prefer the translationese variant. Our dataset and findings provide a resource and benchmark for developing models that produce more natural, idiomatic output in non-English languages.

Keywords: Translationese, Idiomaticity, Machine Translation, Evaluation, Minimal Pair Probes

1. Introduction

Translations have long been known to carry traces of the source language and differ in style and features from texts originally written in the target language; a phenomenon commonly called *translationese* (Gellerstam, 1986). Research shows that translationese is particularly strong in machine translation, leading to simplified and less varied language marked by reduced lexical and morphological richness (Vanmassenhove et al., 2021). Such output is often easily distinguishable from original text (Koppel and Ordan, 2011; Li et al., 2025). Recent work suggests that LLMs produce less literal translations (Raunak et al., 2023). However, even though their outputs show increased lexical diversity compared to specialized machine translation systems, they can still be reliably distinguished from human-written text (Kong and Macken, 2025).

For many languages, it is common to use translated datasets especially for LLM evaluation (Nielsen, 2023; Bandarkar et al., 2024) and instruction tuning (Li et al., 2023; Dac Lai et al., 2023; Holmström and Doostmohammadi, 2023), as there is often no practical alternative. In addition, for low-resource languages, but to a lesser extent even for high-resource languages such as English, large portions of modern web-crawled corpora that constitute the training data of LLMs are translations (Thompson et al., 2024). It is therefore important to analyze this problem and assess its extent, in order to move towards models that produce natural, idiomatic output in non-English languages.

Kunz (2026) investigate conventionalized idiom knowledge and translationese in LLM outputs for Swedish. Analyzing sentence pairs from a book on translationese (Katourgi, 2022), which contains translations with translationese phrasing alongside idiomatic alternatives suggested by the author, they examine whether the model assigns higher probability to the translationese version or to the idiomatic alternative. Building on this work, we construct a dataset for probing translationese preferences. Our dataset is similar in its setup, but addresses several limitations of the previous dataset: it is fully open under a permissive license, includes the English source sentence with preceding context to also study preferences in a translation context, and provides annotations tagging the type of translationese, enabling more fine-grained analysis of model behavior. To our knowledge, this is the first freely available dataset explicitly contrasting translationese with idiomatic alternatives for Swedish.

We provide a detailed description of the dataset, including both qualitative and quantitative analyses of translationese examples in English-to-Swedish translations produced by a smaller specialized machine translation system. We also compare these to translations generated by a state-of-the-art LLM, highlighting that while it does not fully resolve any specific problem, it is much better at generating more idiomatic words and phrasings.

In experiments probing the intrinsic preferences of smaller multilingual LLMs with our dataset, findings show a strong bias toward the translationese phrasing. Notably, models favor the human alternative more often when omitting translation context in the prompt, indicating that exposure to the English source sentence biases them toward the more literal, translationese wording. However, increasing the context window helps the model become less biased towards the translationese variant.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8767–8779
11–16 May 2026.

Release We release our dataset on HuggingFace under https://huggingface.co/datasets/liu-nlp/translationese-opensubtitles. A full version including all annotations with further explanations, including the code to reproduce our experiments, is available on GitHub: https://github.com/jekunz/translationese.

2. Background

Translationese Translated texts often preserve features of the source language while avoiding target-language constructions that diverge more strongly from the source language. This is just one aspect of a phenomenon known in both human and machine translation as *translationese* (Gellerstam, 1986). Translationese is not generally a sign of poor translation quality (Gellerstam, 2005) but the resulting texts have properties that are different from idiomatic language. As Baker (1993) shows, translated texts tend to be more explicit than their originals, are lexically and syntactically simpler, and follow a more conventional style. Koppel and Ordan (2011) show that classifiers can easily distinguish between original and translated text, and even determine the source language of the translated text. Kong and Macken (2025) show that this still holds true with LLMs, as distinguishing human from LLM-translated text works almost perfectly. Vanmassenhove et al. (2021) show that machine translation systems tend to simplify language, losing lexical and morphological variety, but that Transformer-based models preserve more diversity than earlier systems. Similarly, Kong and Macken (2025) find that LLMs generate more varied text than specialized machine translation systems, yet their outputs remain clearly distinguishable from human-written text. Bizzoni et al. (2020) compare human and machine translationese, finding that while human translations have similar properties across modalities (written versus spoken), machine translations exhibit fundamentally different patterns. Li et al. (2025) let annotators mark unnatural parts of translations, distinguishing between rigid sentence structures and literal word choices. They find that LLM outputs still contain much translationese, and trace this problem to translationese in the training data. They propose a polishing step where the model revises its own translation to reduce this effect, while prompts for natural style do not consistently help.

Other studies have examined idiom translation as a common case of overly literal translation. Fadaee et al. (2018) show that systems often translate idioms word by word, leading to semantic errors. Dankers et al. (2022) also find that models tend to treat idioms compositionally, i.e., as literal phrases. When they recognize idioms as non-compositional units, interactions between the idiom's parts and the surrounding context decrease. We annotate "idioms" as one category of error in our dataset to enable further research into this phenomenon.

English-to-Swedish translationese Gellerstam (1986) gives the classic example that English often uses an adjective together with a generic noun (e.g. "an [ADJ] thing"), whereas Swedish prefers to use a pronominal construction ("something [ADJ]") instead. In translations from English, such noun constructions therefore appear more frequently than in original Swedish texts—e.g., "a silly thing happened" may be literally translated as *en fånig sak hände* instead of the more idiomatic *något fånigt hände* ("something silly happened"). Katourgi (2022) documents other typical features of translation-influenced Swedish: Participial forms are less common than in English, e.g., *Ta en bild på dig själv tittandes in i kameran* ("Take a picture of yourself looking into the camera") versus the more idiomatic *när du tittar mot kameran* ("...when you look at the camera"). Also, noun phrases in predicative expressions often use the article when it should be omitted, e.g., *Jag är en översättare* → *Jag är översättare* ("I am a translator" → "I am translator"). Ahrenberg (2021) analyzes adjective usage in English–Swedish translation and reports systematic distributional differences: human translators are more likely to restructure phrases, whereas systems tend to follow the source text more closely. Ahrenberg (2017) compares a human and a machine translation of an article, reporting similar findings: the system adheres closely to the source, while the human employs strategies such as word reordering and sentence splitting, resulting in a longer text with a slightly higher type–token ratio.

Error tags for translations For our dataset and analysis, we developed a custom error-tagging system. The tags were created through an iterative process in order to address the specific linguistic issues observed in our data. A related framework for error classification and assessment is the Multidimensional Quality Metrics (MQM; Lommel et al., 2024). MQM is currently regarded as a standard framework for analytic Translation Quality Evaluation. It consists of two central components: an *error typology*, which provides a hierarchical classification of errors, and a *scoring model*, which defines how identified errors are quantified (with different approaches depending on sample size). The MQM error typology is organized into seven main categories, with each one of them containing subcategories and subtypes of those. The scoring model is a combined method, process, and formula designed to derive overall quality scores from identified errors, either in calibrated or non-calibrated settings. Typically, the evaluation follows guidelines

and specifications defined for a particular task or customer. Identified errors receive a quality score through assigned penalty points or weights, which are then aggregated in a record or scorecard. While the MQM tag set has some overlap with our custom tag system, it also differs in important ways. In particular, it presents limitations for our analysis of idiomatic language use, where we need more fine-grained tags. The relationship between our tag system and the MQM tags is discussed in more detail in Appendix A.

3. Dataset Construction

Our dataset consists of 600 sentences from the English part of OpenSubtitles (Lison and Tiedemann, 2016), a dataset consisting predominantly of spoken dialogue. Sentences were translated to Swedish with OPUS-MT (Tiedemann and Thottingal, 2020) as an example for a neural (but not LLM-based) translation system, and GPT-5 (OpenAI, 2025) as an example for a recent LLM. For each sentence, we provide error tags for each of the machine translations, an alternative translation produced by a human annotator, a contextual explanation, and a problem and solution description.

3.1. Annotation Process

The dataset was created and revised by two cognitive science students, both native Swedish speakers with basic linguistic training. The main annotator sampled random sections of the source document, collecting sentence pairs where OPUS translations showed signs of translationese. For each pair, they added a brief context description, outlined the problem and its solution, and assigned up to three error tags (see Section 3.2). The annotator then proposed more idiomatic alternative translations based on intuition and supported by dictionaries. A third translation by GPT-5 was added and evaluated against the human alternative by the second annotator. GPT-5 translations judged equally good or better were marked and supplemented with comments on their strengths or possible improvements. GPT-5 translations containing errors were tagged using the same scheme as the OPUS translations. The process was repeated several times for quality control, including the removal or replacement of problematic or duplicate entries.

3.2. Error Tags

Each phrase in the dataset is annotated with up to three tags to capture the types of issues encountered in the OPUS translation. For the GPT-5 translation, we mark if the translation is acceptable or even improved in comparison to the human one. We introduce three tags indicating **major errors**: *Grammar* (GR) is used for grammatical or syntactical errors, *missing* (SAK) for missing or not translated parts and *incorrect* (LF) for sentences containing words that are incorrect in the given context. *Loss of meaning* (BET) is used in combination with other tags and marks sentences where errors are critical enough to cause significant loss of the original phrase's meaning. *Additional information* (ADD) is used when there are unnecessary added words in the translation. Multiple tags are used if there is need to capture multiple errors or connections between issues, e.g. if missing words cause loss of meaning. Two **minor error** tags mark up less critical, but still significant issues affecting interpretation. *Semantic* (SEM) indicates subtler changes in meaning that come with a risk of misinterpretation. *Lexical preference* (PR) is used for translations containing less normative, inappropriate or unnatural words from a fluent speaker's perspective. Finally, we use three **descriptive tags** to indicate the presence of certain types of language that are known to cause issues in machine translations: an *idiom* (ID) tag for idioms, a *slang* (SL) tag for informal language, a *style* (ST) tag for when the context requires domain-specific language and a *direct translation* (DIR) tag for when the translation is a noticeable direct translation from the source language. The descriptive tags are used to highlight causes of errors, combined with other tags that specify the issues that are connected to the original phrases of either category, e.g. idioms or slang expressions losing their meaning because of literal translation, or domain-specific terms not being applied where they would be preferrable.

4. Dataset Analysis

Table 1 shows basic token statistics for the Swedish translations from different sources, computed on the lower-cased text with punctuation marks removed. The human translations are longest on average, both in terms of characters and (whitespace-separated) word tokens, and show the highest type–token ratio, which is in line with previous analyses of translationese (e.g., Ahrenberg, 2017). The GPT-5 translations are closer to the human translations than the ones provided by OPUS-MT.

Figure 1 shows the counts of each error tag by translation model. The "minor" error tags SEM and PR are the most common, followed by loss of meaning (BET) or incorrectly-translated words or phrases (LF). The GPT-5 translations consistently have fewer problems than the OPUS translations, with the largest error reductions observed for PR and SAK, indicating that GPT-5 is much better at generating more idiomatic words or phrasings, and less likely to omit words entirely in the translation.

	OPUS	Human	GPT-5
# word tokens	3,860	3,971	3,913
# characters	16,070	17,208	16,854
Type–token ratio	33.03	35.00	34.83

Table 1: Basic statistics for the Swedish translations

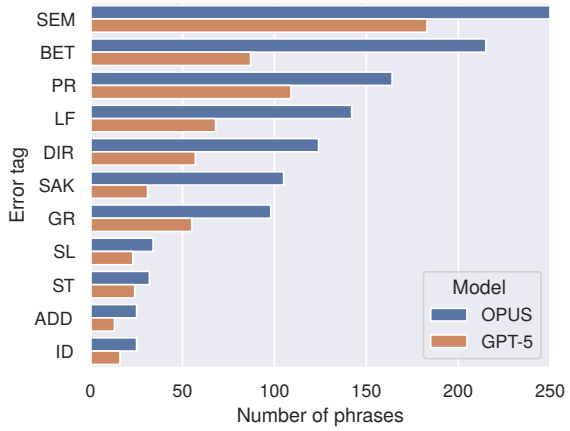

Figure 1: Error tag distribution for the Swedish translations; cf. Sec. 3.2 for an explanation of tags.

Comparing annotator judgments for GPT-5 versus the human translator, we find that the GPT-5 translation is equally acceptable in ca. 40% of cases, and even judged to be an improvement over the human translation for 36 examples (6%).

4.1. Causes of Minor Errors

Semantic shift (SEM) The most common error tag in our dataset is SEM, indicating that the situational or emotional meaning of the phrase is changed in a way that leaves room for misunderstanding. For example, the phrase in (1) is used in a context where the addressee is exhausted but doesn't go to bed, and therefore is told to rest:

(1) You take rest
  a. *Ta det lugnt* 'take it easy' (SEM)
  b. *Du ska vila dig* 'you should/will rest' (SEM)

The OPUS translation (1 a) captures that the person needs to calm down in some manner, but does not communicate the form of rest; the GPT-5 translation (1 b) carries a more aggressive connotation than "care." An idiomatic translation would be *Ta och vila*, roughly translating to 'take some rest.'

Metaphoric language is a common cause of this error type. The phrase in (2) is used in the context of policemen offering help to an alleged criminal, and translated identically by OPUS and GPT-5:

(2) We could make you look clean
  Vi kan få dig att se ren ut (SL, SEM)

The translation here carries a literal meaning and could also be interpreted as drug-related. To express this is in a more idiomatic way, the word *fläckfri* 'spotless' is more precise than *ren* 'clean'.

(3) They just fade in
  a. *De bara bleknar in* (SEM)
  b. *De bara tonar in* 'they just tune in' (LF, BET)

Example (3) is used in the context of black people "fading in" in society. The OPUS translation (3 a) uses the word *bleknar* 'to become pale', which is possible to understand, but not a commonly accepted usage of the word. GPT-5 produces a worse translation (3 b) in this case, using the word *tonar* that can be used in reference to "fading in music" or to "engage in something," which communicates the opposite of the original phrase, and thus gets annotated with different error tags in our dataset. A more idiomatic, and still metaphorical, phrasing that is appropriate in the given context would be *De bara smälter in*, literally 'they just melt in.'

Lexical preference (PR) This tag indicates translations that communicate the meaning but sound unnatural to a native speaker. For example, while *gåva* in Swedish does mean 'gift,' it has a broader and more formal meaning than in English, and sounds unnatural in a more casual context:

(4) Simone, your gift
  Simone, din gåva (PR)

A more appropriate choice here is *present* 'present', which has a less formal connotation in Swedish.

Translations can also be too informal, as in (5), where the English phrase communicates the seriousness of a crime in a police investigation:

(5) We're talking murder, here
  Vi pratar om mord (SAK, PR)
  'We're speaking of murder here'

Both OPUS and GPT-5 pick the word *pratar* 'speak', which is not wrong, but makes the information sound less serious than it is. In a context like this, *talar* 'talk' would be a more formal choice, thus conveying the gravity of the situation better.

(6) Poor little thing
  Stackars lilla sak (DIR, SEM, PR)

The phrase in (6) is said to comfort another person and sympathize with them; the direct, literal, translation uses the word *sak* 'thing', which in Swedish

is only applicable when referring to items, not persons. To call someone an "item" is not something one would do in a serious and emotional situation, so we use SEM to highlight that this sentence could be interpreted as unserious or rude. A common and lexically preferred phrase to would be *Din stackare*, roughly translating to 'you poor one/person.'

4.2. Causes of Major Errors

Loss of meaning (BET) The most common "major" problem we observe is that the phrase's meaning gets lost in the translation. In (7), the word "dope" is used as an intensifier synonymously with "cool" or "awesome," to describe a watch, whereas OPUS literally translates it to "drug watch":

(7) This dope watch

Den här knarkklockan 'this drug watch' (SL, DIR, BET)

A more correct translation would be *den här fräna klockan*, using the adjective *fräna* 'cool, stylish' that is commonly used in such contexts.

Idioms (ID) are a common source of this type of error, as in Example (8), which talks about the structural properties of a house:

(8) It's got good bones

a. *Den har bra ben* (ID, DIR, BET)

b. *Den har bra grundförutsättningar* 'it has good basic prerequisites' (ID, SEM)

The literal OPUS translation (8 a) loses the idiomatic meaning in this context. GPT-5 correctly captures this meaning in (8 b), though the phrasing causes a subtle semantic shift due to the idiomaticity being lost. An idiomatic alternative that is applicable in the context of housing would be *Den har bra stomme* 'it has good framing.'

Incorrect word choice (LF) Similar to (8), Example (9) contains an idiom where the OPUS translation (9 a) loses the meaning by being overly literal. It also translates the word 'pick' as *välja* 'choose', which is not the correct choice in this context, and therefore gets annotated with the LF tag:

(9) I had an old bone to pick with you

a. *Jag hade ett gammalt ben att välja med dig* 'I had an old bone to choose with you' (ID, LF, BET)

b. *Jag hade ett gammalt horn i sidan på dig* 'I had an old horn in your side' (ID, LF, BET)

The GPT-5 translation (9 b) attempts to use the Swedish idiom *ett horn i sidan*, related to the English 'a thorn in the side,' but this is neither fitting nor used correctly here. A Swedish idiom that would fit better here is *Jag hade en oplocked gås med dig*, literally 'I had an unplucked goose with you,' used in the same way as the original phrase.

(10) He's a degenerate gambler

Han är en degenererad spelare 'he is a degenerated player/gambler' (LF, BET, PR)

Example (10) mistranslates 'degenerate' by applying the closely related adjective *degenererad* 'degenerated', thus also losing the meaning. To capture the intent of the original utterance, we could use the phrase *Han är en nedgången speltorsk*, roughly 'he is a worn-out gambling addict.'

Grammatical errors (GR) Example (11) illustrates a case where both OPUS and GPT-5 introduce grammatical errors into the translation by attaching 'all' to the object rather than the subject:

(11) We all miss you

a. * *Vi saknar dig alla* (GR)

b. * *Vi saknar dig allihop* (GR)

Both translations rather suggest the meaning 'We miss you all,' but are grammatically incorrect as *alla* 'all' and *allihop* 'all (together)' are plural pronouns, where *dig* 'you' is exclusively singular. The correct translation should follow the same word order as in English, i.e. *Vi alla saknar dig*.

We also observe examples of incorrect article use that were described by Katourgi (2022), e.g. in predicative constructions describing a person's occupation, where the article *en* needs to be omitted for a grammatically correct translation:

(12) You're a detective

* *Du är en detektiv* (GR)

Missing words (SAK) This category is much more common in the OPUS translations than in GPT-5. In many cases, the missing words do not affect the translation significantly, e.g. when a sentence-initial 'Come on' or 'Listen, ...' is omitted, or in Example (13), where the verb is omitted from the imperative clause, causing no errors or misunderstandings, but still differs from the original.

(13) Put your hands on your head!

Händerna på huvudet! 'Hands on your head!' (SAK)

(14) Why don't you go get some sleep?

a. *Gå och lägg dig.* 'Go to sleep' (SAK, SEM)

b. *Varför går du inte och sover lite?* 'Why don't you go and get some sleep?' (SEM)

Figure 2: Prompting setups. Each box shows a **minimal pair**: Sentence 1 is a translationese variant, sentence 2 is an idiomatic variant of a translation of the same English sentence. We compute the perplexity of each variant to determine which one the model prefers. Translations of text in the figure: **Translate the following sentence to Swedish.* †*Translate the following sentence to Swedish, considering the context.*

Example (14) is used to suggest, in a caring way, that a person should go to bed. The OPUS translation (14 a) is missing a large part of the source sentence and reads more like an order. GPT-5 produces a direct translation (14 b) of the original; however, in Swedish, this shifts the phrase's tone from caring to questioning. A more idiomatic phrasing would be *Ska inte du gå och försöka sova lite?*, roughly 'Shouldn't you go and try sleeping a little?'

4.3. Domain-Specific Language

Some causes of errors that we annotate with their own tag have already been mentioned above, such as slang words (SL; Ex. 7) and idioms (ID; Ex. 8 and 9). Domain-specific language (ST), i.e. phrases that use terminology specific to a certain field or domain, are another common error source in our dataset. An example of this is the legal domain, which often uses very specific terminology:

(15) Not guilty on all counts
Inte skyldig på alla punkter (ST, PR)

Here, the OPUS translation is again very literal and understandable, although better law-specific terminology exists: The human translation in our dataset is *Oskyldig på samtliga åtalspunkter*, using *oskyldig* 'innocent' instead of the literal 'not guilty,' the more formal and therefore lexically-preferred variation *samtliga* 'all,' and the legal term *åtalspunkter*, literally 'counts of indictment.'

Example (16) shows a subtle shift in meaning from the occupational term of 'counselling':

(16) I counsel people with AIDS

a. *Jag ger råd till människor med AIDS*
'I give advice to people with AIDS'
(ST, SEM)

b. *Jag ger stöd till personer med AIDS*
'I give support to persons with AIDS'
(ST, SEM)

Both translations capture the practical aspects of the job, but leave out the professional part of being a counselor. The correct Swedish term here is *kurator* 'counselor,' but this cannot be conjugated into a verb, which means that a nominal construction must be used, e.g. *Jag är kurator åt människor med AIDS* 'I am a counselor for people with AIDS.'

5. Experiments

Prompting setups We evaluate models on our dataset in a minimal pairs setup (translationese versus alternative) to probe intrinsic preferences for idiomatic language. We use two prompting setups to examine how contextual and task framing affect their preferences. Each setup compares two alternatives: a *machine translation* and the *human alternative* version of a sentence. In the **no translation context** setup, the model is simply presented with the Swedish sentence, allowing us to assess general preference without translation instruction. In the **with translation context** setup, the model is instructed to translate the English source sentence into Swedish. We test seven variants: one where the model receives only the target sentence, and six where the prompt even includes 1–10 preceding English sentences as contextual information. The an example for the prompting setups is illustrated in Figure 2.

Models We evaluate on language models across different scales and language coverage. AI Sweden LLaMA-3 8B[1] is a continued pre-training of LLaMA-3 8B (LlamaTeam, 2024) on Scandinavian language data. To quantify the impact of this adaptation, we also include the original **LLaMA-3** 8B as a baseline. EuroLLM-1.7B and 9B (Martins et al., 2024) are open multilingual language models that include Swedish in their pre-training corpus. Finally, we include the Gemma-3 models (GemmaTeam, 2025) (270M, 1B, 4B, and 12B) as multilingual models across a wider size range. We include both base models and instruction tunes (suffix *-it*).

[1] https://huggingface.co/AI-Sweden-Models/Llama-3-8B

Model	Human>OPUS		Human>GPT-5		SGB_{all}		$\text{SGB}_{filtered}$	
	Acc.	ΔLP	Acc.	ΔLP	Acc.	ΔLP	Acc.	ΔLP
LLaMA-3-8B	47.83	1.83	40.33	0.91	38.57	-3.01	56.71	7.13
LLaMA-3-8B-it	49.83	2.01	42.50	1.17	40.33	-2.65	58.95	8.47
ai.se-LLaMA-8B	56.67	5.17	44.50	2.36	49.01	3.67	81.34	16.56
ai.se-LLaMA-8B-it	52.83	3.36	43.50	1.47	34.12	-5.80	58.95	5.60
EuroLLM-1.7B	48.33	1.09	41.17	0.83	41.46	-2.01	58.06	4.70
EuroLLM-1.7B-it	49.17	2.00	40.83	1.05	41.26	-2.34	57.41	4.88
EuroLLM-9B	49.67	2.64	43.00	1.19	42.19	-0.80	61.29	5.41
EuroLLM-9B-it	51.00	nan	40.83	nan	44.26	nan	65.35	nan
Gemma-270M	47.33	0.95	45.83	3.16	42.50	-0.13	56.33	6.22
Gemma-270M-it	49.50	0.61	44.00	2.40	40.95	-1.15	49.29	4.49
Gemma-1B	49.67	2.99	47.33	3.93	42.91	0.11	63.38	7.64
Gemma-1B-it	46.00	0.88	43.50	2.80	40.12	-2.37	51.40	2.43
Gemma-4B	52.50	4.03	46.33	3.83	43.22	-0.36	66.90	7.86
Gemma-4B-it	54.83	6.49	46.00	4.49	38.77	-2.29	56.33	3.91
Gemma-12B	55.33	5.54	45.00	4.39	43.64	0.39	71.12	8.79
Gemma-12B-it	58.33	8.31	46.33	4.09	40.53	-1.94	59.85	6.89

Table 2: Preferences for prompts *without* translation context for human alternatives over OPUS or GPT translations. We compare to the datasets by Kunz (2026) derived from Katourgi (2022) (*SGB*).

Metrics A direct comparison of log-likelihoods across sentences is problematic as the alternatives often differ in length. To account for this, we use the length-normalized mean log probability (MeanLP), defined as

$$\mathrm{MeanLP}(x) = \frac{1}{|x|} \sum_{i=1}^{|x|} \log p(w_i \mid w_{<i}) \qquad (17)$$

where $x = (w_1, \ldots, w_{|x|})$ is the sentence and $|x|$ is its length. We use it for two metrics: The **Accuracy**, i.e., the percentage of examples where the human alternative receives a higher probability than the OPUS or GPT sentence, capturing *how often* the model prefers the human variant, and **ΔLP**, i.e., the average relative difference (%) between the probabilities of the OPUS or GPT and human variants across the dataset. This metric reflects the *magnitude* of the model's preference. A negative value indicates a stronger preference for the translationese variant.

Reference dataset We compare our results to the dataset introduced by Kunz (2026), based on the book *Svenskan går bananer* (Katourgi, 2022), which contrasts Translationese sentences with idiomatic alternatives suggested by the author. Following Kunz (2026), we evaluate two setups: (1) SGB_{all}, which includes all sentence pairs from the book, and (2) $\text{SGB}_{\text{filtered}}$, a version filtered to include only pairs where (a) both sentences are of the same length, since sentence lengths vary considerably in the dataset, and (b) human annotators agreed that the alternative is clearly better than the translationese version. We use this dataset to compare whether similar patterns, such as model rankings, hold using their dataset and in ours.

6. Results

All models we evaluated show a bias toward the machine translated sentences, even for OPUS variants where the translationese wording is obvious. As shown in Tables 2, 3a, and 3b, there are very few instances where models prefer the human translation over the machine translation in most examples.

Preferences without translation context In Table 2, we see that the models often prefer the machine translated sample. The highest selection rate (accuracy) for the human alternative over OPUS is $\mathbf{58.33}$ for the largest model, Gemma-12B-it. However, even though most models prefer the machine-translated sentences in the majority of cases, the **Δ**LP values are positive in all settings except for the SGB_{all} dataset (which contains strong length-based cues favoring the translationese samples). This suggests that when a model *does* prefer the human alternative, it tends to assign it a higher likelihood than it does for machine-translated sentences when the opposite is true. Table 2 also shows that the Human>OPUS setup exhibits some expected trends: scores increase with model size for both Gemma and EuroLLM families, and ai.se-LLaMa-8B gets higher scores than Llama-3-8B. This pattern is similar to that observed for $\text{SGB}_{\boldsymbol{filtered}}$, although the latter has overall higher scores. There are only two exceptions to these trends: in the Human>OPUS setup, accuracy does not increase

Model	0 Sent.		1 Sent.		2 Sent.		3 Sent.		4 Sent.		5 Sent.		10 Sent.	
	Acc.	ΔLP	Acc.	ΔLP	Acc.	ΔLP	Acc.	ΔLP	Acc.	ΔLP	Acc.	ΔLP	Acc.	ΔLP
LLaMA-3-8B	40.50	-1.75	39.33	-1.11	40.50	-0.87	41.50	-0.68	42.50	-0.61	**42.67**	-0.51	42.17	-0.32
LLaMA-3-8B-it	39.33	-2.28	42.00	-1.03	42.83	-0.77	44.17	-0.66	44.00	-0.58	43.83	-0.53	**44.33**	-0.32
ai.se-LLaMA-8B	43.00	-1.58	49.33	-0.10	48.50	0.03	49.67	0.11	51.17	0.12	**51.33**	0.12	50.67	0.08
ai.se-LLaMA-8B-it	44.00	-1.39	44.67	-0.95	45.33	-0.68	45.50	-0.52	46.17	-0.44	46.50	-0.37	**46.83**	-0.21
EuroLLM-1.7B	33.67	-3.09	35.17	-1.69	37.33	-1.31	37.17	-1.05	37.00	-0.94	36.83	-0.83	**38.17**	-0.57
EuroLLM-1.7B-it	35.50	-2.77	38.50	-1.58	**40.50**	-1.21	39.17	-0.96	40.00	-0.85	39.83	-0.77	40.00	-0.53
EuroLLM-9B	41.33	-1.43	41.00	-0.90	42.00	-0.61	44.83	-0.40	42.83	-0.34	**45.67**	-0.26	45.50	-0.15
EuroLLM-9B-it	38.83	-1.90	42.83	-0.90	43.17	-0.64	45.50	-0.39	44.67	-0.31	44.00	-0.23	**47.00**	-0.11
Gemma-270M	35.17	-2.42	35.00	-1.82	34.50	-1.58	35.00	-1.35	**35.50**	-1.23	**35.50**	-1.11	**35.50**	-0.79
Gemma-270M-it	32.17	-3.73	32.00	-2.46	32.17	-2.15	33.33	-1.87	33.00	-1.65	34.00	-1.52	**35.17**	-1.07
Gemma-1B	36.50	-2.41	38.67	-1.19	39.83	-0.91	40.17	-0.77	39.50	-0.69	40.00	-0.62	**42.17**	-0.43
Gemma-1B-it	33.00	-3.60	35.83	-2.04	36.33	-1.47	**38.33**	-1.22	36.83	-1.06	37.83	-0.93	38.00	-0.70
Gemma-4B	40.17	-1.38	41.50	-0.90	42.50	-0.69	42.83	-0.52	43.50	-0.44	**44.67**	-0.39	**44.67**	-0.27
Gemma-4B-it	41.50	-1.53	43.00	-0.76	45.00	-0.58	46.33	-0.43	45.83	-0.32	**46.83**	-0.32	44.33	-0.24
Gemma-12B	45.50	-0.61	44.50	-0.47	46.33	-0.27	46.50	-0.19	**46.83**	-0.17	46.00	-0.14	45.50	-0.11
Gemma-12B-it	48.67	0.79	51.33	0.91	52.00	0.95	54.83	1.01	56.17	0.92	**57.00**	0.88	55.67	0.63

(a) Human > OPUS

LLaMA-3-8B	**33.17**	-2.15	33.00	-1.43	33.00	-1.25	32.50	-1.07	33.00	-0.97	32.67	-0.87	**33.17**	-0.60
LLaMA-3-8B-it	30.67	-3.12	31.67	-1.72	32.17	-1.46	32.83	-1.31	32.83	-1.21	31.67	-1.11	**34.00**	-0.74
ai.se-LLaMA-8B	**34.17**	-1.64	34.00	-0.91	33.83	-0.78	33.50	-0.68	33.83	-0.63	**34.17**	-0.57	**34.17**	-0.42
ai.se-LLaMA-8B-it	30.67	-2.63	30.67	-1.78	30.50	-1.49	31.00	-1.30	30.33	-1.17	30.00	-1.06	**31.67**	-0.71
EuroLLM-1.7B	30.17	-2.33	30.33	-1.41	31.50	-1.15	32.33	-0.96	32.83	-0.86	32.50	-0.79	**33.17**	-0.55
EuroLLM-1.7B-it	29.00	-2.48	30.33	-1.53	31.83	-1.24	32.00	-1.06	**32.83**	-0.96	31.83	-0.88	**32.83**	-0.62
EuroLLM-9B	25.83	-3.04	25.17	-1.98	26.50	-1.62	26.83	-1.37	27.17	-1.23	26.67	-1.09	**28.33**	-0.75
EuroLLM-9B-it	24.17	-3.73	25.00	-2.20	26.00	-1.82	**27.83**	-1.52	26.50	-1.34	26.50	-1.18	27.67	-0.79
Gemma-270M	34.00	-1.13	33.83	-0.92	34.83	-0.76	**36.33**	-0.63	34.33	-0.62	34.67	-0.54	35.50	-0.38
Gemma-270M-it	36.00	-1.62	33.33	-1.20	36.67	-1.02	36.83	-0.89	36.17	-0.77	36.17	-0.69	**37.33**	-0.48
Gemma-1B	33.33	-1.69	34.17	-0.95	**35.50**	-0.78	33.50	-0.67	33.17	-0.63	33.67	-0.55	33.67	-0.43
Gemma-1B-it	30.50	-2.41	31.33	-1.43	32.33	-1.16	33.00	-0.94	32.33	-0.81	33.33	-0.73	**34.83**	-0.51
Gemma-4B	31.67	-1.82	32.00	-1.25	33.17	-1.05	32.83	-0.89	33.83	-0.79	33.83	-0.71	**34.50**	-0.51
Gemma-4B-it	26.33	-3.70	28.17	-2.19	29.17	-1.84	**30.50**	-1.58	30.00	-1.43	29.17	-1.29	30.17	-0.87
Gemma-12B	31.83	-1.89	30.67	-1.27	32.17	-1.02	31.83	-0.90	32.00	-0.81	31.17	-0.74	**32.67**	-0.52
Gemma-12B-it	29.67	-3.63	28.17	-2.47	27.83	-2.09	29.17	-1.82	29.33	-1.63	30.00	-1.47	**30.50**	-0.98

(b) Human > GPT

Table 3: Preferences for machine translations vs. human alternatives for prompts *with* translation context, with 0–10 preceding sentences from the source document. Highest scores for each model are bold.

between Gemma-135M-it and Gemma-1B-it, and in the $SGB_{filtered}$ setup, the ai.se-LLaMa-8B-it does not outperform Llama-3-8B-it.[2] The Human>GPT setup produces generally lower and more variable results than Human>OPUS, similar to those for SGB_{all}, although slightly higher and with positive ΔLP values. For these two datasets, scores do not consistently increase e.g. with model size. The lower scores in the Human>GPT setup align with our analysis in Section 4, where we found that GPT translations are often more acceptable, even if many still contain issues. Consequently, comparisons against GPT are a less reliable indicator of model capability than comparisons against OPUS.

With and without translation context Adding translation context could help the model interpret how an expression is intended, but it may also bias it toward preferring the more literal machine-translated sentence. Comparing the Human>OPUS setup in Table 2 to the best values in Table 3a, we see that when *no* translation context is provided, results are substantially better: accuracy is higher than even in the best translation setups, and ΔLP values are more often positive. We thus see that the English source sentence steers the model toward the literal, translationese option, even with a substantial amount of context.

Translation contexts, with varying context length As visible in Table 3a, adding more context often leads to a clear improvement: accuracy is *always* higher with two or more preceding sentences than with none. This suggests that context helps the model better interpret the intended meaning and, as a result, favor the human translation

[2] Kunz (2026) also found that instruction tuning biases ai.se-LLaMa-8B towards translationese.

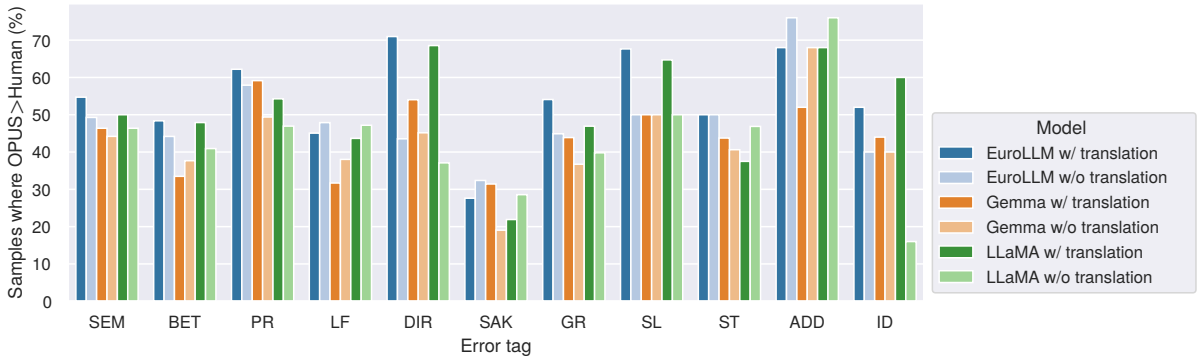

Figure 3: Percentage of samples where models prefer the OPUS-translated sentence over the human alternative, by error tag. See Section 3.2 for explanations of error tags.

in some cases. Five sentences of context is often the best overall setup, as indicated by the bolded scores in Table 3a, followed by ten sentences. This shows that a substantial amount of context generally helps better to avoid Translationese phrasings. Even so, very few models achieve positive ΔLP values (only the base AI Sweden LLaMA for half of the prompts and the Gemma-12B-it model across all prompts). We also see a strong "bigger is better" trend for Human>OPUS. For the Human>GPT values in Table 3b however, the opposite holds true: larger models generally perform worse. There is often a consistent decrease in accuracy with increasing model size, both within the EuroLLM and Gemma families: As the models become more capable, their preference for the GPT-produced translations tends to *increase*. A clear trend for Human>GPT is that the setup with 10 sentences is the best; the bolded values show that it is the best for 12 out of 16 models. There appears to be a number of pairs in the dataset where the model does prefer the human to the GPT-5 translation, but only with the help of substantial context.

Error tags We explore which types of errors are over- and underrepresented in cases where the OPUS translation is preferred over the human alternative. Specifically, we analyze the best-performing models from each family: AI Sweden LLaMA base, and the instruction-tuned versions of EuroLLM-9B and Gemma-12B. The analysis is carried out using the best-performing setups: (i) without a translation prompt and (ii) with a translation prompt including a 10-sentence context. As shown in Figure 3, direct translations (DIR) stand out as strongly overrepresented when translation context is provided (54–71% OPUS preference), a sharp increase from the no-context setting (37–45%), suggesting that context may lead models to translate too literally. Slang (SL) also leads to many errors, particularly with context (up to 68% for EuroLLM), showing that such expressions are hard to translate idiomatically. Even the minor error categories lexical preference (PR) and semantic (SEM) show high OPUS preference rates, especially when provided with translation context (up to 62% and 55%, respectively), indicating that these subtler types of errors are challenging for the models. In contrast, the major error categories missing parts (SAK) and grammatical (GR) errors show consistently low OPUS preference rates (19–32% and 37–45% without context, respectively), suggesting that clear mistakes are easier to avoid.

7. Conclusion

We introduce the first freely available, manually annotated dataset contrasting translationese from machine translations with human-written idiomatic alternatives for Swedish. In creating this dataset, we conducted a detailed analysis of translationese in English-to-Swedish translations produced by both smaller specialized machine-translation systems and LLMs. The dataset includes the English source sentence, preceding context, a fine-grained analysis of each sample, and tags for different types of translation problems. It is a resource for studying translationese in LLM outputs and ultimately for developing models that produce more natural, idiomatic translations in non-English languages.

Our experimental analyses reveal that the smaller multilingual LLMs we probe consistently exhibit a bias toward translationese phrasing; at best the human alternative is assigned a higher likelihood in a small majority of cases. Human alternatives are chosen more frequently when the English source sentence is *not* provided, suggesting that models tend to follow the source closely in translations, although even without translation context, the models often prefer the translationese variant. Including preceding context in translation prompt guides models towards the human alternative for some samples, but the overall preference toward translationese phrasing remains strong.

Limitations

Our dataset is based on a single source dataset and therefore lacks domain diversity. We chose subtitles as the domain because spoken dialogue is particularly challenging for machine translation, with many hard-to-translate expressions such as idioms and slang. Complementing it with other domains, particularly written ones, would however be valuable future work as it would give insights to what extent those are affected by translationese.

As our dataset is manually constructed and includes detailed annotations of each sample, it is relatively small. While it could be expanded using methods like back-translation, this would reduce control of the samples included in the dataset.

Choosing an appropriate set of error tags is also a trade-off. The annotators were not fully satisfied with the final tag set and sometimes found it difficult to make precise decisions. A more detailed tag set could capture finer distinctions, but it would also make annotation and interpretation more complex.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback, which helped improve this paper. We also thank our student assistant Oskar Erkstam, who worked on the dataset annotation together with author AJ. This research was supported by TrustLLM funded by Horizon Europe GA 101135671. The computations were enabled by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

8. Bibliographical References

Lars Ahrenberg. 2017. Comparing machine translation and human translation: A case study. In *Proceedings of the Workshop Human-Informed Translation and Interpreting Technology*, pages 21–28, Varna, Bulgaria. Association for Computational Linguistics, Shoumen, Bulgaria.

Lars Ahrenberg. 2021. Translation competence in machines: A study of adjectives in English-Swedish translation. In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pages 57–65, online. Association for Computational Linguistics.

Mona Baker. 1993. Corpus linguistics and translation studies: Implications and applications. In *Text and Technology*. John Benjamins.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The Belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Yuri Bizzoni, Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How human is machine translationese? Comparing human and machine translations of text and speech. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 280–290, Online. Association for Computational Linguistics.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv e-prints*, pages arXiv–2307.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? Analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2018. Examining the tip of the iceberg: A data set for idiom translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

Martin Gellerstam. 2005. Fingerprints in translation. In Gunilla Anderman and Margaret Rogers, editors, *In and Out of English: For Better, For Worse*, chapter 13, pages 201–213. Multilingual Matters, Bristol, Blue Ridge Summit.

GemmaTeam. 2025. Gemma 3 technical report.

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.

Alexander Katourgi. 2022. *Svenskan går bananer: en bok om översättningar som syns*. Lys förlag.

Delu Kong and Lieve Macken. 2025. Decoding machine translationese in English-Chinese news: LLMs vs. NMTs. In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 99–112, Geneva, Switzerland. European Association for Machine Translation.

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.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. Bactrian-X: Multilingual replicable instruction-following models with low-rank adaptation.

Yafu Li, Ronghao Zhang, Zhilin Wang, Huajian Zhang, Leyang Cui, Yongjing Yin, Tong Xiao, and Yue Zhang. 2025. Lost in literalism: How supervised training shapes translationese in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12875–12894, Vienna, Austria. Association for Computational Linguistics.

LlamaTeam. 2024. The Llama 3 herd of models.

Arle Lommel, Serge Gladkoff, Alan Melby, Sue Ellen Wright, Ingemar Strandvik, Katerina Gasova, Angelika Vaasa, Andy Benzo, Romina Marazzato Sparano, Monica Foresi, Johani Innis, Lifeng Han, and Goran Nenadic. 2024. The multirange theory of translation quality measurement: MQM scoring models and statistical quality control. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: Presentations)*, pages 75–94, Chicago, USA. Association for Machine Translation in the Americas.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. EuroLLM: Multilingual language models for Europe.

Dan Nielsen. 2023. ScandEval: A benchmark for Scandinavian natural language processing. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 185–201, Tórshavn, Faroe Islands. University of Tartu Library.

OpenAI. 2025. Introducing GPT-5. Accessed: 2025-09-01.

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.

Brian Thompson, Mehak Dhaliwal, Peter Frisch, Tobias Domhan, and Marcello Federico. 2024. A shocking amount of the web is machine translated: Insights from multi-way parallelism. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1763–1775, Bangkok, Thailand. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

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.

9. Language Resource References

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

A. A Comparison of Our Error Tags and MQM

To address the similarities and differences between the MQM Error Typology and the self-developed error tags used in this article, we compare the two using the CORE Typology definitions provided on the MQM website[3]. Readers familiar with the MQM error classification may notice similarities between MQM categories and our tags. The purpose of this section is therefore to compare the two systems and highlight how our tagging scheme captures relationships between language type, error causes, and error effects, as well as how translation errors influence the receiver's understanding.

A.1. Motivation Behind our Custom System

Based on the MQM documentation and typology design, the framework appears primarily oriented toward organizational and technical written communication, where terminology management and adherence to conventions are central. In contrast, our error tags were designed to highlight the specific translation error patterns observed in our dataset, which consists of conversational language from the OpenSubtitles corpus. When addressing issues in transcribed spoken dialogue, translation errors often relate to linguistic nuance, emotional tone, and idiomatic expression. These aspects play a central role in conversational language but are less emphasized in frameworks primarily designed with technical or professional translation contexts in mind. Although it would have been possible to extend the MQM framework with additional categories to better suit our needs, we decided that developing a custom tagging system improves the clarity of our analysis. Some of the errors we identify could be captured within the MQM framework. However, other nuances are difficult to classify without explicitly linking error causes and their effects, particularly when the analysis does not rely on MQM's scoring-based evaluation scheme. Our error tags were developed through an iterative annotation process. Sentences containing translation issues were manually examined, and as recurring patterns emerged, additional tags were introduced to capture these nuances. While MQM is designed to measure translation quality, our tagging system focuses on identifying systematic error patterns in the dataset. The tags allow us to capture relevant aspects of the linguistic context, identify the underlying causes of translation issues, and determine whether these issues result in loss of meaning, semantic shifts, or merely less preferable lexical choices.

[3] https://themqm.org/the-mqm-full-typology/

A.2. Comparison of Individual Tags

Our *Style* (ST) tag, which captures instances of domain-specific language, is similar to the MQM *Terminology* category in that both relate to the appropriate use of specialized vocabulary. Our tag covers issues corresponding to the MQM *Terminology* subcategories (*Inconsistent with terminology resource*, *Inconsistent use of terminology*, and *Wrong term*), since domain-specific communication typically requires the correct use of established terminology that can often be verified through domain resources. However, the role of the tag differs between the two systems. In MQM, *Terminology* functions as an error category and is treated as a direct cause of translation errors, typically evaluated against predefined terminology guidelines or resources. In contrast, our *Style* (ST) tag is not itself an error tag and is not tied to a specific terminology resource. Instead, it marks sentences where domain-specific language is present and where appropriate terminology or professional language conventions must be considered. The *Style* tag therefore extends beyond terminology. It is also used to identify sentences that require attention to linguistic appropriateness and preferred forms of expression within professional or institutional contexts. Errors occurring in sentences marked with *Style* may therefore vary in type and severity. These may include lexically preferred ways of expressing a concept within a field (Example 15), translations that only partially capture the intended meaning (Example 16), or cases where specialized terms are translated incorrectly, resulting in a loss of meaning (BET).

Our *lexical error* (LF) tag captures the same errors as the "mistranslation" subcategory of MQM-*Accuracy*, specifically addressing cases where certain words are translated incorrectly. In cases where words are missing, aligning with the "under-translation","omission", and "untranslated" subcategories of MQM-*Accuracy*, our *Missing* (SAK) tag is used to highlight when target language sentences lack words from the source sentence or when words remain in the source language. In cases where unnecessary words are present in the translation, our *Added words* (ADD) tag is used, aligning with the "addition" subcategory of MQM-*Accuracy*. Our use of tags in the dataset captures causes of errors, but the outcomes can be of various types. For example, "Yeah, all right" may be translated simply as "Okej" (Okay), which is missing (SAK) several words and changes the semantics (SEM) of the affirmative response. If the translation is strongly affected by missing, added, or incorrectly translated words, the most severe outcome is *loss of meaning* (BET), which is tagged accordingly. If the meaning is somewhat altered and the sentence undergoes a semantic shift, the *semantic* (SEM) tag is used.

If there is instead a more preferable way to express something, our *Lexical preference* (PR) tag is applied. Words may also be missing from or added to a sentence without making any substantial difference, and therefore without requiring additional tags, as shown in Example 13.

On a general level, the *Semantic* (SEM) tag is used to capture translations that exhibit some form of semantic shift, change in energy, or emotional nuance when compared to the source language. This may result from specific errors, but the tag can also be used independently in situations where there is no clear or severe source of the semantic shift, and the difference instead arises from the particular word choices in the translation. An example from the dataset is the response "Totally", used as a positive and agreeable reply to participating in an activity, being translated as "Helt och hållet". This expression is closer in meaning to *completely* or *entirely*, and shifts the response's tone to something more serious than in the source language. MQM does not provide a direct equivalent to *Semantic* (SEM) that explicitly captures semantic shifts in tone, energy, or emotional nuance, although the "undertranslation" and "overtranslation" subcategories (MQM-*Accuracy*) partially cover shifts that have semantic effects.

Our *Loss of meaning* (BET) tag also captures "mistranslation" (MQM-*Accuracy*), but in terms of *error effect* rather than *error cause*. The BET tag is never used by itself, but instead serves as an indicator that another error is severe enough to cause the sentence to carry a different, or indistinguishable, meaning in the target language compared to the source language.

Direct translation (DIR) captures literal word-by-word translations. This becomes relevant to tag because such translations can cause different types of errors, most often related to preferential or semantic issues, but they may also lead to grammatical problems or even loss of meaning in strongly idiomatic speech. No equivalent type of translational specification is found in MQM.

Our *Grammar* (GR) tag is used to capture grammatical and syntactic errors, aligning with the "grammar", "textual conventions", and "spelling" subcategories of MQM-*Linguistic conventions*. Severe grammatical errors can cause *Loss of meaning* (BET) or sometimes *Semantic* (SEM) shifts if the grammatical structure affects how a sentence is interpreted. An example from the dataset is the sentence "You'll toast but not drink", which is translated as "Du skålar, men inte dricker", roughly meaning *You toast/raise your glass, but don't drink*. The syntactic structure of the translation is incorrect and also contributes to making the phrase appear more commanding and somewhat unpleasant.

Our *Idiom* (ID) and *Slang* (SL) tags are used to identify sentences containing idiomatic or slang expressions. These are not direct errors themselves, but can be causes of different types of translation errors or issues. In MQM, the "Unidiomatic style" subcategory of *Style* is used for situations where the target language uses grammatically correct but unnatural language. In our dataset, however, several factors related to strongly idiomatic language use are important to capture, which requires separating the type of language used from error causes and error effects. Idioms or slang are often handled through *Direct translation* (DIR) or through *lexical errors* (LF), and can result in *Loss of meaning* (BET), as seen in Examples 8 and 9. Other examples can also be found in the dataset, such as the slang expression "Nuts!", used to express that something is crazy, which is directly translated (DIR). Because there is no equivalent expression in Swedish, this results in a loss of meaning (BET). Translating idiomatic expressions into more literal sentences may capture the general meaning, but can still cause semantic shifts (SEM). An example of this occurs in the dataset where the phrase "get well soon" is translated directly (DIR); in Swedish, however, this formulation becomes more assertive and commanding, which is not the intended tone when wishing someone a recovery.

Our *Lexical preference* (PR) tag is also somewhat related to the "Unidiomatic style" subcategory (MQM-*Style*), in the sense that a lexically preferred expression may also be a more idiomatic one. However, in our dataset it is important to distinguish between cases where established idiomatic or slang expressions are used and cases where the issue concerns lexical preference from a native speaker's perspective.