Diversity and Homogenisation in Generative AI Translation: A Comparative Study of English-Dutch Translation Across Domains

Dimitar Shterionov
Department of Intelligent Systems,
Tilburg University
d.shterionov@tilburguniversity.edu

Noa van Helleman
Damen Naval
noavhelleman@gmail.com

Eva Vanmassenhove
Department of Computational Cognitive Science,
Tilburg University
e.o.j.vanmassenhove@tilburguniversity.edu

Abstract

Generative AI tools, such as ChatGPT, are applied to a wide range of language-related tasks, including translation. Despite their current popularity among users and researchers and the impressive results obtained on several benchmarks (Kocmi et al., 2024; Deutsch et al., 2025), their potential side-effects on languages and translations are still understudied (Vanmassenhove, 2025). The paradigm shift from Machine Translation (MT) to Generative AI Translation (GAIT) likely calls for a reconsideration of our assessment and evaluation metrics and practices. In this work, we focus on GAIT by analyzing translations from four multilingual large language models (MLLMs), *mBART*, *Jamba-1.5-large*, *GPT 4o* and *DeepSeek R1* applied to three different domains (news, literature and poetry) for the English-Dutch language pair. Focusing on metrics related to lexical and textual diversity, we find that while GAIT text for literature is of significantly high lexical and grammatical richness, that is not the case for news and poetry. We also assess the homogeneity of AI-generated text through a set of clustering and classification experiments. In addition to a clear separation between human- and AI-generated content, our results indicate that GAIT output is more homogeneous among MLLMs.

1 Introduction

Generative AI tools like ChatGPT[1], DeepSeek[2] or Claude[3] took the world by storm (Vujović, 2024). Thanks to their impressive performance on a wide range of tasks and ease of access (through publicly available APIs or chat interfaces), the initial media attention was followed by their widespread adoption (Liang et al., 2025). These tools and their underlying technology (Large Language Models (LLMs) and other foundation models) have caused a paradigm shift in the general domain of AI (Schneider et al., 2024) and in various subdomains, including Machine Translation (MT), where they increasingly complement or replace dedicated systems (Gain et al., 2026).

Although multilingual LLMs are not explicitly trained for translation, recent evaluations show that they can match or outperform traditional end-to-end systems across many language pairs and domains (Gain et al., 2026; Kocmi et al., 2024; Deutsch et al., 2025). Their strong performance has led to their increasing integration into professional and everyday, non-professional translation practices, impacting the translation sector, workflows, and translator education (Wang and Wu, 2026). At the same time, this shift raises concerns about the (linguistic, stylistic, and so on) characteristics of LLM-generated translations, particularly regarding standardization and loss of diversity (Vanmassenhove, 2025).

The aforementioned developments call for a more detailed analysis of the (long-term) impact of LLM-based translation on language and translation practices, a complex and still underexplored

[1] https://openai.com/blog/chatgpt
[2] https://chat.deepseek.com/
[3] https://www.claude.ai

challenge. With this paper, we take a step towards addressing this by (i) assessing the lexical and grammatical diversity of LLM-generated translations (EN→NL) across three domains in line with previous work on (post-edited) machine-translated texts (Vanmassenhove et al., 2019; Toral, 2019; Vanmassenhove et al., 2021); and (ii) comparing the output of different LLMs (across architectures, generations,and sizes) as a proxy for cross-model convergence (Jiang et al., 2025) and potential language homogenization, and assess the extent to which such convergence renders model outputs indistinguishable.[4]

A Brief Note on Terminology In this work, we particularly focus on Generative Artificial Intelligence Translation (GAIT). While the term Machine Translation (MT) is also used nowadays to refer to applications where generative or foundation models are used for translation, we opt to use the term GAIT in this paper to make a more clear distinction between the so-called more *traditional* AI approaches to MT (e.g. Neural Machine Translation (NMT)) which do not rely on foundation models and those that do.

2 Related Work

LLMs mark a substantial shift in how language is produced, circulated, and reused. Unlike earlier Natural Language Processing (NLP) systems that were typically developed for narrow, task-specific applications, contemporary foundation models operate across domains and increasingly participate *directly* in text production (Vanmassenhove, 2025). As a result, they do not only support writing and translation, but also potentially directly shape the linguistic material that later re-enters model training pipelines. This raises concerns that the widespread reuse of model-generated text may gradually alter the distribution of (written) language itself and eventually also affect the quality of our models. A growing body of work frames this risk in terms of *model collapse*: when models are trained on data increasingly contaminated by their own outputs, low-probability events tend to disappear first, causing distributions to narrow over time (Shumailov et al., 2023). In computer vision, such iterative degeneration has been visualized as progressively distorted artifacts (Alemohammad et al., 2023). In language, the effects are subtler, but arguably more consequential, they concern not only output quality, but also the preservation of lexical, syntactic, and semantic variation (Vanmassenhove, 2025).

This broader perspective is relevant as MT- and GAIT-generated texts are no longer peripheral artifacts: they increasingly populate the web and thereby become part of the data from which future systems learn. Thompson et al. (2024) show that a substantial amount of multilingual web content is machine translated, especially for lower-resource languages such content may constitute a large fraction of the available web text. This is important beyond data quality concerns alone as it suggests that model-produced translations are already feeding back into the multilingual textual environment on which MT systems and LLMs depend.

While concerns regarding convergence and collapse have been gaining some traction, these concerns do not originate from research on translation with LLMs. Earlier work in MT already showed that statistical and neural models systematically favor more frequent lexical and morphological forms, often at the expense of more rare alternatives. Vanmassenhove et al. (2019) and Toral (2019) were among the first to argue that MT output tends to be lexically less rich than human translation. Subsequent work confirmed this tendency across language pairs, domains, and architectures. For instance, Vanmassenhove et al. (2021) show that both SMT and NMT reduce lexical and morphological diversity relative to the original training data. Related studies similarly report that MT often diverges from original data in terms of linguistic features (e.g. average sentence length, ngram features and lexical diversity) (de Clercq et al., 2021). Aranberri and Pascual (2025) simulate multiple generations of MT by training sequential systems on previously generated outputs and shows an initial loss of lexical diversity that continues (at a much slowed pace) in following generations. Structural, specifically morphological, variation remains comparatively stable, suggesting that convergence primarily operates at the lexical level rather than uniformly across linguistic structure. In parallel, other research on structural divergence finds that MT output often remains closer to the source text (more on-to-one alignments), shows less morphosyntactic diversity

[4]Test sets and code are available at: `https://github.com/dimitarsh1/LLM_LexRichness`. Our evaluation code includes additional evaluation results which could provide further insights to the interested reader.

and more convergent patterns (Luo et al., 2024).

Taken together, these findings position MT as an early empirical case of what is now being discussed more broadly as convergence, homogenization or diversity reduction in generative models. Recent work extends these concerns to LLMs. Guo et al. (2025) show that LLM-generated texts generally exhibit lower diversity than human-authored texts, although the extent of this depends on the level of analysis: lexical, syntactic, and semantic diversity do not necessarily move together. Another line of research aligned with these distortions introduced by LLMs studies millions of PubMed abstracts (Kobak et al., 2024). They identify abrupt post-ChatGPT increases of words such as *delve*, *crucial*, and *significant*, arguing that LLM usage is measurably altering academic writing. Extending this perspective from writing to speech, Geng et al. (2025) examine both papers and conference presentations and show that LLM-associated lexical patterns are present across written and spoken academic discourse. Similarly, Yakura et al. (2024) provide large-scale evidence that AI-associated lexical choices are increasingly appearing in human spoken communication, suggesting that model-preferred phrasing is not confined to mediated writing environments but can diffuse into speech as well. For LLM-generated translations, Zhang et al. (2025) report that literary translations produced by LLMs remain more literal and less diverse than human translations, even when they can be seen as competitive according to more conventional automatic metrics.

We ought to note that research in NLP often relies on pairwise comparisons between a single AI-generated text and a single human-written text. Given that LLMs are trained on large-scale corpora reflecting the linguistic patterns of many speakers, their outputs can exhibit greater surface-level variation than those of an individual human text (as reported in e.g. Brglez and Vintar (2022) or Castilho and Resende (2022)). Such findings may lead to misleading conclusions about overall lexical richness or diversity. Most comparisons overlook variation across human authors and are therefore not well-suited for assessing distributional properties of language as a whole (Vanmassenhove, 2025). More broadly, this connects to recent evidence of convergence in generative models: Jiang et al. (2025) describe an *Artificial Hivemind* effect, where both individual models (intra-model homogenization) and different model families (inter-model homogenization) produce increasingly similar outputs. In the context of translation, this implies that apparent 'increased' diversity at the instance level may coexist with reduced diversity at the system level, as models repeatedly favor high-probability lexical and structural choices.

Last, this literature is complemented by broader evidence that linguistic diversity itself may be shrinking under LLM-assisted writing conditions. Sourati et al. (2025) report declines in linguistic diversity associated with LLM use and argue that this has implications not only for expression, but also for the kinds of social and psychological signals language carries. Together, these studies seem to support a feedback-loop hypothesis: LLMs draw on human-produced language, amplify some forms over others, and then feed these preferences back into human writing and speaking, where they may gradually normalize. Rresearch on literary translation and creativity indicates that these systems tend to favor high-probability lexical and structural choices, resulting in reduced stylistic variation compared to human translations, even under different prompting strategies (Du et al., 2025; Arenas and Toral, 2022). This apparent discrepancy can be partly explained by methodological differences: most evaluations rely on pairwise comparisons and are therefore not well-suited for assessing distributional properties of language (Vanmassenhove, 2025).

In this paper, we examine whether translations produced by generative AI differ from human-generated text in terms of linguistic richness, and how such differences relate to translation quality. In contrast to prior work, we compare multiple LLMs and assess a broad set of lexical and grammatical indicators, while explicitly interpreting diversity in relation to translation quality rather than as an unconditional good in itself. Along with the comparison between LLM outputs and human-generated texts in terms of translation quality evaluation metrics and lexical and grammatical diversity metrics, we also conduct an empirical evaluation in terms of separability between GAIT and human-generated text. We use these findings to assess whether LLMs produce homogeneous translations. For this purpose we train and evaluate classification and clustering models. First, LR classifiers are trained for each domain to predict the origin of

the translation, i.e. whether it is human-generated, mBART, Jamba, GPT or DeepSeek. We analyse these models through the precision, recall and F1 scores on held-out test sets. Second, we trained k-means models (one for each domain) to cluster the data in 5 classes. We analysed the distribution of the data into the derived clusters. The analysis of the classification performance on the held out test sets and of the cluster distributions shows that human-generated text is clearly separable from the GAIT text and that GAIT translations are less separable between themselves.

3 Experiments

To test the translation capacity and lexical richness of the generated translations of the selected LLMs, we choose three EN–NL test sets (news, poetry, and literature) and translate them using best practices for prompting and parameter setup. Before further detailing the datasets and overall experimental setup, we would like to provide some background on the specific translation domains as they differ in their constraints on form, meaning, and overall variation.

3.1 Translation Domains

News prioritizes adequacy, fluency, and consistency. MT for the news domain has been extensively studied in the WMT shared tasks[5], where MT systems achieve strong performance, although claims of human parity (Hassan et al., 2018) remain contested since these conclusions are sensitive to the evaluation design, annotator design and the presence of translationese in the source (Toral et al., 2018). Since 2021, the submissions to the WMT shared task show a shift toward GAIT for the news domain ((Símonarson et al., 2021; Wu and Hu, 2023; Kocmi et al., 2024; Kocmi et al., 2025) which, in recent editions show, hes turned into the standard approach (Símonarson et al., 2021; Kocmi et al., 2024; Kocmi et al., 2025).

Poetry is sometimes constrained by formal properties such as rhythm and rhyme, has a higher semantic ambiguity and is often more culture-dependent than other domains making it particularly challenging for MT (Ghazvininejad et al., 2018; Humblé, 2020; Dunder et al., 2020; Seljan et al., 2020). Recent work shows some improvements (professional human and automatic evaluation metrics) in automatic poetry translation through domain-specific prompting strategies (Wang et al., 2024).

Literature, similarly to poetry, places a strong(er) emphasis on preserving not only meaning but also style, tone, and the overall reading experience, making it particularly demanding for MT systems (see a.o. (Toral and Way, 2015; Toral and Way, 2018; Guerberof-Arenas and Toral, 2020; Besacier and Schwartz, 2015; Karpinska and Iyyer, 2023)). Prior work shows that MT outputs tend to exhibit reduced creativity and stylistic variation compared to human translations, even when post-edited, and that maintaining coherence and register remains a major challenge (Arenas and Toral, 2022; Fonteyne et al., 2020). Recent work on GAIT for literary translation has explored the impact of prompting strategies on translation quality and creativity (Du et al., 2025). While such approaches can improve output quality, findings show that even 'creative' LLM-generated translations - measured using established creativity metrics (Guerberof-Arenas and Toral, 2020; Toral and Way, 2018) – do not reach human translation performance in terms of creativity and stylistic variation (Du et al., 2025). Related work investigates the use of LLMs for post-editing literary MT, showing that while such models can increase lexical variation, their edits are often more surface-level and less effective at correcting translation errors than those of professional translators (Macken, 2024).

3.2 Dataset

We use the Dutch Parallel Corpus[6] (DPC) (Macken et al., 2011) to extract news and literature texts. For poetry, we use the PoetryTranslationEMNLP2021[7] dataset (Chakrabarty et al., 2021) dataset.

The DPC is a parallel corpus with over 10 million words containing high quality aligned sentences from five different types of texts. It is balanced with respect to direction of translation and type of text. The PoetryTranslationEMNLP2021

[5]See https://www.statmt.org/

[6]Available from: https://taalmaterialen.ivdnt.org/download/tstc-dutch-parallel-corpus-niet-commercieel/

[7]Available from: https://github.com/tuhinjubcse/PoetryTranslationEMNLP2021

corpus contains over 190,000 high quality translations for six languages, including Dutch.

Both corpora consist entirely of human-produced data, but differ in how original texts and translations are represented. The DPC includes both original texts and their human translations in parallel. The poetry dataset consists of original (human-authored) texts (in Dutch, Russian...) paired with their English (human-produced) translations. From these datasets we extracted 1000 English-Dutch parallel sentences per category (news, literature and poetry) as test sets.

3.3 Translation Models

We compared different models – the "old-school", encoder-decoder mBART, the Transformer-based, decoder-only GPT's 4o, and DeepSeek r1, as well as the hybrid Transformer-Mamba model – Jamba-1.5 (Jamba Team et al., 2024).

We used the largest mBART-large-50 model, available through the Huggingface Transformers library (Wolf et al., 2020), including both the mBART tokenizer and the mBART model. The jamba-1.5-large model we used, was accessed through the AI21's API. Similarly, we used GPT-4o via the OpenAI's API. The DeepSeek model we used was the reasoning R1, 70b model accessed from the Ollama [8] platform.

Each model is evaluated under conditions aligned with its intended usage and recommended decoding configuration. To account for architectural differences, each model was evaluated using prompting and decoding settings aligned with its recommended usage. Instruction-tuned language models require explicit task prompts, whereas the encoder–decoder model (mBART) operates without prompting. Decoding temperatures were selected based on standard practices for each model type. The prompts and temperature settings are:

- For **mBART** we used no prompt, but simply decoded the English sequence (after tokenization) into Dutch.
- **jamba-1-5-large** *Translate the following text from English to Dutch, showing only the translation:* followed by the sentence to translate. Temperature is set to 0.0. General recommendations for translation is to set the temperature value between 0.0 and 0.2; we set it to 0.0.[9]
- **gpt-4o** *Show only translation, English: [SRCLINE] = Dutch:*, where *SRCLINE* is the source sentence. We used the default temperature, which is equal to 1.0[10] and considered the medium temperature.
- **deepseek-r1:70b** *Translate the following text from English to Dutch while preserving the original meaning, tone, and context. Maintain proper grammar and natural fluency as if written by a native speaker. Don't show the source nor your reasoning. Only show the final translation. The text is: [SRCLINE]* where *SRCLINE* was the source sentence. Temperature is set to 0.6[11]

We set the `role` parameter to *user* for the three LLMs (jamba-1.5-large, gpt-4o and deepseek-r1:70b). We manually inspected and cleaned all outputs using regexes and correction of UTF encoding errors (See Appendix A).

3.4 Evaluation

Standard Automatic Metrics & Diversity Metrics We evaluate our results using (i) automatic MT evaluation using BLEU (Papineni et al., 2002), TER (Snover et al., 2006), chrF (Popović, 2015) from the Sacrebleu package (Post, 2018), Comet (Rei et al., 2020) and BLEURT (Sellam et al., 2020) (ii) lexical diversity evaluation through TTR, Yule's I and MTLD, Simpson and Shannon diversity scores; we also assessed the number of lemmas and singleton lemmas, and vocabulary sizes following the methodology of (Vanmassenhove et al., 2021).

Classification & Clustering Additionally, we analyse model-specific output characteristics by training logistic regression classifiers (precision, recall, F1) to predict the source model and apply k-means clustering to assess the separability of model outputs. The rationale behind it being that if translations produced by different models are difficult to distinguish or form overlapping clusters, this indicates higher similarity across the different translation systems.

Prior to training the classification and clustering models, we pre-processed the data by removing

[8] `https://ollama.com/`

[9] `https://docs.ai21.com/docs/prompt-engineering`

[10] `https://developers.openai.com/api/reference/resources/chat`

[11] as recommended in `https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B\#usage-recommendations`

brackets, accolades, commas, semiclons, @'s and duplicate and other special characters followed by stopwords removal using nltk (Bird et al., 2009). The resulting dataset, consisting of 12000 translations, was then split into a training, validation and test set with the ratios 0.56, 0.14 and 0.3. This data was then vectorised using count vectorizer and the tfidf transformer from scikit-learn (Pedregosa et al., 2011).

We trained LR classifiers on the combined training and validation set and evaluated them on the test set. We evaluate the classification models using precision, recall and F1. We trained three sets of classification models for different $l1\_ratio$ values. The $l1\_ratio$ parameter controls the type of regularisation: $l1\_ratio$ of 0 implies $l2$ (or ridge) regularisation, while a ratio of 1 implies $l1$ or lasso regression. While ridge regression is typically considered better for text data as it utilises all data and handles correlated features, we experimented also with $l1$ which could guide the algorithm to focus on more important features (i.e. words / tokens), zero-ing out less important ones; we also use a combination of both $l1$ and $l2$ at ratio 0.5.

4 Results

The automatic MT evaluations are presented in Section 4.1, the lexical diversity evaluation can be found in Section 4.2. Section 4.3 covers the classification and clustering experiments.

4.1 Automatic MT Evaluation

Table 1 shows the scores of both statistical and neural evaluation metrics attained per domain.

Domain	Model	BLEU↑	chrF↑	TER↓	BLEURT↑	Comet↑
News	mbart	34.81	62.19	50.57	60.81	89.03
	jamba	38.02	65.43	48.71	61.72	90.11
	gpt	**39.79**	**66.72**	**47.19**	**63.21**	**90.71**
	deepseek	35.33	63.99	51.30	61.52	90.02
Poetry	mbart	20.63	46.39	69.41	49.14	70.83
	jamba	27.67	52.95	62.61	53.04	74.10
	gpt	**31.67**	**54.79**	**58.37**	**55.68**	**75.77**
	deepseek	25.24	51.11	65.05	53.40	74.10
Lit.	mbart	22.01	51.02	65.02	51.14	81.50
	jamba	27.84	55.93	60.71	55.69	85.19
	gpt	**29.53**	**57.76**	**58.97**	**57.23**	**85.91**
	deepseek	25.55	54.99	63.59	54.64	84.75

Table 1: Evaluation metrics scores: BLEU, chrF, TER, BLEURT and Comet.

For all three categories and according to all statistical and neural metrics GPT-4o outperforms the rest and mBART performs the worst. Jamba ranks second and DeepSeek third. The absolute difference in the values of all evaluation metrics between those of GPT and Jamba is the smallest (as compared to the rest). We also observe a high difference between the values for the different translation domain (within one domain and through each domain, the rankings are the same as noted above).

4.2 Lexical and Grammatical Diversity

We present (i) standard lexical diversity metrics: Yule's I, Type-token ratio (TTR) and Measure of textual lexical diversity (MTLD); (ii) grammatical diversity scores: Shannon and Simpson metrics, along with the number of lemmas and singleton lemmas and (iii) vocabulary sizes in Table 2.

These results show that with exception to the translations for the literature domain, the references (i.e. the human-generated text) are of higher lexical and grammatical richness. However, it is worth noting that only for poetry, the variability in the metrics is high. It is also worth noting that mBART generally has the lowest scores.

4.3 Classification and clustering

The precision, recall and F1-scores of evaluating our classifiers are summarised in Figure 1. The figure shows the progression from $l2$ regularisation to $l1$ regularisation for all three metrics.

To assess the clustering algorithm, we looked at the distribution of the different translations along the 5 clusters, summarised in Table 3 in percentages.

5 Analysis

5.1 Translation quality metrics

The results summarised in Section 4.1 show a very consistent performance across domains and metrics for all models. GPT generates the best translations for all domains, followed by Jamba and DeepSeek (which swap rankings according to BLEURT and COMET) for the Poetry domain, and mBART which performs the worst. These suggest that GPT generalises best across domains.

The highest scores are on News translations, e.g. Comet scores are between ± 86 and ± 91 compared to ± 81 to ± 86 for the Literature domain and to ± 70 – ± 76 for the Poetry domain. These scores reflect the degree of difficulty in translating News text (structured and well standardised, easiest to translate) as compared the more stylistic and cre-

		Standard Metrics			Grammatical Metrics				Vocabulary	
Domain	model	Yule's I ↑	TTR↑	MTLD↑	Num. of all lemmas↑	Num. of single lemmas↑	Shannon↑	Simpson↓	Avg. sent length	Voc size
News	Ref	**5.4398**	**0.2321**	103.2077	**3959**	**3434**	**0.9388**	**0.0943**	20.257	17 995
	mbart	4.2812	0.2106	97.1528	3516	3007	0.9332	0.1035	19.979	17 641
	jamba	5.0875	0.2266	103.915	3744	3215	0.9347	0.1012	19.876	17 849
	gpt	5.0562	0.2248	100.7675	3697	3152	0.9317	0.1053	19.952	17 974
	deepseek	5.0386	0.2229	**104.172**	3730	3194	0.9351	0.0999	20.159	18 120
Poetry	Ref	**12.4715**	**0.3384**	**182.6479**	**2138**	**1922**	**0.9497**	**0.079**	7.444	6 651
	mbart	3.8876	0.2704	122.7931	1814	1578	0.9363	0.1005	8.204	6 815
	jamba	3.8601	0.2882	152.2353	1975	1752	0.9433	0.0897	8.272	6 828
	gpt	7.1788	0.3015	159.1791	1971	1744	0.943	0.0903	7.920	6 860
	deepseek	7.5999	0.3027	162.377	1972	1727	0.9383	0.0968	8.106	7 050
Literature	Ref	4.1752	0.2063	131.4198	3578	3069	0.9314	0.1068	21.226	18 326
	mbart	3.6372	0.1966	116.9994	3490	2971	0.9293	0.1100	21.637	18 532
	jamba	4.1934	**0.2086**	127.043	**3714**	**3192**	**0.9330**	**0.1046**	21.534	18 638
	gpt	4.2391	0.2073	127.5569	3675	3126	0.9288	0.1103	21.751	18 909
	deepseek	**4.3307**	0.2077	**132.4669**	3713	3183	0.9305	0.1078	21.838	18 949

Table 2: Lexical and grammatical diversity measurements of tokenized output (using Moses tokenizer from SacreMoses: `https://github.com/hplt-project/sacremoses`).

	News					**Poetry**					**Literature**				
Model	C0	C1	C2	C3	C4	C0	C1	C2	C3	C4	C0	C1	C2	C3	C4
ref	**78.3**	9.5	6.2	0.9	5.0	0.1	0.4	2.8	1.4	**95.3**	1.8	1.7	2.7	5.2	**88.6**
mbart	76.0	10.4	6.0	1.1	6.5	0.1	0.5	5.0	1.8	92.6	1.7	2.3	3.2	5.3	87.5
jamba	75.9	10.2	6.5	1.1	6.3	0.3	0.5	4.6	1.9	92.7	1.6	2.8	3.1	5.2	87.3
gpt	76.2	9.9	6.6	1.1	6.2	0.3	0.5	4.2	1.9	93.1	1.8	1.8	3.2	5.2	88.0
deepseek	77.6	8.5	6.3	0.9	6.6	0.3	0.5	3.0	2.1	94.1	1.7	1.6	3.4	5.4	87.9
Silhouette score:	0.0029					0.0047					0.0049				

Table 3: Cluster distribution (in %) and silhouette across domains and models. Bold values indicate the dominant cluster per domain. Silhouette scores are close to zero, indicating poor clustering for the selected number of clusters.

ative literary text (medium) and the ambiguous Poetry translations (hardest to translate).

The variability of the performance across domains of the investigated models suggests that even though these LLMs achieve high-quality translations they are still domain-sensitive.

5.2 Lexical diversity

We analyse the lexical and grammatical diversity of GAIT and human-generated text over the three domains. Across the News and Poetry domains the human-generated texts exhibit consistently higher diversity than GAIT outputs, as reflected in Yule's I, TTR, MTLD (except for the case of DeepSeek translations for the News domain), Shannon, and Simpson scores. This indicates that LLMs produce more concentrated lexical distributions than human translations for these domains.

The differences are strongly domain dependent with the Poetry domain translations differences being the most pronounced ones: all models, especially mBART and Jamba, yield a substantially lower diversity than human-generated text, reflecting their lack of capacity to address the stylistic and interpretative requirements for this domain. As noted in Section 3.1 the reference text (i.e. the Dutch text), which we compare the GAIT output to, is human-authored and not translated; the English text in the Poetry domain, which is the translation source, is human translated. The fact that in the case of the Poetry domain the translations are compared to original human-authored text may (partially) explain the results and deserves further analysis. This raises an additional research question which interleaves human and machine translation and goes beyond the scope of this work. We leave it, therefore, for future work.

For the literature domain, the results slightly contrast those for the News and Poetry domains. While the differences are small, the lexical and grammatical diversity scores for the human-generated text are ranked 3^{rd} or 4^{th}. This suggests that GAIT text approximates or even surpasses human-generate text in terms of diversity on the literature domain. This contrasts with the Poetry, where the gap is largest, and with the News domains. These rankings (within the literature domain) are also reflected in the average sentence length and the vocabulary sizes with the human-generated text having the lowest counts for both.

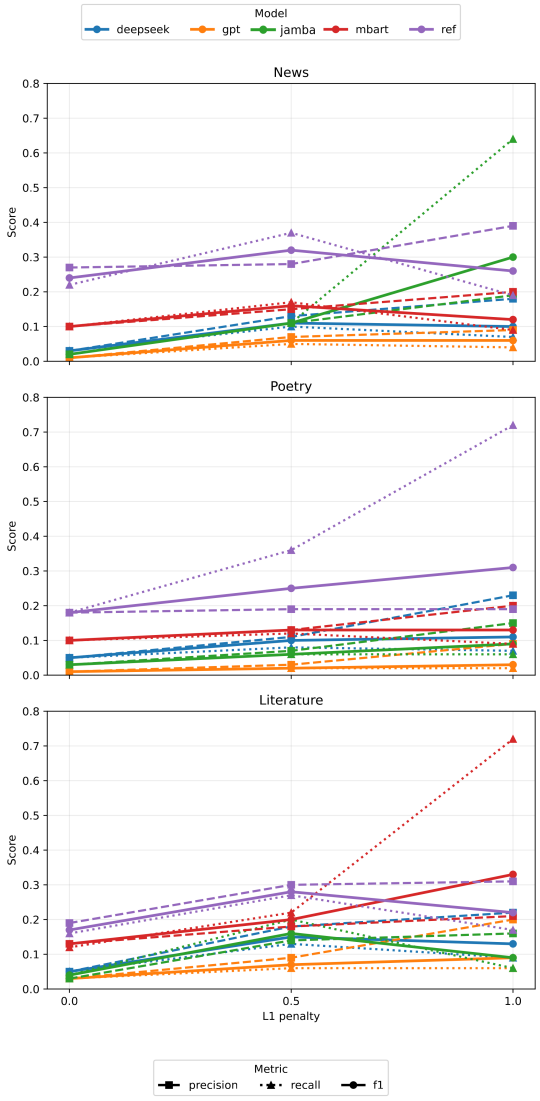

Figure 1: Precision, recall and F1-score for the classifier trained and evaluated on the data for different $l1$ values.

Looking at these values for the Poetry domain we observe that the rankings are the same (the human-generated text has the smallest average sentence length and smallest vocabulary size). These results hint towards a more biased use of certain words / tokens leading to a flatter tail in the frequency-of-use distribution for these translations (aligning with the observations by (Vanmassenhove, 2025)). Comparing the different models, mBART consistently yields the lowest diversity. It should be noted that Jamba, GPT-4o and DeepSeek produce higher and, most importantly, closely aligned scores. With exception to the output of Jamba for the Poetry domain, the texts generated by the LLMs, excluding mBART, have a very small standard deviation from the mean scores, i.e. have a very small inter-model variance. This suggests *similar distributional behavior across architectures*.

When comparing the diversity metrics to the translation quality evaluation metrics, the results reveal an inconsistency between quality and diversity. Model rankings according to the translation quality evaluation metrics do not consistently align with the rankings according to the diversity metrics. This is very obvious for the Literature domain where, for example, DeepSeek translations are scored highest according to Yule's I and MTLD, even surpassing human-generated text (in terms of diversity) while they are constantly ranked third according to the evaluation metrics. This observation is aligned with previous work and supports the statement that evaluation and lexical / grammatical diversity metrics are not interchangeable and should both be taken into account, especially in cases such as literary or poetry translation.

5.3 Separability between AI and human translations

The classification and clustering experiments provide us with an idea about the degree of separability of the reference in comparison to the rest. We assess how pronounced the predictions are in favour of the references, i.e. the human-generated text. In the context of classification that would imply a more separable / concentrated distribution represented by higher precision –which indicates cleaner boundaries– higher recall –which indicates higher number of correctly identified– and higher f1 – which indicates a higher separability, overall. In the context of clustering we analyse the cluster distributions.

Classification Across all domains, higher precision, recall and f1 scores are observed for the values of $l1\_ratio$ equal to 0.0 and 0.5. This illustrates that the classifier more reliably distinguishes the human reference class under L2 and elastic-net regularisation. Interstingly, and especially for $l1\_ratio = 0.0$, we observe that classifying mBART translations stands out, not because it has the second highest scores, but because the values for the other models are much closer together. When the classifier is trained with an $l1$ regulariser (i.e. $l1\_ratio = 1.0$) we notice that together with the increase in the scores, also the differences between precision and recall become higher.

This indicates a bias in the classifier towards over-predicting this class – as in the case of the Jamba translations of the News domain, the references for the literature domain and the mbart translations for the poetry domain, the model predicts more sentences as correct for that category, some of which are simply falsely classified; in contrast, for the news and the poetry domains, the $l1$ classifier has a high precision (but lower recall) in classifying the reference translations, which would indicate more conservative prediction which mostly is correct.

When analysisng the results per domain, for the **news domain** the reference class is clearly dominant at $l1\_ratio$ = 0 and 0.5. At $l1\_ratio$ = 1.0 jamba overtakes with higher recall (0.64) but a low precission (0.19) indicating over-generalisation. Given that overall, the results for the reference translations are of higher f1 and are more balanced and stable, indicating that human translations are more separable in the news and poetry domains, though this effect is weaker for the Literature domain. When it comes to the **poetry domain**, the results are more distinguished, despite the high spike in recall for the reference translations. This strongly supports separability, although, with the high $l1\_ratio$ value the classifier risks over-generalisability. With respect to the **literature domain** the differences are less pronounced and at $l1\_ratio$ of 1.0 we see that the classifier performs better for mBART with the classification of the references ranking second (f1 for mBART = 0.33 and for ref = 0.22). This is an interesting observation on its own as it indicates that mBART translations are more separable than the rest. This is a pattern observable for all domains. Still, for $l1\_ratio$ of 0.0 and 0.5, the scores for the reference translation class are higher. Furthermore, there are smaller differences between precision and recall (as compared to mBART) which indicates a more stable performance of the classifier and supports the hypothesis of separability for the Literature domain. We notice, however, that this separability is less pronounced than for the other domains.

We observe a consistent relationship between lexical and grammatical diversity and separability: domains in which human-generated texts exhibit substantially higher diversity (e.g. Poetry) also show higher classification separability, whereas domains with overlapping diversity profiles (e.g. Literature) exhibit reduced separability.

Clustering The clustering analysis reveals a strong concentration of translations within a single dominant cluster for every domain (C0 for News, C4 for Poetry, C4 for Literature). In addition, both human-generated text and GAIT outputs (for all models) exhibiting highly similar distributions across clusters. These indicate that the clusters capture shared distributional structures rather than model-specific characteristics, and that the data points are not easily separable in this representation space. However, the Jamba, GPT-4o and DeepSeek translations fall in nearly identical cluster distributions across all domains, suggesting a high degree of cross-model convergence. While human-generated texts also fall within the same dominant clusters, the differences are more pronounced and the similarity among GAIT outputs, and especially between Jamba, GPT-4o and DeepSeek is higher. This strengthens the observation from the diversity and classification analyses that these models produce closely aligned outputs.

We also notice that the degree of concentration varies by domain: poetry exhibits the strongest skew, with over 90% of instances assigned to a single cluster. This indicates highly constrained and homogeneous texts for all GAIT and human-generated texts. In contrast, the clusters from the News and Literature domains are slightly more disperse.

While the clustering for all models output and human-generated text show substantial overlap in cluster assignments, GAIT outputs (and especially Jamba, GPT-4o and DeepSeek) are more tightly clustered and less distinguishable from one another than the rest, suggesting that these are more homogenous texts than the other two (mBART and human-generated texts).

6 Conclusions and future work

This work aims to investigate the translation capabilities of four multilingual LLMs – mBART, Jamba-1.5-large, GPT-4o, and DeepSeek – across three domains: News, Poetry, and Literature. We evaluated their performance on an English-to-Dutch translation task using both standard translation quality evaluation metrics metrics (statistical and neural) and measures of lexical and grammatical diversity. In addition, we employed classification and clustering techniques to assess (i) the separability of human-generated translations from model outputs, and (ii) the degree of similarity

among model-generated translations.

The automatic MT evaluation metrics showed that GPT4-o can be considered the best model for English to Dutch translation for all categories.

The results from automatic translation evaluation metrics indicate that GPT-4o performs best across all domains. However, the analyses of lexical / grammatical diversity, classification, and clustering reveal a more nuanced picture. First, they show that human-generated texts are more clearly separable in domains such as News and Poetry, where they also are scored higher in terms of lexical and grammatical diversity metrics. In contrast, this is not the case for the Literature domain – GAIT texts more closely resemble human-generated texts in their distributional properties (e.g. clustering) and even surpass them according to diversity metrics.

Similar to other research, our results highlight a divergence between translation quality and lexical / grammatical diversity, i.e. models that achieve high scores according to both statistical and neural evaluation metrics, do not necessarily follow a similar trend when compared to each other and to the human-generated text according to the lexical and grammatical diversity metrics. This is particularly obvious in stylistically demanding domains such as Literature and Poetry. This augments the insight that current evaluation metrics primarily capture adequacy and fluency, while overlooking stylistic variation and exacerbates the need for more elaborate evaluation strategies.

Comparing the different models, our analysis showed that the GPT-4o, DeepSeek, and Jamba-1.5-large model outputs are consistently less distinguishable from one another than from the old-school mBART. In addition to the classification and the diversity findings, the clustering results show substantial overlap across all systems – similar distributions over clusters for both human- and GAIT-generated texts. These findings suggest that, despite architectural differences (Jamba is Mamba-based, GPT4-o is a based on a dense transformer architecture and DeepSeek is based on Mixture of Expert (MoE) models), there is a strong inter-model overlap, likely reflecting shared training paradigms and data sources. Although our experiments do not answer definitively whether LLMs lead to language homogenization, they offer consistent evidence of increased similarity among LLM outputs compared to human-generated texts. By combining evaluation metrics with diversity, classification, and clustering analyses, we demonstrate a methodological framework for assessing the distributional characteristics of LLM-based translation and their potential impact on language.

Our study has several limitations. First, we use a relatively small dataset that does not originate from a standardized benchmark. Second, we evaluate models under a fixed prompting setup, without exploring variations in temperature or prompt design. While we acknowledge that experimenting with varying prompt and temperature settings facilitate a more complex analysis, we chose this setup deliberately. These settings reflect common use-case conditions, where one would follow specific guidelines according to settings tested, validated and recommended in other work (that specifically targets the assessment of these variables). Third, the Poetry dataset differs in how it was built which may hinter the comparison with respect to the other domains. We note that this asymmetry raises an interesting question which interleaves human and machine translation analysis worth exploring in the future. Fourth, our classification and clustering experiments rely on relatively simple representations (TF-IDF). This could be extended with more robust embedding-based approaches (e.g. contextual embeddings such as BERTje).

Future work will address the above limitations, i.e. using larger and more diverse datasets (also scaling to other languages/language pairs), exploring different prompt and decoding options, utilising more advanced clustering and classification approaches, as well as involving a thorough human evaluation with in-depth error analysis. In addition, task-specific fine-tuning and distillation may provide further insights into the capabilities and limitations of LLM-based translation systems.

References

Alemohammad, Sina, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G Baraniuk. 2023. Self-consuming generative models go mad. *arXiv preprint arXiv:2307.01850*, 4:14.

Aranberri, Nora and Jose A Pascual. 2025. Propagating machine translation traits to predict potential impact on the target language. *Natural Language Processing*, 31(6):1450–1469.

Arenas, Ana Guerberof and Antonio Toral. 2022. Creativity in translation: machine translation as a constraint for literary texts. *CoRR*, abs/2204.05655.

Besacier, Laurent and Lane Schwartz. 2015. Automated translation of a literary work: A pilot study. In Feldman, Anna, Anna Kazantseva, Stan Szpakowicz, and Corina Koolen, editors, *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 114–122, Denver, Colorado, USA, June. Association for Computational Linguistics.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Brglez, Mojca and Špela Vintar. 2022. Lexical diversity in statistical and neural machine translation. *Information*, 13(2).

Castilho, Sheila and Natália Resende. 2022. Post-editese in literary translations. *Information*, 13(2):66.

Chakrabarty, Tuhin, Arkadiy Saakyan, and Smaranda Muresan. 2021. Don't go far off: An empirical study on neural poetry translation. In Moens, Marie-Francine, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7253–7265, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

de Clercq, Orphée, Gert de Sutter, Rudy Loock, Bert Cappelle, and Koen Plevoets. 2021. Uncovering Machine Translationese Using Corpus Analysis Techniques to Distinguish between Original and Machine -Translated French. *Translation Quarterly*, (101):21–45.

Deutsch, Daniel, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. Wmt24++: Expanding the language coverage of wmt24 to 55 languages & dialects. *arXiv preprint arXiv:2502.12404*.

Du, Shuxiang, Ana Guerberof Arenas, Antonio Toral, Kyo Gerrits, and Josep Marco Borillo. 2025. Optimising ChatGPT for creativity in literary translation: A case study from English into Dutch, Chinese, Catalan and Spanish. In Bouillon, Pierrette, Johanna Gerlach, Sabrina Girletti, Lise Volkart, Raphael Rubino, Rico Sennrich, Ana C. Farinha, Marco Gaido, Joke Daems, Dorothy Kenny, Helena Moniz, and Sara Szoc, editors, *Proceedings of Machine Translation Summit XX: Volume 1*, pages 578–591, Geneva, Switzerland, June. European Association for Machine Translation.

Dunder, Ivan, Sanja Seljan, and Marko Pavlovski. 2020. Automatic machine translation of poetry and a low-resource language pair. In Koricic, Marko, Karolj Skala, Zeljka Car, Marina Cicin-Sain, Vlado Sruk, Dejan Skvorc, Slobodan Ribaric, Bojan Jerbic, Stjepan Gros, Boris Vrdoljak, Mladen Mauher, Edvard Tijan, Tihomir Katulic, Predrag Pale, Tihana Galinac Grbac, Nikola Filip Fijan, Adrian Boukalov, Dragan Cisic, and Vera Gradisnik, editors, *43rd International Convention on Information, Communication and Electronic Technology, MIPRO 2020, Opatija, Croatia, September 28 - October 2, 2020*, pages 1034–1039. IEEE.

Fonteyne, Margot, Arda Tezcan, and Lieve Macken. 2020. Literary machine translation under the magnifying glass: Assessing the quality of an NMT-translated detective novel on document level. In Calzolari, Nicoletta, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3790–3798, Marseille, France, May. European Language Resources Association.

Gain, Baban, Dibyanayan Bandyopadhyay, Asif Ekbal, and Trilok Nath Singh. 2026. Bridging the linguistic divide: A survey on leveraging large language models for machine translation.

Geng, Mingmeng, Caixi Chen, Yanru Wu, Yao Wan, Pan Zhou, and Dongping Chen. 2025. The impact of large language models in academia: from writing to speaking. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19303–19319.

Ghazvininejad, Marjan, Yejin Choi, and Kevin Knight. 2018. Neural poetry translation. In Walker, Marilyn A., Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 67–71. Association for Computational Linguistics.

Guerberof-Arenas, Ana and Antonio Toral. 2020. The impact of post-editing and machine translation on creativity and reading experience. *Translation Spaces*, 9(2):255–282.

Guo, Yanzhu, Guokan Shang, and Chloé Clavel. 2025. Benchmarking linguistic diversity of large language models. *Transactions of the Association for Computational Linguistics*, 13:1507–1526, 11.

Hassan, Hany, Anthony Aue, C. Chen, Vishal Chowdhary, J. Clark, C. Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, W. Lewis, M. Li, Shujie Liu, T. Liu, Renqian Luo, Arul Menezes, Tao Qin, F. Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and M. Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *ArXiv*, abs/1803.05567.

Humblé, Philippe. 2020. Machine translation and poetry. the case of english and portuguese. *Ilha do Desterro*, 72:41–56, 08.

Jamba Team, Barak Lenz, Alan Arazi, Amir Bergman, Avshalom Manevich, Barak Peleg, Ben Aviram, Chen Almagor, Clara Fridman, Dan Padnos, Daniel Gissin, Daniel Jannai, Dor Muhlgay, Dor Zimberg, Edden M Gerber, Elad Dolev, Eran Krakovsky, Erez Safahi, Erez Schwartz, Gal Cohen, Gal Shachaf, Haim Rozenblum, Hofit Bata, Ido Blass, Inbal Magar, Itay Dalmedigos, Jhonathan Osin, Julie Fadlon, Maria Rozman, Matan Danos, Michael Gokhman, Mor Zusman, Naama Gidron, Nir Ratner, Noam Gat, Noam Rozen, Oded Fried, Ohad Leshno, Omer Antverg, Omri Abend, Opher Lieber, Or Dagan, Orit Cohavi, Raz Alon, Ro'i Belson, Roi Cohen, Rom Gilad, Roman Glozman, Shahar Lev, Shaked Meirom, Tal Delbari, Tal Ness, Tomer Asida, Tom Ben Gal, Tom Braude, Uriya Pumerantz, Yehoshua Cohen, Yonatan Belinkov, Yuval Globerson, Yuval Peleg Levy, and Yoav Shoham. 2024. Jamba-1.5: Hybrid transformer-mamba models at scale.

Jiang, Liwei, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. 2025. Artificial hivemind: The open-ended homogeneity of language models (and beyond). *arXiv preprint arXiv:2510.22954*.

Karpinska, Marzena and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In Koehn, Philipp, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore, December. Association for Computational Linguistics.

Kobak, Dmitry, Rita González-Márquez, Emőke-Ágnes Horvát, and Jan Lause. 2024. Delving into chatgpt usage in academic writing through excess vocabulary. *arXiv preprint arXiv:2406.07016*.

Kocmi, Tom, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA, November. Association for Computational Linguistics.

Kocmi, Tom, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinthór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China, November. Association for Computational Linguistics.

Liang, Weixin, Yaohui Zhang, Mihai Codreanu, Jiayu Wang, Hancheng Cao, and James Zou. 2025. The widespread adoption of large language model-assisted writing across society. *Patterns*, 6(12).

Luo, Jiaming, Colin Cherry, and George Foster. 2024. To diverge or not to diverge: A morphosyntactic perspective on machine translation vs human translation. *Transactions of the Association for Computational Linguistics*, 12:355–371.

Macken, Lieve, Orphée De Clercq, and Hans Paulussen. 2011. Dutch parallel corpus: a balanced copyright-cleared parallel corpus. *META*, 56(2):374–390.

Macken, Lieve. 2024. Machine translation meets large language models: Evaluating ChatGPT's ability to automatically post-edit literary texts. In Vanroy, Bram, Marie-Aude Lefer, Lieve Macken, and Paola Ruffo, editors, *Proceedings of the 1st Workshop on Creative-text Translation and Technology*, pages 65–81, Sheffield, United Kingdom, June. European Association for Machine Translation.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL 2002)*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Popović, Maja. 2015. chrF: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT@EMNLP 2015)*, pages 392–395, Lisbon, Portugal.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on*

Machine Translation: Research Papers, pages 186–191, Belgium, Brussels, October. Association for Computational Linguistics.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation.

Schneider, Johannes, Christian Meske, and Pauline Kuss. 2024. Foundation models: a new paradigm for artificial intelligence. *Business & Information Systems Engineering*, pages 1–11.

Seljan, Sanja, Ivan Dunder, and Marko Pavlovski. 2020. Human quality evaluation of machine-translated poetry. In Koricic, Marko, Karolj Skala, Zeljka Car, Marina Cicin-Sain, Vlado Sruk, Dejan Skvorc, Slobodan Ribaric, Bojan Jerbic, Stjepan Gros, Boris Vrdoljak, Mladen Mauher, Edvard Tijan, Tihomir Katulic, Predrag Pale, Tihana Galinac Grbac, Nikola Filip Fijan, Adrian Boukalov, Dragan Cisic, and Vera Gradisnik, editors, *43rd International Convention on Information, Communication and Electronic Technology, MIPRO 2020, Opatija, Croatia, September 28 - October 2, 2020*, pages 1040–1045. IEEE.

Sellam, Thibault, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, July. Association for Computational Linguistics.

Shumailov, Ilia, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross J. Anderson. 2023. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2307.07003*.

Símonarson, Haukur Barri, Vésteinn Snæbjarnarson, Pétur Orri Ragnarson, Haukur Jónsson, and Vilhjalmur Thorsteinsson. 2021. Miðeind's WMT 2021 submission. In Barrault, Loic, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 136–139, Online, November. Association for Computational Linguistics.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation of the Americas (AMTA 2006). Visions for the Future of Machine Translation*, pages 223–231, Cambridge, Massachusetts, USA.

Sourati, Zhivar, Farzan Karimi-Malekabadi, Meltem Ozcan, Colin McDaniel, Alireza Ziabari, Jackson Trager, Ala Tak, Meng Chen, Fred Morstatter, and Morteza Dehghani. 2025. The shrinking landscape of linguistic diversity in the age of large language models. *arXiv preprint arXiv:2502.11266*.

Thompson, Brian, 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.

Toral, Antonio and Andy Way. 2015. Machine-assisted translation of literary text: A case study. *Translation Spaces*, 4:240–267, 12.

Toral, Antonio and Andy Way. 2018. What level of quality can neural machine translation attain on literary text? *CoRR*, abs/1801.04962.

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, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 113–123. Association for Computational Linguistics.

Toral, Antonio. 2019. Post-editese: an exacerbated translationese. In Forcada, Mikel, Andy Way, Barry Haddow, and Rico Sennrich, editors, *Proceedings of Machine Translation Summit XVII: Research Track*, pages 273–281, Dublin, Ireland, August. European Association for Machine Translation.

Vanmassenhove, Eva, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 222–232, Dublin, Ireland, August. European Association for Machine Translation.

Vanmassenhove, Eva, 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, April. Association for Computational Linguistics.

Vanmassenhove, Eva. 2025. Losing our tail–again: On (un) natural selection and multilingual large language models. *arXiv preprint arXiv:2507.03933*.

Vujović, Dušan. 2024. Generative ai: Riding the new general purpose technology storm. *Ekonomika preduzeća*, 72(1-2):125–136.

Wang, Huashu and Ying Wu. 2026. Technologies are reshaping translation. *AI Through LLMs: Transforming Translation Practice and Teaching*, page 3.

Wang, Shanshan, Derek Wong, Jingming Yao, and Lidia Chao. 2024. What is the best way for ChatGPT to translate poetry? In Ku, Lun-Wei, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14025–14043, Bangkok, Thailand, August. Association for Computational Linguistics.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Wu, Yangjian and Gang Hu. 2023. Exploring prompt engineering with GPT language models for document-level machine translation: Insights and findings. In Koehn, Philipp, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 166–169, Singapore, December. Association for Computational Linguistics.

Yakura, Hiromu, Ezequiel Lopez-Lopez, Levin Brinkmann, Ignacio Serna, Prateek Gupta, Ivan Soraperra, and Iyad Rahwan. 2024. Empirical evidence of large language model's influence on human spoken communication. *arXiv preprint arXiv:2409.01754*.

Zhang, Ran, Wei Zhao, and Steffen Eger. 2025. How good are LLMs for literary translation, really? literary translation evaluation with humans and LLMs. In Chiruzzo, Luis, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10961–10988, Albuquerque, New Mexico, April. Association for Computational Linguistics.

A Post-processing of translations

The following regex matches were used to clean the translated sentences and leave only the translated, Dutch, output:

1. `([^=]+)=`
2. `=$`
3. As well as looking for leading `English:` , `Dutch:` ,`Nederlands:` or `Nederlandse:`

We also corrected utf encoding errors by replacing the following non-ascii characters with their utf-8 versions:

Generated token / sequence	Corrected token / sequence
Ã©, Â©, Ã³	é
Ã¯, iÌ^	ï
Ã´	ô
Ã§	ç
Ã«, eÌ^, Ì€	ë
Ã¨	è
Ã¼	ü
Ã¤	a
Ã·	á
Ã¶	ö
eÌ□eÌ□n	één
vÃƒÂ³ÃƒÂ³r	véér

Table 4: Non-ASCII characters replacements