Mitigating Translationese with GPT-4: Strategies and Performance

Maria Kunilovskaya[1], **Koel Dutta Chowdhury**[1], **Heike Przybyl** [1], **Cristina España-Bonet**[2], and **Josef van Genabith**[1,2]
[1]Saarland University, Saarland Informatics Campus, Germany
[2]German Research Center for Artificial Intelligence (DFKI)
maria.kunilovskaya@uni-saarland.de

Abstract

Translations differ in systematic ways from texts originally authored in the same language. These differences, collectively known as translationese, can pose challenges in cross-lingual natural language processing: models trained or tested on translated input might struggle when presented with non-translated language. Translationese mitigation can alleviate this problem. This study investigates the generative capacities of GPT-4 to reduce translationese in human-translated texts. The task is framed as a rewriting process aimed at modified translations indistinguishable from the original text in the target language. Our focus is on prompt engineering that tests the utility of linguistic knowledge as part of the instruction for GPT-4. Through a series of prompt design experiments, we show that GPT4-generated revisions are more similar to originals in the target language when the prompts incorporate specific linguistic instructions instead of relying solely on the model's internal knowledge. Furthermore, we release the segment-aligned bidirectional German–English data built from the Europarl corpus that underpins this study.

1 Introduction

There has been a surge of interest in the impact of translationese on the performance of natural language processing (NLP) applications. Translationese has been shown to have tangible effects on the outcomes of various cross-lingual tasks, potentially leading to biased results and decreased or artificially inflated performance, especially in evaluating machine translation (MT) models (Zhang and Toral, 2019; Graham et al., 2020), but also in the natural language inference tasks when using translated datasets and cross-lingual transfer scenarios (Artetxe et al., 2020). While translationese is viewed as an inalienable property of translated language, preferences may lean toward translation variants that are closer to target language patterns provided that the meaning and usefulness of the message in the source language (SL) are retained. The task of reducing translationese by making translations less deviant from the originally authored text in the target language (TL) is a newly recognised and relevant NLP problem. At the same time, only a few studies actively address it, including Dutta Chowdhury et al. (2022) who remove translation bias in latent representation space, as well as Jalota et al. (2023) and Wein and Schneider (2024), debiasing translations at the surface text level.

Our work is the first to explore the utility of linguistically informed prompts to harness the generative capabilities of large language models (LLMs) in the task of translationese mitigation. This approach is inspired by the successful application of LLMs to a range of text adaptation tasks including simplification (Feng et al., 2023), style transfer (Suzgun et al., 2022; Reif et al., 2022), and translation (post-)editing (Chen et al., 2023; Raunak et al., 2023). To the best of our knowledge, only Chen et al. (2023) uses LLMs to address translationese reduction. We extend this line of research.

Specifically, we focus on exploring the impact of linguistic knowledge, made available to

the LLM via prompts, on the outcomes of translationese reduction. The key research question is **what type of information is required in the prompts to effectively guide the model through the rewriting process**. We propose two approaches: (i) a self-guided approach, which probes the ability of the LLM to solve the task independently using its internal knowledge versus (ii) a feature-guided approach, which relies on detailed linguistically-informed instructions to edit the input. The instructions are based on the observed deviations of each individual segment from the expected TL norm. We define the expected TL norm as the type of language that can be expected in the target culture in a comparable communicative situation. It is represented by the average feature values from the register-comparable corpus of TL documents produced by native speakers of the TL (hereinafter referred to as originals).

The contributions of this work are as follows:

- We formulate the translationese mitigation task in an LLM-prompting setup, where an LLM is expected to remove the translation process artefacts and generate a 'less translated' version for an existing human translation (HT).
- We demonstrate the importance of detailed linguistically-informed instructions in formulating prompts, individually tailored for each segment.
- We release the document- and segment-level aligned corpus created from Europarl for this study and the multiparallel datasets for English–German and German–English contrastive samples including LLM generated versions aligned with the inputs[1].

These contributions collectively address our research question and advance our understanding of the impact of linguistic knowledge available to the LLM via prompts on the outcomes of translationese reduction. The remainder of this paper is organised as follows: Section 2 discusses related work. In Section 3, we introduce our prompt generation approaches. Section 4 details our experimental settings, including the rationale behind our linguistic feature design, feature extraction and selection methods, data description and our evaluation strategy. Section 5 presents and discusses the results. We conclude with a summary in Section 6.

[1] https://github.com/SFB1102/b7-b6-prompting-eamt2024

2 Related Work

Translationese artefacts exert a substantial influence on diverse downstream tasks. In MT, Toral et al. (2018) and Edunov et al. (2020) found that source sentences that were already the result of a translation were easier to translate than non-translated sources returning higher BLEU scores. Graham et al. (2020) and Zhang and Toral (2019) also showed that translationese in test sets could lead to inflated and inaccurate evaluation scores and recommended non-translated sources in MT evaluation to avoid these biases. The influence of translationese on MT goes beyond evaluation. For example, Riley et al. (2020) trained the translationese classifier to tag the sentences in training data to control the output domain: translationese ("Tr") or original/natural text ("Nt."). In other cross-lingual applications, Singh et al. (2019) showed that substituting original training samples with their translations from another language improves performance on natural language inference tasks. Clark et al. (2020) introduced a translation-free question-answering dataset to avoid having inflated gains from translation artefacts in transfer-learning tasks. Artetxe et al. (2019) found that cross-lingual models suffered from induced translation artefacts when evaluated on translated test sets.

Active attempts to level out translationese bias include a method that can be applied in the *translate-train*[2] cross-lingual setup (Yu et al., 2022). They created a mapping from the original to the translated language, projecting original and translated text into a shared multilingual embedding space and minimising the distance between the mapped representations of the originals and translations. To mitigate translationese effects in translated data, Dutta Chowdhury et al. (2022) extended the Iterative Null Space Projection algorithm (Ravfogel et al., 2020) originally designed to mitigate gender attributes, to *debias* translationese artefacts, and not directly on the text itself, which makes them less interpretable. Wein and Schneider (2024) reduced translationese deviations at the surface level of text using Abstract Meaning Representation (AMR) proposed by Banarescu et al. (2013)) as an intermediate form to abstract away from translationese artefacts. In another line of research, Jalota et al. (2023) reframed the

[2] In this setting, the training is based on translated data instead of originally authored data.

task as a self-supervised monolingual translation-based style transfer task, aiming to make human-translated text closely resemble original texts in the TL. However, whether current out-of-the-box LLMs are able to mitigate translationese from text without removing traces of other variables remains unexplored. Apart from the key related works in translationese mitigation, we elaborate on other contemporary studies that have used LLMs for manipulating text, sometimes with goals related to refining translations or removing undesired information from text representations. Vilar et al. (2023) benchmarked the capabilities of LLMs to translate, and Kocmi and Federmann (2023) and Lu et al. (2023) to evaluate translations. Along the same line, Hendy et al. (2023) extensively analyses the translation output of LLMs to demonstrate that GPT-enabled translation achieves high quality when utilised for the translation of high-resource languages. However, it still falls short in terms of translation quality for underrepresented languages. Likewise, Raunak et al. (2023) investigated these differences in terms of the literalness of translations produced by standard NMT and ChatGPT-3.

Contemporaneously to the present work, Chen et al. (2023) propose a simple way to refine translations iteratively with LLMs based on automatic post-editing that imitates human corrections.

3 Prompt Generation

Our experiments are designed to explore the effectiveness of including various types of information in prompts that influence the generative behaviour of an LLM in the task of translationese mitigation. The study is based on a bidirectional German-English subset of Europarl data. Each translation direction is aligned at the segment level, meaning that depending on the syntactic arrangement of the same content the source or the target side of the parallel data can have more than one sentence. We experimented with two prompting approaches: self-guided and feature-guided, each with two modes (min and detailed). The full prompt examples for each of these four prompting setups appear in Appendix C. The four setups vary in the degree of independence in decision-making given to the model and in the level of linguistic instruction. Below we provide a description for each setup.

1. **Self-guided modes**: These modes rely on the model's discretion in solving the task.

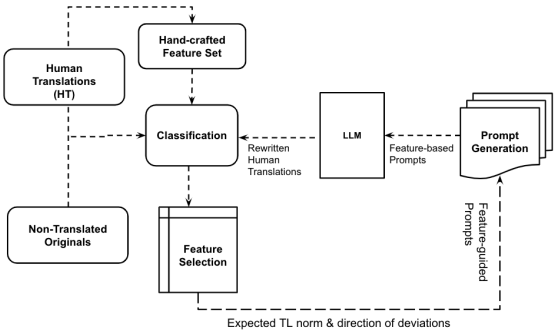

Figure 1: An overview of our pipeline based on the feature-guided approach.

min: In this mode, the prompt formulates the translationese reduction task without any reference to the concept of translationese or any other linguistic knowledge, in layman's language: *Your task is to re-write a human translation in a more natural way if necessary*. Importantly, the model is given the option to return the input if it does not detect any traces of translationese, i.e. if the translation already sounds *like a text originally produced in the target language*.

detailed: Unlike the previous setup, the prompt contains a concise paragraph (186 words) explaining the concept of translationese as discussed in translation studies (Volansky et al., 2015; Hu and Kübler, 2021). It describes the known trends in translator behaviour and typical translationese indicators established in the literature. The option to return the input translation in case the model could not detect translationese deviation is kept. Figure 1 shows an overview of our pipeline for the feature-guided approach.

2. **Feature-guided modes**: The prompts include specific linguistic instructions that limit the model to a set of required transformations for each input translation. The list of instructions is tailored for each segment and addresses the most prominent deviations of this segment from the expected TL norm based on a number of linguistically motivated handcrafted features (Section 4.1). The TL norm for each feature is calculated as the average across all segments in the original text category. The instructions for a particular feature are included in the prompt if the feature met the following criteria: (i) deviated more than

2.5 times from the TL norm in the direction observed in HT (e.g. in German translations, the frequency of additive connectives was lower than in non-translations, while translations into English had significantly more additive connectives than comparable non-translations in English), (ii) it was among the top 15 translationese indicators as flagged by SVM feature weights for each translation direction. If none of the features exceeded the 2.5-ratio threshold, the segment was not sent to the model and remained unchanged. The instructions for all segments were pre-compiled based on the threshold calculations and formatted as a newline-separated list appended to the task statement, source segment, and target segment (i.e. HT). The two variations of this setup were only different in how detailed the description of each instruction was.

min: The model was given a task *to re-write a human translation in a more natural way* by following the pre-compiled instructions. The instructions were formulated in a very concise manner. For example, *Make causative-consecutive relations between parts of the sentence more explicit*.

detailed: The task and the instructions were explained in more detail, offering descriptions of the linguistic concepts. Where possible, we provided lists of TL-specific examples for linguistic categories. Those prompts started with a brief definition of translationese followed by instructions like *Make causative relations between parts of the sentence more explicit. This can be done by using connectives like: because, therefore, so that, for this reason, as a result, after all, for that reason, hence, consequently, to this end*. In formulating the descriptions we relied on the definitions from the UD framework.[3]

In summary, in the two self-guided modes, the LLM's behaviour is not constrained by specific rewriting instructions. The model had to make self-guided decisions not only on how to rewrite a segment but also on whether any transformation was necessary at all. In contrast, the two feature-guided modes closely supervised the model by specifying linguistic properties to be transformed in the rewriting process. All prompt types contained the source segment and its human translation. Preliminary experiments indicated that when the model was not constrained by the source segment, the re-writing process was highly volatile. Throughout this study, we only considered segments longer than eight words.

LLM Specifications. For our experiments, we use the GPT-4 model through the OpenAI API.[4] This model returned more consistent results than GPT-3.5-turbo in a preliminary study. Our best results are obtained with GPT-4 and the default temperature (0.7). Although we attempted to suppress noise[5] in the GPT-4's output by appending formatting instructions to each prompt (e.g. *Do not add any meta-phrases or quotation marks*), the rewritten versions required extensive cleaning. The model's comments were varied and the output had to be manually curated. Interestingly, even though the instructions were provided in English, the model added meta-comments either in German or in English when working on re-writing translations into German.

4 Experimental Settings

4.1 Linguistic Features

We propose to capture translationese with a set of morpho-syntactic features and text measures extracted from the Universal Dependencies (UD) annotation of the data. Unlike surface features like ngrams and neural network-based feature-learning approaches to translation detection, explicit discrete structural features have a lower risk of capturing irrelevant topical differences between the categories (Volansky et al., 2015; Borah et al., 2023). They are more interpretable and can be incorporated into human-readable rewriting instructions for an LLM. The initial feature set included 58 features and was motivated by previous research in language-pair-specific translationese (Evert and Neumann, 2017; Kunilovskaya and Lapshinova-Koltunski, 2020) and contrastive studies (Konig and Gast, 2007), as well as multilingual analysis (Hu and Kübler, 2021). In Appendix A, the fea-

[3] https://universaldependencies.org/u/dep/index.html

[4] https://platform.openai.com/docs/guides/gpt. The final version of the re-written translations analysed here was obtained between 08 and 10 March 2024.

[5] refers to undesirable outputs in model-generated text, including unwanted copies of the input, additional quotes and meta-comments from the model like: 'Here is the revised translation:'

tures are categorised according to the type of linguistic units they capture. Our feature set contains grammatical forms, morphological word classes, clause types, syntactic dependencies, word order patterns, discourse elements, and textual measures. Generally, we gave preference to the features that:

- captured relatively frequent linguistic items to minimise sparsity as much as possible, especially at the segment level,
- were suggested as contrastive for the given language pair and/or were expected (or known) to generate translationese deviations from the TL norm.

Feature Extraction. For most features (37 out of 58), the extraction was straightforward and directly dependent on the accuracy of automatic annotation. The annotation quality is comparable across our languages, according to the official report for the models[6] used. Six features of the remaining 21 features (various discourse marker types and adverbial quantifiers) relied on external pre-defined lists which were compiled using previous research in language variation for each language (Biber, 1988; Nini, 2015; Evert and Neumann, 2017), while the other 15 features included (i) straightforward metrics such as sentence length in tokens, word length, number of simple sentences, number of clauses per sentence, the ratio of core verbal arguments expressed by nouns, (ii) mean hierarchical distance and mean dependency distance (Jing and Liu, 2015), (iii) type-to-token ratio calculated as the ratio of part-of-speech-disambiguated content word types to their tokens, lexical density calculated as the ratio of disambiguated content word types to all tokens, (iv) and six word-order patterns that were discussed as English-German contrasts (Konig and Gast, 2007). All features were estimated and normalised at the sentence level and mean-aggregated for segments or documents. The highly correlated features were excluded (cutoff=0.65 for both languages).

Feature Evaluation and Importance. Table 2 shows that the proposed feature set demonstrated relatively high classification results at the document level. The feature selection did not yield considerable gains in performance: the improvements on the optimal 29 and 45 features (reported in Table 2) were in the fractional part of the scores. This suggests that the proposed feature set does not include irrelevant features and is effective in capturing translationese. None of the features could reliably distinguish the categories on its own, demonstrating that translationese is a subtle phenomenon, which is better captured through feature patterns, in a multi-variate setup.

4.2 Data

We use the Europarl-UdS preprocessing pipeline[7] to extract parliamentary speeches[8] delivered in German and English by native speakers and their translations into English and German respectively. Our rewriting approach required parallel data, therefore, we report the details on sentence alignment quality. The documents were automatically aligned with LF Aligner[9], a wrapper over the *hunalign* library (Varga et al., 2005), using domain-specific bilingual glossaries built from IATE dictionaries.[10] The resulting parallel corpus was limited to the documents with an average document-level similarity score returned by the alignment tool over 0.3 and 0.5 for German-to-English and English-to-German directions, respectively. The manual evaluation of the automatic alignment, performed by a compensated research assistant on 80 document pairs (750 sentence pairs) randomly extracted for each direction, revealed that the resulting parallel corpus contained at most 4.5% (German-to-English) and 1.8% (English-to-German) of misaligned segments.

For this study, the corpus was balanced across translation directions by taking 1500 random document pairs that contained at least 450 tokens in the source language. The document length filter excluded short documents containing formulaic exchanges between the Chair and the participants of the debates in the European Parliament. All textual data were automatically parsed with the default Stanza packages for German and English (Qi et al., 2020). The quantitative parameters of the research data are given in Table 1.[11]

[6] https://stanfordnlp.github.io/stanza/performance.html

[7] https://github.com/chozelinek/europarl

[8] It is well known that translation direction and register are the two major factors that influence the properties of translations (Redelinghuys, 2016; Evert and Neumann, 2017; Kunilovskaya and Lapshinova-Koltunski, 2020; Kunilovskaya and Pastor, 2021). Europarl data is convenient because it helps control for these factors.

[9] https://sourceforge.net/projects/aligner/

[10] https://iate.europa.eu/search/standard

[11] The datasets are available as an indexed long table here: https://zenodo.org/records/11127626

		docs	segs	tokens
DE	original	1500	38,305	967,385
	translated	1500	36,078	924,919
EN	original	1500	36,078	927,045
	translated	1500	38,305	1,060,295

Table 1: Parameters of the entire research corpus (after filtering and annotation). EN (English) and DE (German) stand for the language of the comparable samples of originally authored and translated text. All translations are from the other language in the language pair. For example, DE translated are translations into German from English. DE original are texts in German by German native speakers.

The corpus in Table 1 was further distilled to obtain a contrastive sample of 200 documents in each TL that concentrated the translationese-related phenomena. To this end, we ran a 10-fold binary document-level translationese classifier using the features described in Section 4.1 and classification setup from Section 4.3. The results of this classification can be found in Table 2. For comparison, we report results for the full feature set and the optimal set of features (see details on feature selection in Section 4.3).

	feats	docs	F1
DE	29	3000	**88.83**±1.99
	58		88.39±2.54
EN	45	3000	**80.05**±1.68
	58		79.66±2.05

Table 2: The quality of the document-level translationese classifications across the two languages in the 10-fold cross-validation setup. The average document length in the translated text categories is around 700 tokens, 25.5 segments.

The contrastive subset was defined as 100 'most translated' and 100 'most original' documents based on the probability over 0.99 of belonging to their true class returned by the classifier on the best-performing 29 and 45 features for German and English, respectively. This data filtering step was required to meaningfully downsize the data to a subset manageable in the prompting experiments. Given the relatively high quality of the translationese classification (F1 score of 88% for German and 80% for English in Table 2), we have good reasons to believe that the selected documents bring into focus the contrasts between translations and non-translations while being naturally-occurring texts containing cohesive sequences of sentences. The parameters of this experimental subset appear in Table 3.

		segs	tokens	seg_len ± std
DE	original	1908	59,942	31.4±17.6
	translated	1934	57,492	29.7±14.1
EN	original	1987	55,128	27.7±13.0
	translated	1919	65,065	33.9±19.6

Table 3: Parameters of the contrastive subset for rewriting experiments. Note that the originals here are not the sources for the translations in the other language. Instead, they are the top documents predicted as originals by the classifier (Table 2).

4.3 Evaluation

Translationese Classification. Our main translationese mitigation evaluation method is segment-level[12] text classification. If a rewriting strategy is effective, the accuracy scores for classifying translationese on the rewritten output should be lower compared to classification on HT (human-translated) text. In other words, there should be a negative difference in accuracy scores between the rewritten output and the initial HT, indicating that the rewritten versions blended better with the TL norm than the existing HT. For all experiments, we used a simple Support Vector Machine (SVM) with a linear kernel (C=1) in a 10-fold cross-validation setup. Linear SVM was preferred because it allows access to feature weights. The feature weights were used to identify a set of 15 most informative features. These features were used in prompt engineering and for evaluation purposes. The feature selection was performed using Recursive Feature Elimination technique with a linear SVM as implemented in the scikit-learn library.[13] All classification results are reported for the top 15 features and for the full feature set. Although the number of instances per category was almost the same, we report a macro F1 score throughout to avoid any impact of the data imbalance on the results.

Re-translation (RT). As a sanity check for the rewriting approaches outlined in Section 3, we ran a re-translation mode (referred to as RT) to ensure that in the rewriting setups, the model follows our instructions and edits the existing translation, rather than returning a new translation. Here, we prompt the model to re-translate an existing HT if it detects any translationese deviations.

[12]Rewriting experiments on documents resulted in cropped GPT-4 output and therefore segment level was preferred.

[13]`https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html`

Statistical Analysis. The analysis of the classifiers' performance was supported by tracking the shifts in the feature values observed in the generated text against original texts and HTs. This helped us understand whether the model managed to level out the existing translationese deviations and whether it introduced new tendencies. The significance of differences between originals in the TL and rewritings was estimated using the two-tailed Mann-Whitney U Test for independent samples. The results are considered significant at the confidence level of 5%.

Content Preservation. We evaluate the quality of the GPT-4 outputs in preserving the meaning of the input translations using COMET (Rei et al., 2022). We use two variants of COMET for this purpose: (a) `R`, reference-based (`wmt22-comet-da`) and (b) `QE`, the reference-free COMETQE (`wmt20-comet-qe-da`).

Manual Analysis. The automatically edited translations and re-translations were evaluated by one of the authors of this paper, a German-native professional translator with English and German as their working languages. The evaluator reviewed a random sample of 25 generated rewritten segments for each mode and translation direction. These segments were presented in the context of the source segment and the professional HT. Their task was to assess translation faithfulness to the source (accuracy) and lexicogrammatical acceptability (fluency) using a 1-6 scale (higher is better) for each output mode. Additionally, they checked whether the rewritten translations were compliant with the provided instructions (feature-guided modes only) to see whether the model followed the instructions. The expert was not asked to pass judgments about the translationese properties of the items in their sample. We maintain that translationese is a property of language that is visible to a machine rather than a human.

5 Results and Discussion

Translationese Classification. The results of our baseline SVM segment-level classification between originals and HTs from the contrastive sample (see Section 4.2) in each TL are reported in Table 4. We report F1 scores on the top 15 features and on the full feature set to throw the performance on the top 15 features into perspective.

The main observations from Table 4 are: (i) HTs into German contain more machine-detectable deviations from non-translations than translations into English, (ii) the reduced 15-feature set returns results comparable to the full 58-feature set, especially in German. We address these strong translationese predictors in the GPT4-based rewriting pipeline.

	feats	segs	F1
DE	15	3842	81.06±0.76
	58		81.51±1.79
EN	15	3906	75.60±1.87
	58		78.30±1.42

Table 4: Segment-level classification results on human translations from the contrastive 200-document sample using linear SVM. EN and DE stand for the target language.

To assess the impact of rewriting on translated segments from the contrastive sample, we conduct another set of translationese classifications using the same original texts and their GPT4-rewritten versions on the top-15 subsets of translationese indicators addressed in the rewriting process and on the full-58 feature set. Table 12 shows the differences in F1 scores. Below we show some observations from these results. Recall that lower translationese classification accuracy would suggest that rewritten segments became less distinguishable from originals after editing. The negative differences in Table 12 confirm that GPT-4 can be conditioned through prompting to address the task, even if the overall gains are small on the segments from the contrastive 200-documents sample. The rewriting task is more successful in English than in German. All attempted approaches decrease the prominence of translationese in the English translations by at least 0.58 points. In particular, when given detailed instructions based on the linguistic features (*Feature-guided Detail* mode), we observe a substantial 7.63 and 4.07 percentage

		Rewriting Setups				
		RT	Self-guided		Feature-guided	
		–	Min	Detail	Min	Detail
DE	15	-0.28	-0.27	-1.01	**-2.39**	-2.21
	58	0.10	0.53	**-0.56**	0.06	-0.28
EN	15	-3.32	-2.70	-4.10	-3.18	**-7.63**
	58	-0.58	-1.40	-1.61	-1.61	**-4.07**

Table 5: Differences in F1 scores between the segment-level results on the rewritings and on human translations from the contrastive sample (Table 4). The best results for each feature set are shown in bold.

point decrease in classification results for the top-15 subset and for the full-58 feature set, respectively, working with the segments from the contrastive sample of 100 originals and 100 translations.

For German, the best-performing modes are the *Feature-guided Min* for 15 features and the *Self-guided Detail* setup for 58 features. Table 12 shows that the results are better for the 15 strong translationese predictors, specific for each language, even for GPT4 rewriting modes that did not rely on features. It means that the model effectively picked and reduced the most prominent translationese deviations even when it was not prompted to do so. The modes with the linguistic explanation (Detail) seem to be better than *Min* mode regardless of whether the model was presented with a list of specific rewriting instructions or was left to decide how to tackle this text adaptation task (except the *Feature-guided* approach for German on 15 features). Feature-guided modes were on average more successful than self-guided modes, especially for English. The performance on the features that were addressed in the instructions shows that the instructions were carried out in the rewriting.

Finally, the comparison with the SVM classification outcomes for the re-translation task indicate that the model did not simply return a new translation of the source. Although the model reduced translationese in the re-translation task, the explicit editing tasks performed better (cf. RT column to Detail columns in Table 12). Overall, the properties of rewritten documents are shifted towards being more similar to original texts. This effect is visible in Figure 2 which displays Kernel Density Estimation (KDE) plots for the values on the 'translationese' component obtained through Principal Component Analysis (PCA) of the 58-dimensional feature space. These plots capture the distribution of the values on this PCA component for original, human-translated and LLM-rewritten segments. Figure 2 shows that the rewritten documents (in red) are shifted from the area taken by the translated texts on the right (green line) to the non-translations' left side of the graph.

Statistical Analyses. First, we find an imbalance in the segments affected by the diverse rewriting approaches across TLs. Recall that in the self-guided and re-translation modes the model was given the option to return the input unmodified while in the feature-guided modes, the segments that did not exhibit deviations above a 2.5-ratio threshold were not sent to the model. In self-guided modes and re-translation, the model was more willing to dismiss segments as requiring no editing in German than in English. Moreover, for German the number of automatically bypassed segments was close to the number of segments that were skipped in the feature-guided modes, while for English there was a strong contrast in this respect. The translationese filter used in the feature-guided prompt generation considered about 29.28% of HTs into English sufficiently complying with the TL norm, while only less than 1% were not changed in self-guided setups.[14] This means that GPT-4 was more ready to edit a text in English than in German. Note that unchanged segments were included in the data underlying classification results in Table 12 to maintain comparability with the baseline.

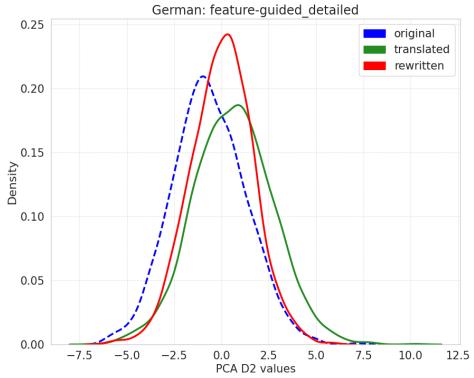

Figure 2: KDE plot for values on the 'translationese' dimension of a PCA-transformed data for German translations rewritten in feature-guided detailed mode (on 58 features).

Second, we looked into the changes in the feature frequencies in the rewritten segments against the TL non-translations and grouped the features according to their contribution to the task. The *expected* outcome is a reduction of significant deviations from the TL norm. Other possible developments include *no change* compared to the input and some *new trends* absent in HTs. Table 6 has the number of features in each group counted from the full results of statistical tests given in Appendix B.

Table 6 shows that the feature-guided modes had different effectiveness across the translation direc-

[14]The full account of these differences can be found in Appendix B, Table 7.

	shift	Feature-guided Min	Feature-guided Detail
DE	expected	3 (0)	6 (3)
	new trend	19 (7)	16 (4)
	no change	36 (8)	36 (8)
EN	expected	15 (2)	16 (3)
	new trend	29 (9)	26 (9)
	no change	14 (4)	16 (3)

Table 6: Analysis of changes in feature frequencies and significance of differences: Number of features by the direction of frequency change after rewriting in feature-guided modes. The number in brackets shows how many of them were among the features addressed in the instructions.

tions. In German, the expected changes were observed only for a few features (3 and 6 for Min and Detail modes), while most features remained unaffected (36 for both Min and Detail modes). In English, most features (29 and 26 for Min and Detail modes) demonstrated new deviations from the TL norm. Two-thirds of these emerging trends were over-normalising tendencies, i.e. the features started to deviate from the TL norm in the direction opposite what is typically observed in translations. This effect can hardly be linked to the number of times each feature appeared in the instructions. We hypothesise that the unexpected outcomes were collateral to the other requested transformations which counteracted the specific instructions to favour or avoid specific structures. Except for over-normalisation, the rewritten versions occasionally exhibited deviations on the features where there were no statistical differences between HTs and non-translations.

In almost all cases the non-significant lack or overuse of a specific item was intensified by rewriting. For example, in feature-guided detailed mode on German the number of clauses per sentence (numcls) and specifically of clausal complement without own subjects (xcomp) went further down as compared to HT. In English, the lower frequency of coordinated elements (conj) and higher frequency of simple sentences (simple) reached levels of statistical significance. These deviations, however, were not large and/or consistent enough to build new patterned distinctions between GPT4-edited translations and the TL norm, at least not along the same translationese properties. The rewriting pipeline effectively removed the targeted translationese signals without introducing new deviations, at least those captured by our features. It should be noted that there seems to be a certain limit to the effective number of instructions that could be passed to the rewriting pipeline. In the reported feature-guided setups, the number of instructions per segment was at most 7 for German and 9 for English, with averages about 2.4 and 2.3, respectively. An attempted alternative approach that generated more instructions per segment was less successful. That approach considered all features with the statistical differences between originals and translations (about 43-44 out of 58 features) if their frequencies for a given translated segment were two standard deviations away from the expected TL norm in the 'translationese' direction. This approach generated more varied and longer lists of instructions: the average number of instructions per segment was 3.4, and the number of features addressed in the instruction was twice higher than in the reported approach (21 and 30 for German and English).

Content Preservation. Even if the rewriting pipeline seems to achieve the goals of translationese reduction, we need to make sure that it outputs acceptable translation variants.

		Rewriting Setups				
		RT	Self-guided		Feature-guided	
		–	Min	Detail	Min	Detail
DE	R	0.63	0.87	0.86	0.84	0.85
	QE	0.16	0.49	0.48	0.40	0.44
EN	R	0.85	0.85	0.84	0.80	0.82
	QE	0.46	0.46	0.45	0.33	0.39

Table 7: Average COMET scores for the generated sentences from each of our four rewriting techniques for translationese reduction, compared against the original sentences as references.

Table 7 shows that for German the rewriting setups consistently outperform GPT4-translated sentences in terms of COMET scores for both reference-based (R) and reference-free (QE) evaluations. Specifically, for reference-based (R) evaluation, the COMET scores range from 0.84 to 0.87 across different rewriting setups, indicating a high level of content preservation. This suggests that the rewriting techniques effectively maintain the meaning of the original English sentences. The results for the English pipeline evaluation indicate that (i) GPT-4 is probably much more skilled in translating into English than into German, and that (ii) the rewriting setups, especially in the feature-guided modes, generate less semantically similar translation candidates, even if they seem to be less

deviating from the TL norm on some frequency-based features.

Manual Analysis. The manual analysis by a translation expert was carried out to assess the quality of the re-written output in addition to automatic COMET scores. The human evaluation (Table 13) returned consistently high scores for both accuracy and fluency, giving better results in the German-to-English direction than English-to-German.

		Rewriting Setups				
		RT	Self-guided		Features	
		–	Min	Detail	Min	Detail
DE	A	5.9	5.8	5.8	5.1	5.4
	F	5.4	5.7	5.6	5.4	5.4
EN	A	5.9	5.7	5.9	5.2	5.4
	F	6	6	5.9	5.6	5.8

Table 8: Results of human evaluation for accuracy (A) and fluency (F) in a 1-6 Likert scale.

Both self-guided modes were rated higher than the feature-guided modes. This is in line with the automatic results on content preservation (cf. Table 7). Although the feature-guided instructions were generally followed (92-96% of observations in DE, 96% of observations in EN), it was noticed that they were applied excessively leading to overtransformed renditions as in Example 2 (Appendix D). Human and machine translation preserved one long sentence, showing traces of translationese. The GPT4-rewritten output in the self-guided modes returned 2-3 short sentences whereas the instruction to *make the sentences shorter* resulted in 4 and even 5 shorter sentences for the same input. A similar tendency can be observed in Example 1, where the instruction to *use more adverbial modifiers* in rewriting translations into German in the feature-guided modes resulted in the overuse of adverbials (underlined in the example) and also intensification of the message and therefore decline in accuracy.

6 Conclusion

In this paper, we explore the potential of using LLM-prompts to reduce translationese-related differences between translated and non-translated texts. We evaluate four types of prompts based on either a high-level explanation of the translationese mitigation task or on a micro-managing approach to prompting where the model received segment-tailored instructions to increase or reduce the frequency of prominent translationese predictors. Our findings demonstrate that GPT-4 **was able to edit human translations to make them less distinguishable** in an automatic classification setup from non-translations in both self-guided and feature-guided LLM-rewriting modes. The best results were seen for English on the prompts containing feature-guided instructions with a linguistic description of special terminology, showing that **the prompting approach benefited from including linguistic knowledge**.

For German the results were less straightforward but the advantages of detailed task information and specific linguistic instructions were visible. The inferior results on the re-translation task provide further evidence in favour of linguistic features for the translationese reduction task. In our experiments, prompting was more effective in the German-to-English translation direction even though the difference between translated and non-translated documents in German was more detectable to start with (as indicated by 5% higher classification results). We can tentatively explain this result by **the language of instruction (English)**, which might prime the model for better performance when generating English output. Future work may need to extend this research by including tasks with instructions in German, especially when the model rewrites translations into German.

Finally, we have seen that even though rewritten translations exhibited some new individual deviations from non-translations on some individual features, they did not coalesce into patterns picked be a classifier. This conclusion is supported by high results from content preservation metrics and from the manual analysis for accuracy and fluency of translations. While our translationese classification-based evaluation shows that LLM-rewriting is effective, in our paper we focus on the tip of the iceberg, i.e. the segments from 200 most contrastive documents in our data set. Furthermore, manual evaluation and, to some extent automatic evaluation, show that content preservation under LLM-rewriting needs more attention, and we will focus on this in our future research.

7 Acknowledgments

This research was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1102 Information Density and

Linguistic Encoding, Project-ID 232722074.

References

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy, jul. Association for Computational Linguistics.

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online, November. Association for Computational Linguistics.

Banarescu, Laura, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Biber, Douglas. 1988. *Variations Across Speech and Writing*. Cambridge University Press.

Borah, Angana, Daria Pylypenko, Cristina Espana-Bonet, and Josef van Genabith. 2023. Measuring spurious correlation in classification: 'Clever Hans' in translationese.

Chen, Pinzhen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2023. Iterative translation refinement with large language models. *arXiv preprint arXiv:2306.03856*.

Clark, Jonathan H, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: a benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Dutta Chowdhury, Koel, Rricha Jalota, Cristina España-Bonet, and Josef Genabith. 2022. Towards debiasing translation artifacts. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3983–3991, Seattle, United States, jul. Association for Computational Linguistics.

Edunov, Sergey, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2836–2846, Online, July. Association for Computational Linguistics.

Evert, Stefan and Stella Neumann. 2017. The impact of translation direction on characteristics of translated texts: a multivariate analysis for English and German. *Empirical Translation Studies: New Methodological and Theoretical Traditions*, 300:47–80.

Feng, Yutao, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2023. Sentence simplification via large language models. *arXiv preprint arXiv:2302.11957*.

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

Hendy, Amr, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? A comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Hu, Hai and Sandra Kübler. 2021. Investigating translated Chinese and its variants using machine learning. *Natural Language Engineering*, 27(3):339–372.

Jalota, Rricha, Koel Chowdhury, Cristina España-Bonet, and Josef van Genabith. 2023. Translating away translationese without parallel data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7086–7100.

Jing, Yingqi and Haitao Liu. 2015. Mean hierarchical distance augmenting mean dependency distance. In Nivre, Joakim and Eva Hajicova, editors, *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 161–170, 24–26 August.

Kocmi, Tom and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *24th Annual Conference of the European Association for Machine Translation*, page 193.

Konig, Ekkehard and Volker Gast. 2007. *Understanding English-German Contrasts*. Erich Schmidt Verlag.

Kunilovskaya, Maria and Ekaterina Lapshinova-Koltunski. 2020. Lexicogrammatic translationese across two targets and competence levels. In Calzolari, Nicoletta, Frederic Bechet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, and And Others, editors, *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4102–4112. The European Language Resources Association (ELRA).

Kunilovskaya, Maria and Gloria Corpas Pastor. 2021. Translationese and register variation in English-to-Russian professional translation. In Wang, Vincent X., Defeng Li, and Lily Lim, editors, *New Frontiers in Translation Studies*, pages 133–180. Springer Nature Singapore Pte Ltd.

Lu, Qingyu, Baopu Qiu, Liang Ding, Liping Xie, and Dacheng Tao. 2023. Error analysis prompting enables human-like translation evaluation in large language models: a case study on ChatGPT. *arXiv preprint arXiv:2303.13809*.

Nini, Andrea. 2015. Multidimensional Analysis Tagger (v. 1.3).

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: a Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Raunak, Vikas, Arul Menezes, Matt Post, and Hany Hassan. 2023. Do GPTs produce less literal translations? In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Ravfogel, Shauli, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256.

Redelinghuys, Karien. 2016. Levelling-out and register variation in the translations of experienced and inexperienced translators: a corpus-based study. *Stellenbosch Papers in Linguistics*, 45(0):189–220.

Rei, Ricardo, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Reif, Emily, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 837–848.

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

Singh, Jasdeep, Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2019. XLDA: cross-lingual data augmentation for natural language inference and question answering. *arXiv preprint arXiv:1905.11471*.

Suzgun, Mirac, Luke Melas-Kyriazi, and Dan Jurafsky. 2022. Prompt-and-Rerank: a method for zero-shot and few-shot arbitrary textual style transfer with small language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2195–2222.

Toral, Antonio, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium, October. Association for Computational Linguistics.

Varga, Dániel, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2005. Parallel corpora for medium density languages. *International Conference Recent Advances in Natural Language Processing, RANLP*, pages 590–596.

Vilar, David, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: assessing strategies and performance. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Volansky, Vered, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Wein, Shira and Nathan Schneider. 2024. Lost in translationese? Reducing translation effect using abstract meaning representation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Yu, Sicheng, Qianru Sun, Hao Zhang, and Jing Jiang. 2022. Translate-train embracing translationese artifacts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 362–370, Dublin, Ireland, May. Association for Computational Linguistics.

Zhang, Mike and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy, aug. Association for Computational Linguistics.

Appendix A. Linguistic features

Table 7: Types of linguistic information by language level captured with the UD features. The 15 features identified as strong translationese predictors at sentence level for German as a target language appear in bold, for English – in *italics*.

	type	number	list of features [shorthand code]
1	word forms	5	**finite verb [fin]**, ***past tense, including conjunctive forms [pastv]***, infinitive [inf], passive voice form [aux:pass], deverbal noun [deverb]
2	word classes	9	noun [nn], *personal [ppron]*, **possessive [poss]**, reflexive [self] and demonstrative [demdet] pronouns, adverbial quantifier [advqua], coordinate and subordinate conjunctions ([cconj], [sconj]), adposition [prep]
3	discourse markers	5	*adversative [advers]*, ***additive [addit]***, **causative-consecutive [caus]**, temporal-sequential [tempseq] connectives and epistemic stance markers [epist]
4	types of clauses	7	clause with modal predicates [mpred], adjectival clause, including relative clauses [acl], ***adverbial clause [advcl]***, clausal complement with or without own subjects ([ccomp], [xcomp], respectively), **asyndetically joined elements in a sentence [paratax]**, *negative clause [negs]*
5	other dependencies	17	adjective in attributive function [amod], ***adverbial modifier [advmod]***, auxiliary verb [aux], appositional modifier [appos], *conjunctive relation [conj]*, copula verb [cop], three types of relations within multi-word expressions (*[compound]*, [fixed], [flat]), discourse element [discourse], subordinate clause marker [mark], nominal subject [nsubj], direct object [obj], **indirect object [iobj]**, *non-core argument [obl]*, numeric modifier [nummod], ***nominal dependent of a noun [nmod]***
6	sentence complexity and word order	10	**mean hierarchical distance [mdd] and *mean dependency distance [mhd]***, *number of clauses per sentence [numcls]*, ***ratio of nouns or proper names as core verb arguments to the total of these arguments [nnargs]***, ratio of head-verb preceding noun-object to all objects in a clause [vo_noun], inversion in main clause (in affirmative sentences) [vs_noun], ratio of oblique object preceding direct object to clauses with both dependencies [obl_obj], adverbial modifier preceding head-verb to all adverbial modifiers in a clause [adv_verb], any dependencies except subject preceding the main verb [vorfield], prepositional phrases at the end of the finite clauses [nachfield]
7	textual properties	5	**lexical type-to-token ratio [ttr]** and lexical density [dens] (based on disambiguated content types), number of simple sentences [simple], ***sentence length [sent_len]*** and word length [wdlen]
	TOTAL	58	

Appendix B. Changes in feature frequencies and feature importance

Table 7: The expected TL thresholds (i.e. the average feature values in TL originals) and the significance of differences between originals, on the one hand, and HT/rewritten outputs for each feature, on the other hand. The upward and downward departures from the expected TL norm are shown by arrows. The asterisks indicate a lack of statistical significance for the difference based on the two-tailed Mann-Whitney test for unpaired samples. The 15 features identified as strong translationese predictors at sentence level for German as a target language appear in bold, for English – in highlighted rows.

			Rewriting Setups							Rewriting Setups				
			RT	Self		Feature				RT	Self		Feature	
	TL	HT	–	Min	Det	Min	Det	TL	HT	–	Min	Det	Min	Det
	English-to-**German**							German-to-**English**						
addit	0.02	↓	↓	↓	↓	↓	↓	0.002	↑	↑	↑	↑	↑*	↓
advcl	0.312	↑	↑	↑	↑	↓	↓	0.552	↑*	↓	↓	↓	↓	↓
advmod	3.327	↓	↓	↓	↓	↓	↓	1.112	↑	↑	↑	↑	↓	↓
caus	0.012	↓	↓	↓	↓	↓	↓	0.002	↑	↑	↑	↑	↑*	↑*
fin	2.673	↓	↓	↓	↓	↓	↓	2.289	↑	↓	↓	↓	↓	↓
iobj	0.153	↑	↑	↑	↑	↓	↓*	0.01	↑	↓*	↓*	↑*	↓*	↑*
mdd	3.512	↑	↓	↓	↓	↓	↓	2.668	↑	↓*	↑*	↓	↓	↓
sent_len	29.222	↓*	↓	↓	↓	↓	↓	27.503	↑	↓	↓	↓	↓	↓
mhd	3.552	↑	↑	↑	↓*	↓	↓	3.857	↓*	↓	↓	↓	↓	↓
nmod	1.257	↑	↑	↑	↑	↓	↓*	1.562	↓	↓	↓	↓	↓	↓
nnargs	0.378	↑	↑	↑	↑	↑	↑	0.584	↓	↓	↓	↓	↓	↓
paratax	0.173	↓	↓	↓	↓	↓	↓	0.059	↑	↑	↑	↑	↓*	↓*
pastv	0.238	↑	↑	↑	↑	↑	↑	0.966	↑	↓*	↓*	↓	↓	↓
poss	0.006	↑	↑	↑	↑	↓	↑*	0.012	↓*	↓	↓	↓*	↓*	↑
ttr	0.958	↑	↑	↑	↑	↑	↑	0.964	↑*	↑	↑	↑	↑	↑
acl	0.407	↑	↑	↑	↓*	↓*	↓*	0.372	↓	↓	↓	↓	↓	↓
advers	0.003	↑*	↑	↑	↑*	↑	↑	0.002	↑	↑	↑	↑	↑	↓*
adv_verb	0.157	↓	↓	↓	↓	↓	↓	0.117	↑	↓	↓	↓	↓*	↓*
advqua	0.023	↓	↓	↓	↓	↓	↓	0.008	↑	↑	↑	↑	↓	↓*
amod	1.288	↑	↑	↑	↓	↓*	↓*	1.702	↑	↓*	↓*	↓	↓	↓
appos	0.163	↓	↓	↓	↓	↓	↓	0.06	↑	↑	↑	↑	↓	↑
aux	0.959	↓	↓	↓	↓	↓	↓	0.853	↑	↓*	↓*	↓	↓	↓
aux:pass	0.24	↑	↑	↑	↓*	↑	↑	0.248	↑	↑	↑	↓*	↓*	↓*
ccomp	0.468	↓	↓	↓	↓	↓	↓	0.294	↑	↑	↑	↓	↓	↓
cconj	0.034	↓	↓	↓	↓	↓	↓	0.035	↓	↓	↓	↓	↓	↓
compoun	0.082	↑*	↓*	↓*	↓*	↓*	↓*	1.012	↓	↓	↓	↓	↓	↓
conj	1.169	↓	↓	↓	↓	↓	↓	1.139	↓*	↓	↓	↓	↓	↓
cop	0.454	↓	↓	↓	↓	↓	↓	0.529	↑	↓	↓*	↓	↓	↓
demdets	0.012	↑	↑	↑	↑	↑	↑	0.017	↓*	↓*	↓*	↓	↓*	↓*
dens	0.41	↓	↓	↓	↓	↑*	↓*	0.423	↓	↓*	↓	↓*	↑	↑*
deverb	0.016	↑	↑	↑	↑	↑	↑	0.025	↓	↓*	↓*	↓	↑	↑
discourse	0.0	↓*	↓*	↓*	↓*	↓*	↓*	0.003	↑	↑*	↑	↑	↓*	↑
epist	0.005	↓	↓	↓	↓	↓	↓	0.003	↑	↑	↑	↑	↓*	↑
fixed	0.011	↓*	↓*	↓*	↓*	↓*	↓*	0.098	↑	↓*	↑	↓*	↓	↓
flat	0.097	↑	↑	↑	↑	↓	↑	0.076	↑	↑	↑	↑	↑	↑
inf	0.008	↓*	↑	↑	↑	↑	↑	0.019	↓	↓*	↑*	↑	↓	↓*
mark	1.03	↓*	↓*	↓*	↓*	↓	↓	1.32	↑	↓*	↓*	↓	↓	↓
mpred	0.6	↓	↓	↓	↓	↓	↓	0.048	↑	↑	↑	↑*	↓*	↑*

nachfeld	0.362	↓*	↓	↓*	↓	↓	↓	0.095	↑*	↓*	↓*	↓*	↓*	↓*
negs	0.012	↓	↓	↓	↓	↓	↓	0.009	↓	↓	↓	↑	↓*	↓*
nn	0.152	↑	↑	↑	↑	↑	↑	0.199	↓	↓	↓	↓	↓*	↓
nsubj	2.356	↓	↓	↓	↓	↓	↓	1.896	↑	↓*	↓*	↓	↓	↓
numcls	1.406	↓*	↓	↓	↓	↓	↓	1.356	↑*	↓	↓	↓	↓	↓
nummod	0.107	↑*	↓*	↓*	↓*	↓*	↓*	0.238	↓	↓	↓	↓	↓	↓
obj	1.273	↑	↑	↑	↓*	↓	↓	1.306	↑*	↓*	↓*	↓*	↓	↓
obl	1.335	↑	↓	↓	↓	↓	↓	1.304	↑	↓	↓	↓	↓	↓
obl_obj	0.097	↑*	↑	↑*	↓*	↓*	↓*	0.07	↓*	↓*	↓*	↓	↓*	↓*
ppron	0.057	↓	↓	↓	↓	↓	↓	0.046	↑	↑	↑	↑	↓	↓*
prep	0.153	↑	↑	↑	↑	↑	↑	0.108	↓	↓	↓	↓	↓	↓
sconj	0.023	↓*	↓*	↓*	↓	↓	↓	0.024	↑	↑	↑	↑	↓	↓*
self	0.003	↑	↑	↑	↑*	↑	↑	0.0	↑	↑	↑	↓*	↑*	↑*
simple	0.273	↓	↓*	↓*	↑	↑	↑	0.273	↑*	↑	↑	↑	↑	↑
tempseq	0.011	↓	↓	↓	↓	↓	↓	0.004	↑	↑	↑	↑	↓*	↓
vo_noun	0.107	↑*	↑*	↓*	↑	↑	↑	0.629	↑*	↓	↓	↓	↓	↓
vorfeld	0.467	↓*	↓	↓	↓	↓*	↓	0.434	↑	↑	↑	↓*	↓	↓
vs_noun	0.044	↑	↑	↑	↑	↑	↑	0.0	↓*	↓*	↓*	↓*	↓*	↓*
wdlen	5.6	↑	↑	↑	↑	↑	↑	4.742	↓	↑	↑	↑	↑	↑
xcomp	0.269	↓*	↓*	↓	↓	↓	↓	0.369	↑	↑	↑	↑	↓*	↓*

Table 7: Percentage of segments that did not undergo changes in the re-writing pipeline because no translationese was detected in them either by the model or by feature analysis.

	Rewriting Setups				
	RT	Self-guided		Feature-guided	
	–	Min	Detail	Min	Detail
DE	7.92	5.32	0.78	6.24	
EN	0.05	0.05	0.16	29.28	

Appendix C. Examples of prompts by approach and mode

1. **Self-guided approach**: the model has to decide on itself whether a segment contains translationese or not. The same instruction was passed for each pair of segments.

    - Min mode:
      `Your task is to re-write a human translation in a more natural way if necessary.`
      `Here is an original English text: "'In six short months, the presidency has conspired to undermine the Stability Pact, has shown contempt for the European Union's policy towards Russia and offended Canada."'`
      `This is its human translation into German: "'In sechs kurzen Monaten ist es dem Ratsvorsitz gelungen, den Stabilitätspakt zu unterminieren, die Politik der Europäischen Union gegenüber Russland zu missachten und Kanada zu beleidigen."'`
      `If this translation can be revised to sound more like a text originally produced in the target language, return a revised version. If this translation sounds natural enough, return the input translation.`
      `Do not add any meta-phrases or quotation marks. Do not copy the original text.`
    - Detail mode:
      `Your task is to reduce translationese in a human translation by re-writing it in a more natural way where possible.`
      `Translationese refers to any regular linguistic features in the translated texts that make them distinct from texts originally produced in the target language, outside the communicative situation of translation. These features are typically detected by statistical analysis and are explained by the specificity of the translation process. Human translators are known to simplify the source language content and to make it more explicit. Translations can exhibit a tendency to conform to patterns which are typical of the target language, making the output less varied than in comparable non-translations in the target language. The more obvious sign of translationese is interference, which can be defined as over-reliance on the intersection of patterns found in source and target languages. Translationese is manifested in the inflated frequencies of specific linguistic items such as function words (especially connectives and pronouns), unusual frequencies of some parts of speech (especially nouns and adverbs) or grammatical forms (especially forms of verbs), in reduced lexical variety and unexpected lexical sequences, in less natural word order, in longer and more complex sentences as well as lack of target language specific items and structures.`
      `Here is an original English text: "'In six short months, the presidency has conspired to undermine the Stability Pact, has shown contempt for the European Union's policy towards Russia and offended Canada."'`
      `This is its human translation into German: "'In sechs kurzen Monaten ist es dem Ratsvorsitz gelungen, den Stabilitätspakt zu unterminieren, die Politik der Europäischen Union gegenüber Russland zu missachten und Kanada zu beleidigen."'`
      `If you can detect any translationese deviations in this translation, revise this translation to make it sound less translated and return the revised version. If no translationese is detected, return the input translation.`
      `Do not add any meta-phrases or quotation marks. Do not copy the original text.`

2. **Feature-guided approach**: the model is 'micro-managed' in how the translation needs to be adapted, if at all. Each pair of segments gets individual instructions, based on features that were found to strongly deviate from the expected TL norm in this translated segment.

   - Min mode:
     Your task is to re-write a human translation in a more natural way.
     Here is an original English text: "'In six short months, the presidency has conspired to undermine the Stability Pact, has shown contempt for the European Union's policy towards Russia and offended Canada."'
     This is its human translation into German: "'In sechs kurzen Monaten ist es dem Ratsvorsitz gelungen, den Stabilitätspakt zu unterminieren, die Politik der Europäischen Union gegenüber Russland zu missachten und Kanada zu beleidigen."'
     Re-write this translation following the instructions:
     Use pronouns instead of nouns as verbal arguments where possible.
     Avoid constructions with indirect objects.
     Do not add any meta-phrases or quotation marks. Do not copy the original text.
   - Detail mode:
     Your task is to reduce translationese in a human translation by re-writing it in a more natural, less translated way.
     Translationese refers to any properties of translations that make them statistically distinct from texts originally produced in the target language.
     Here is an original English text: "'In six short months, the presidency has conspired to undermine the Stability Pact, has shown contempt for the European Union's policy towards Russia and offended Canada."'
     This is its human translation into German: "'In sechs kurzen Monaten ist es dem Ratsvorsitz gelungen, den Stabilitätspakt zu unterminieren, die Politik der Europäischen Union gegenüber Russland zu missachten und Kanada zu beleidigen."'
     Revise this translation following the instructions which reflect deviations of this segment from the expected target language norm:
     Use pronouns instead of nouns or proper names as verbal arguments where possible.
     Avoid constructions with indirect objects. An indirect object of a verb is any nominal phrase that is an obligatory argument of the verb but is not its subject or direct object. The prototypical example is the recipient (dem Kind) with verbs of exchange: Die Frau gibt dem Kind einen Apfel.
     Do not add any meta-phrases or quotation marks. Do not copy the original text.

3. **Re-translation approach**: the model is tasked with re-translation instead of re-writing/revising the existing translation. We keep the prompt as similar as possible to *Self-guided min* setup, while changing the core task formulation. The aim is to compare revised and re-translated versions to see whether the model makes a difference between them.

   Your task is to re-translate a human translation to make it more natural in the target language if necessary.
   Here is an original English text: "'In six short months, the presidency has conspired to undermine the Stability Pact, has shown contempt for the European Union's policy towards Russia and offended Canada."'
   This is its human translation into German: "'In sechs kurzen Monaten ist es

dem Ratsvorsitz gelungen, den Stabilitätspakt zu unterminieren, die Politik der Europäischen Union gegenüber Russland zu missachten und Kanada zu beleidigen.“‘
If this translation can be re-translated to sound more like a text originally produced in the target language, return a re-translated version. If this translation sounds natural enough, return the input translation.
Do not add any meta-phrases or quotation marks. Do not copy the original text.

Appendix D. Examples of source segments, human translation as well as all output modes

Ex.	Type	Segments
1	Source:	Does he regard the committee whose advice he rejects as incompetent, unable to judge what is good for animals or dangerous for human health?
	HT:	Hält er den Ausschuß, dessen Empfehlung er ablehnt, für nicht kompetent, für unfähig zu entscheiden, was für die Tiere gut oder für die menschliche Gesundheit gefährlich ist?
	RT:	Betrachtet er den Ausschuss, dessen Ratschlag er ablehnt, als inkompetent und unfähig zu beurteilen, was gut für Tiere oder gefährlich für die menschliche Gesundheit ist?
	Self-guided Min:	Glaubt er etwa, der Ausschuss, dessen Ratschläge er ignoriert, sei inkompetent und unfähig zu beurteilen, was gut für Tiere ist oder der menschlichen Gesundheit schaden könnte?
	Self-guided Detail:	Hält er den Ausschuss, dessen Ratschläge er ablehnt, für inkompetent und unfähig zu beurteilen, was gut für Tiere ist oder dem menschlichen Wohlbefinden schaden könnte?
	Feature-guided Min:	Betrachtet er den Ausschuss, dessen Rat er entschieden ablehnt, tatsächlich als inkompetent, als gänzlich unfähig zu beurteilen, was ausgerechnet für Tiere gut oder potenziell für die menschliche Gesundheit schädlich ist?
	Feature-guided Detail:	Hält er tatsächlich den Ausschuß, dessen Empfehlung er konsequent ablehnt, für gänzlich inkompetent, für absolut unfähig, um zu beurteilen, was letztendlich für die Tiere gut oder womöglich für die menschliche Gesundheit gefährlich ist?
2	Source:	I am pleased that our own legal services – who say Parliament must be very wary of attempts from representatives of other institutions to limit our powers – have confirmed that there is nothing in the rules to stop us having a parliamentary inquiry, so long as the terms of reference differ from any question before the courts in a Member State.
	HT:	Ich bin froh, dass unsere eigenen juristischen Dienste – die sagen, das Parlament müsse vor den Versuchen von Vertretern anderer Institutionen, unsere Befugnisse zu beschneiden, sehr auf der Hut sein – bestätigt haben, dass es in der Geschäftsordnung keine Bestimmung gibt, die uns von einer parlamentarischen Untersuchung abhalten könnte, solange die Angelegenheit nicht bei den Gerichten eines Mitgliedstaates anhängig ist.
	RT:	Ich freue mich, dass unsere eigenen Rechtsdienste - die warnen, dass das Parlament bei Versuchen von Vertretern anderer Institutionen, unsere Befugnisse einzuschränken, äußerst vorsichtig sein muss - bestätigt haben, dass es in den Regeln nichts gibt, was uns daran hindern könnte, eine parlamentarische Untersuchung durchzuführen, solange die Untersuchungsgegenstände sich von denen unterscheiden, die vor Gericht in einem Mitgliedstaat verhandelt werden.
	Self-guided min:	Es freut mich, dass unsere eigenen Rechtsdienste bestätigt haben, dass es keine Regel gibt, die uns daran hindert, eine parlamentarische Untersuchung durchzuführen. Sie warnen jedoch, dass wir sehr vorsichtig sein müssen, wenn Vertreter anderer Institutionen versuchen, unsere Macht zu begrenzen. Dies gilt natürlich nur, solange die Untersuchungsthemen nicht bereits vor den Gerichten eines Mitgliedslandes verhandelt werden.
	Self-guided Detail:	Es freut mich, dass unsere eigenen Rechtsdienste bestätigt haben, dass nichts in der Geschäftsordnung uns daran hindert, eine parlamentarische Untersuchung durchzuführen - solange die Untersuchungsgegenstände nicht bereits vor Gericht in einem Mitgliedsstaat verhandelt werden. Sie warnen das Parlament jedoch, wir sollten wachsam gegenüber Versuchen von Vertretern anderer Institutionen sein, unsere Macht einzuschränken.

Feature-guided Min:	Die Freude ist groß, die eigenen juristischen Dienste haben bestätigt. Sie warnen das Parlament, behutsam zu sein. Vertreter anderer Institutionen könnten versuchen, Befugnisse zu beschneiden. Keine Bestimmung in der Geschäftsordnung könnte eine parlamentarische Untersuchung verhindern. Dies gilt, solange die Angelegenheit nicht bei Gericht eines Mitgliedstaates vorliegt.
Feature-guided Detail:	Es freut mich, dass ein Bestätigung von den eigenen juristischen Diensten vorliegt. Sie warnen das Parlament, wachsam gegenüber Versuchen von Vertretern anderer Institutionen zu sein, unsere Befugnisse zu beschränken. In der Geschäftsordnung gibt es keine Regel, die eine parlamentarische Untersuchung verhindern könnte. Dies gilt solange der Sachverhalt nicht vor den Gerichten eines Mitgliedstaates liegt.

		Rewriting Setups				
		Baseline	Self-guided		Feature-guided	
			Min	Detail	Min	Detail
DE	15	81.06	-0.27	-1.01	**-2.39**	-2.21
	58	81.51	0.53	**-0.56**	0.06	-0.28
EN	15	75.60	-2.70	-4.10	-3.18	**-7.63**
	58	78.30	-1.40	-1.61	-1.61	**-4.07**

Table 12: Differences in F1 scores between the segment-level results on the rewritings and on human translations from the contrastive sample (Table 4). The best results for each feature set are shown in bold.

		Rewriting Setups			
		Self-guided		Feature-guided	
		Min	Detail	Min	Detail
DE	R	0.87	0.86	0.84	0.85
EN	R	0.85	0.84	0.80	0.82

		Rewriting Setups			
		Self-guided		Feature-guided	
		Min	Detail	Min	Detail
DE	A	5.8	5.8	5.1	5.4
	F	5.7	5.6	5.4	5.4
EN	A	5.7	5.9	5.2	5.4
	F	6	5.9	5.6	5.8

Table 13: Results of human evaluation for accuracy (A) and fluency (F) in a 1-6 Likert scale.