Towards Tailored Recovery of Lexical Diversity in Literary Machine Translation

Esther Ploeger❄ **Huiyuan Lai**☆ **Rik van Noord**☆ **Antonio Toral**☆
❄Department of Computer Science, Aalborg University, Denmark
☆CLCG, University of Groningen, The Netherlands
espl@cs.aau.dk {h.lai,r.i.k.van.noord,a.toral.ruiz}@rug.nl

Abstract

Machine translations are found to be lexically poorer than human translations. The loss of lexical diversity through MT poses an issue in the automatic translation of literature, where it matters not only *what* is written, but also *how* it is written. Current methods for increasing lexical diversity in MT are rigid. Yet, as we demonstrate, the degree of lexical diversity can vary considerably across different novels. Thus, rather than aiming for the rigid *increase* of lexical diversity, we reframe the task as *recovering* what is lost in the machine translation process. We propose a novel approach that consists of reranking translation candidates with a classifier that distinguishes between original and translated text. We evaluate our approach on 31 English-to-Dutch book translations, and find that, for certain books, our approach retrieves lexical diversity scores that are close to human translation.

1 Introduction

With the introduction of neural machine translation (NMT), the performance of high-resource automatic translation has improved substantially. Especially since the introduction of the Transformer architecture (Vaswani et al., 2017), state-of-the-art NMT systems have outperformed previous approaches considerably (Lakew et al., 2018), with some works even claiming human parity (Popel et al., 2020). However, these claims are based mostly on accuracy and fluency measures, while style is often overlooked. In fact, according to expert evaluation, machine translation (MT) did actually not reach human parity (Toral et al., 2018; Fischer and Läubli, 2020). For instance, MT models have been found to exacerbate linguistic patterns that occur frequently, while underrepresenting patterns that are found less commonly (Vanmassenhove et al., 2019). As a result, automatically translated texts are found to be lexically poorer than human translations (HT). This 'artificially impoverished language' has previously been referred to as *machine translationese* (Vanmassenhove et al., 2021).

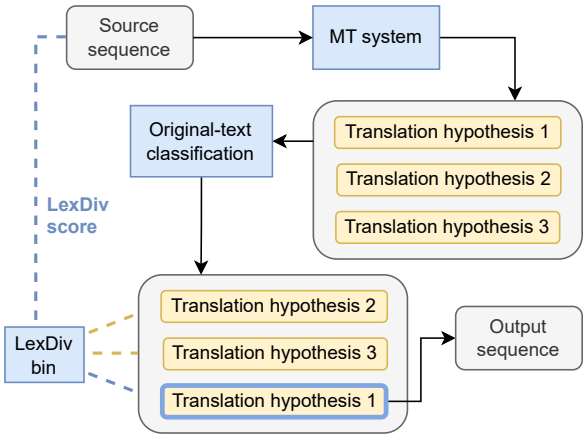

Figure 1: Reranking translation hypotheses based on the probability they are originally written in the target language, where the chosen rank is based on the lexical diversity score of the original book, and could be lower than the most lexically diverse option.

In this paper, we focus on the translation of novels. Contrary to technical domains, where meaning preservation is the main criterion for acceptable translations, literary translations have the additional criterion of style. This is because apart from meaning preservation (*what* is written), maintaining a certain reading experience (*how* it is writ-

Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1), pages 286–299
June 24-27, 2024 ©2024 European Association for Machine Translation

ten) is vital for novels (Toral and Way, 2015). Importantly however, writing style (and linguistic complexity) can vary considerably between books. Some books contain repetitive language use, while others are written in embellished language (see Section 3). Current approaches that aim to mitigate the loss of lexical diversity do not accommodate this. State-of-the-art previous work (Freitag et al., 2019; Freitag et al., 2022) increases lexical diversity in a rigid way, not allowing for flexibility at inference time.

Contributions *(i)* We show that lexical diversity varies considerably across books, and argue that this should be taken into account in MT; *(ii)* We introduce a novel flexible method for recovering lexical diversity in MT, informed by the diversity of the original. *(ii)* We evaluate our method on 31 English novels which are translated to Dutch, and find that our approach is effective when it comes to book-tailored promotion of lexical diversity.

2 Related Work

Literary MT NMT has been argued to hold potential for literary texts, for instance in assisting professional translators or improving the immediate accessibility of untranslated foreign language books (Matusov, 2019). However, MT has been shown to decrease lexical diversity (Vanmassenhove et al., 2019; Vanmassenhove et al., 2021). This is an issue, because literary works can be viewed as a special domain in translation. Typically, literary translators are expected to preserve not only literal elements from the source, such as the plot, but also some sense of creative value (Riera, 2022). In other words, a goal of literary translation could be to recreate the 'aesthetic intentions or effects' that are possibly present in the source book (Delabastita, 2011). Such 'aesthetic intentions' can for instance be voice and metaphor, but also repetition (Wright, 2016). Repetitive use of language is commonly a conscious choice by the writer, and has a function, such as drawing attention or establishing a pattern (Boase-Beier, 2011). Given that lexical diversity can be an intentional writing choice, it should be apparent that an approach that aims at recovering lexical diversity in MT should not be boundless. Therefore, it is our aim to inform recovery with the degree of relative lexical diversity of the source text.

Machine *Translationese* Following recommendations from Jiménez-Crespo (2023), we will largely refrain from using the term *translationese* in the rest of this paper. However, it is important to note that previous work that aims to increase lexical diversity in MT has mostly been framed as part of 'machine translationese' reduction (Freitag et al., 2019; Freitag et al., 2022; Dutta Chowdhury et al., 2022; Jalota et al., 2023). Translations have been found to differ from original texts in a number of ways. For one, Baker (1993) argues that human translations into a language tend to be lexically simpler than text originally written in that language. Automatic classification approaches have been effective in detecting this difference (Baroni and Bernardini, 2005; Koppel and Ordan, 2011; Volansky et al., 2015; Rabinovich and Wintner, 2015; Pylypenko et al., 2021). More recently, work has investigated linguistic differences between MT and HT (van der Werff et al., 2022). Thus, it seems that modelling characteristics of original versus translated texts has a direct link to lexical diversity. Previous work (Freitag et al., 2022) leveraged these detectable differences in their approach to increase the naturalness of output translations. We take inspiration from their lexical diversity evaluation methods, and implement their method as a baseline.

Reranking Methods Reranking hypotheses in text generation originated before the age of neural paradigms (Shen et al., 2004; Collins and Koo, 2005). In essence, reranking entails re-ordering the set of candidate outputs according to some criterion, with the aim of providing a final output that adheres better to that criterion. Such methods have been applied for various tasks, such as summarization (Liu and Liu, 2021) and semantic parsing (Yin and Neubig, 2019). In machine translation, previous approaches include discriminative reranking (Lee et al., 2021) and reranking with energy-based models (Arcadinho et al., 2022).

3 Why Recover Rather Than Increase Lexical Diversity?

In this paper, we argue for tailored recovery of lexical diversity. In this section, we first discuss support for this idea from the field of literary studies. Then, we provide empirical evidence by applying lexical diversity metrics to our test set.

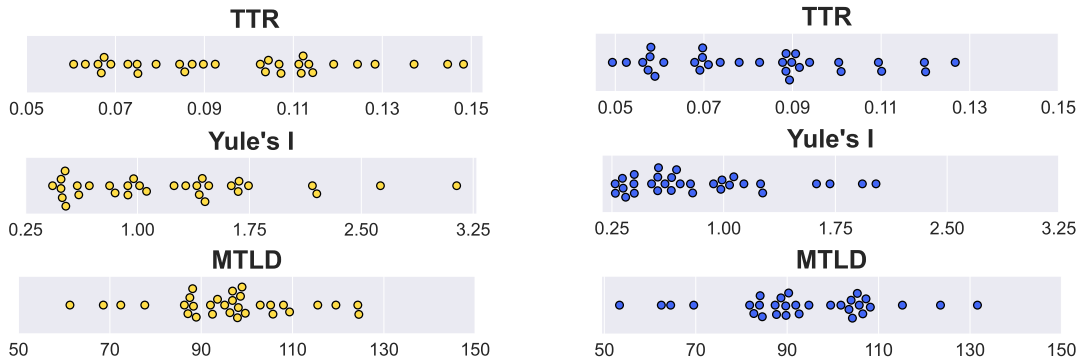

Figure 2: Range and spread of lexical diversity metrics for HT (left, yellow) and original English (right, blue).

3.1 Theoretical Support

Previous work on writing style in novels acknowledges that some books exhibit more lexical diversity than others. As an example, Heaton (1970) finds that no word in the original (i.e. English) version of in *The Old man and the Sea* by Ernest Hemingway contains more than six syllables. Additionally, Hemingway tends to stick to particular words, even when there are more diverse options: in 184 situations of direct speech, he chooses to use the word 'said' 170 times instead of for example 'asked', 'remarked', 'noticed' or 'yelled'. An example from the other end of the spectrum is James Joyce's *Ulysses*. This work is known for its experimental techniques and unorthodox language use. Trotta (2014) illustrates this by highlighting Joyce's use of neologisms, such as 'He *smellsipped* the cordial juice' and 'Davy Byrne *smiledyawnednodded* all in one'. Moreover, Joyce repeatedly uses non-verbs as verbs, like in 'I am *almosting* it.' and even writes long sequences in unconventional spelling (*Ahbeesee defeegee kelomen opeecue rustyouvee doubleyou*. Boys are they?'). These examples make it clear that books can be written with vastly different 'aesthetic intentions'. Thus, for preserving these intentions, MT approaches should not render them equally diverse in terms of lexicon.

3.2 Empirical Support

We empirically verify whether these findings hold for our data specifically, by estimating the lexical diversity of the 31 books in our test set, which we introduce in Section 5.1. We calculate three measures of lexical variety (type-token ratio; TTR, Yule's I (Yule, 1944), and MTLD (McCarthy, 2005)) for each book in our test set. We further elaborate on these metrics in Section 6. Next, we apply the same metrics to the human translations of those same books. Figure 2 shows that there is indeed a wide range of diversity across books, for both HT and original text. For example, in both settings, we find that the highest MTLD value is almost two times as large as the lowest. This emphasises why it is not our aim to generate the highest possible lexical diversity for every book. While we observe similar ranges and distributions in HT vs. original, the HT metrics are slightly higher. However, this does not necessarily mean that HT contains more embellished language. We note that the languages in our study, Dutch and English, are relatively similar (both in terms of genealogy and linguistic typology), but they differ in ways that can influence diversity metrics. For instance, Dutch contains compound nouns while English does not, making a higher TTR for Dutch more likely.

This discrepancy means that we cannot compare our Dutch MT to the original English book diversity directly. Instead, here we compare MT with HT. To verify whether this is sensible, we assess the relationship between HT and the English originals, by computing Pearson's correlation on the corresponding diversity metrics. The results are listed in Table 1, and the corresponding regression plots are found in Appendix B. We observe strong correlations that are all statistically significant. This is important, because as the source diversity is a reliable indicator of HT diversity, it makes sense to use the source scores to approach HT (see Section 4).

Metric	Correlation coefficient	p-value
TTR	0.971	< 0.00001
Yule's I	0.929	< 0.00001
MTLD	0.953	< 0.00001

Table 1: Pearson correlation coefficients for HT and OR lexdiv metrics, rounded to three decimals.

4 Reranking Method

As illustrated in Figure 1, our approach consists of two parts: hypothesis generation and hypothesis reranking. Firstly, we generate the *n* best translation candidates for each source sentence in the test set with a vanilla domain-specific MT system (Section 5.1). Note that we decode all books separately, instead of concatenating all test set books. Then, for each book, we apply a classifier (Section 5.2) to the translation hypotheses and, through a softmax layer, obtain the probability for each candidate that it is an original Dutch sequence. Based on these probabilities, we rerank the translation candidates. In order to obtain the (expected) most lexically rich candidate, we would then choose the rank with the highest original-text probability. However, note that this simple approach is flexible in the sense that, instead of choosing the most original-like option, we have the option to choose a lower original-text rank.

We leverage this flexibility for tailoring rank selection to the lexical diversity of the original English book. First, for each original book, we calculate a *LexDiv* score, which consists of the average of the normalized TTR, Yule's I and MTLD scores (see Section 6). Then, we bin the books according to their *LexDiv* score, relative to the total distribution. That is, given a list that is sorted based on *LexDiv*, we categorize these into groups, where the number of groups depends on the number of *nbest* candidates in decoding. For example, for $n = 5$, we bin the books into 5 different groups of 6 books (adding any remainders into the last bin). The bin per book corresponds to the original-text rank that is selected. As such, the selected rank for each book depends on the lexical diversity of its source, relative to the other books. Reranking translation candidates is a suitable solution to our task, because it accomodates flexibility, which is tunable at inference time. There is no need to train a separate model per diversity setting, saving computational expenses. Additionally, our approach is model-agnostic: reranking can be applied to any MT model that can generate multiple translation candidates.

5 Experimental set-up

5.1 Vanilla MT System

Data We use the dataset by Toral et al. (2024), which contains 531 books that were originally written in English and manually translated into Dutch. We use 495 books for training, 5 for development and 31 as a test set. The genres of the books vary: they include literary fiction, popular fiction, non-fiction and children's books from over 100 authors. We do not make a distinction between literary and 'unliterary' novels, as we believe this to be a subjective judgment.[1]

Training Firstly, we align the sentences of the English and Dutch versions of each book using Vecalign (Thompson and Koehn, 2019). For the books in the test set, we manually discard sentences for which there existed no proper alignment, such as front matter sentences. Additionally, we discard sentences with a cosine distance higher than 0.7 (2.3% of all sentences). Then, we normalise all punctuation using the MOSES toolkit.[2] We then apply SentencePiece (Kudo and Richardson, 2018) subword segmentation to the data. For this, we train a SentencePiece unigram model with a joint vocabulary for both languages and a vocabulary size of 32,000.

We train a Transformer-based translation model using the Fairseq toolkit (Ott et al., 2019). More specifically, we use the *transformer_iwslt_de_en* architecture. This is a Transformer base model with 6 encoder and decoder layers and an embedding dimension of 512. During training, we use an Adam optimiser, a learning rate of 5e-4, the loss function cross entropy with label smoothing 0.1 and the batch size is 64. Each model is trained until convergence with a patience of 3 epochs, using the BLEU score as a maximisation metric for finding the best checkpoint.

Decoding Strategies By default, we use beam search for decoding. Reranking approaches rely heavily on the diversity of the translation hypotheses: if the hypotheses are all very similar, reranking them is not likely to have a large effect. To ensure diverse hypotheses, we use a beam size of 20. Additionally, we experiment with decoding through diverse beam search (Vijayakumar et al., 2016). We follow Vijayakumar et al. (2016) by using 3 groups, with a beam size of 21. Beyond beam search, we investigate the effects of top-k and top-p sampling, with the default parameters and sampling size 10.

[1] A full list of author names, titles, genres and publishing years of the test set books can be found in Appendix A, Table 8.
[2] http://www.statmt.org/moses/

System development (90%)				
Split	**Orig.**	**# Books**	**# Sentences**	**# Words**
Train (80%)	Dutch	1,291	8,576,756	10,425,656
	Other	1,291	12,470,149	165,263,466
Dev (10%)	Dutch	162	1,005,832	12,533,406
	Other	162	1,546,057	19,723,706
Test (10%)	Dutch	162	1,189,690	14,721,914
	Other	162	1,573,499	20,968,346
Original-text Classification (10%)				
Split	**Orig.**	**# Books**	**# Sentences**	**# Words**
Train (80%)	Dutch	143	982,114	11,528,789
	Other	143	139,0351	17,951,613
Test (20%)	Dutch	36	261,151	2,974,873
	Other	36	340,950	4,283,604
Total		3,588	29,336,549	376,130,733

Table 2: Monolingual data set division and size.

5.2 Original-Text Classification

Data We use a monolingual dataset of more than 7,000 Dutch books from varying original languages, authors and genres (Toral et al., 2024). For each book, we annotate whether it was originally written in Dutch.[3] We discard 2,182 books for which the original language is unclear or that were not prose. We make sure to avoid overlap with the parallel data set by removing any books that are also part of the parallel data. Finally, we randomly sample 1,794 of the remaining 2,190 books as to match the total number of translated books, ensuring an equal distribution. In total, we are left with over 3,500 books and over 29M sentences. We further divide these into data for system development and data for original-text classification. We use this data for reproducing previous work (Freitag et al., 2022) and for training our classifier. Additionally, we translate the classifier section of the monolingual data set using a reverse-direction trained version of the vanilla MT system (NL $\rightarrow$ EN), and then perform round-trip-translation (RTT) back to Dutch with the vanilla MT system, to obtain an MT version of the monolingual classifier data. The full data size statistics and division in training, development and testing splits are listed in Table 2.

Training Currently, state-of-the-art performance for original-text detection is based on BERT (Devlin et al., 2019), as demonstrated by Pylypenko et al. (2021). We implement a similar system that distinguishes between original text and MT by training a binary classification model. We fine-tune Dutch language model BERTje (de Vries et al., 2019). We train each model on the training split of the original-text classification data (see Table 2). We train models with batch size 128, accumulating gradients over 8 update steps, using the Adam optimiser (Kingma and Ba, 2015) with a learning rate of 3e-5. We use early stopping (patience 3) if validation performance does not improve. On the held-out test set, the classifier achieves an accuracy of 85.9%. It obtains a precision of 90.6%, a recall of 80.2% and the F1 score is 85.0%.

5.3 Baselines

APE Freitag et al. (2019) introduced Automatic Post-Editing (APE) as a post-hoc method to increase the 'naturalness' of MT output. Following their approach, we train a post-processor that 'translates' synthetic Dutch sequences into more natural Dutch sequences. For training this system, we use the same data that was used to train the classifier (Section 5.2), consisting of RTT Dutch (which we use as source) and original Dutch (which we use as target). We train a model with the same architecture as the vanilla MT system. We apply the post-processor to the output of the vanilla MT system, in an attempt to obtain a translation with a lexical diversity that is closer to HT.

Tagging Our second baseline is based on Freitag et al. (2022). We train an MT system that learns to differentiate between original and translated text during training. This method requires both translated and original Dutch target samples. The translated target samples are found in our parallel dataset. We use the same original Dutch samples that are used in training the translationese classifier. Following Freitag et al. (2022), we then prepend $<orig>$ to the English source sentences that have original Dutch on the target side, and $<trans>$ for the source sentences that have translated Dutch. We train an MT system (same parameters as vanilla MT) on this data set. For inference, we prepend the source with $<orig>$, which prompts the model to produce a translation that exhibits characteristics that are often found in original Dutch. Note that, in contrast to APE, this method cannot be applied post-hoc.[4]

[3]The full annotation workflow can be found in Appendix C

[4]Note that our implementation differs from Freitag et al. (2022) in that they automatically differentiate natural and unnatural samples from a large parallel corpus using contrasting language models.

6 Evaluation

We introduce three classes of metrics. Firstly, we look at general text metrics, which are commonly used for evaluating lexical diversity. Secondly, we use translation-specific metrics. Lastly, we evaluate the general translation quality.

6.1 General Text Metrics

TTR The type-token ratio is the ratio of types (set of words) to tokens (actual words). A higher TTR indicates that more (different) words are used, which in turn indicates a higher lexical diversity. While this method is known to be influenced by the length of the text it is applied to, we report it because it is easy to interpret and widely used.

Yule's I As a metric that is less sensitive to variation in text length, we use Yule's I (Yule, 1944). We calculate this value as stated in Equation 1, where V is the size of the vocabulary (number of types) and *t(i,N)* denotes the frequency of types which occur i times in a sample of length N.

$$\text{Yule's I} = \frac{V^2}{\sum_{i=1}^{V} \times t(i,N) - V} \quad (1)$$

MTLD As an additional metric that has proven to be robust to document length variety, we use the measure of textual lexical diversity (MTLD), which is sequentially calculated as the 'average length of sequential word strings in a text that maintain a given TTR value' (McCarthy, 2005). We use the same TTR threshold (0.72) as Vanmassenhove et al. (2021).

We calculate these values using the *LexicalRichness* Python library (Shen, 2022).

6.2 Translation-specific Metrics

Vanmassenhove et al. (2021) introduce a novel automatic evaluation method for measuring lexical diversity in translations: Synonym Frequency Analysis (SFA). It provides an insight into the diversity of lexical choices in translations. For English words that have multiple translations in Dutch, it takes into account the frequency of these translation options. We re-implement this method, as it was not implemented for our language pair before. We first lemmatise each word in the source (English) side of our test set, using SpaCy (*nl_core_news_lg*).[5] Next, we extract all possible translation options for the English adjectives, nouns and verbs by using a English-to-Dutch bilingual dictionary.[6] Next, for each translation option, we count the number of occurrences in the MT output for each system. The result is a vector which contains the occurrence frequency of each translation synonym for an English word.

PTF The primary translation frequency (PTF) is the average percentage (over all relevant source words) of times the most frequent translation option was chosen, from all translation options. The assumption is that if the output contains more secondary candidates, the text is more lexically diverse. We report the average PTF of all source words.

CDU The CDU is the cosine distance between the output vector for each source word and a vector of the same length with an equal distribution for each translation option (with the same total). We take the average CDU over all relevant source words to compute a final CDU.

SynTTR Lastly, we compute the SynTTR by dividing the number of types (the length of the set of all translation options) by the number of tokens (the sum of all translation options vectors).

6.2.1 Translation Quality

We also calculate a general measure of translation quality, because the 'naturalness' of a translation does not necessarily imply that a translation is a faithful representation of the source. A randomly generated string sequence might be very lexically diverse, but likely does not carry the source meaning. Firstly, we calculate BLEU (Papineni et al., 2002), as implemented in SacreBLEU (Post, 2018). We use the default settings, which are case-sensitive. Secondly, to account for the fact that BLEU does not necessarily evaluate meaning preservation, we additionally evaluate with COMET (Rei et al., 2020). English and Dutch are relatively high-resource languages, so we can use multilingual language embeddings. We report *comet-score*, calculated with the default *wmt22-comet-da*. Still, it should be noted that these automatic metrics do not necessarily correlate strongly with human judgements, especially for literary translation.

[5] https://spacy.io/models/nl#nl_core_news_lg

[6] We use the dictionary from `https://freedict.org/downloads`. As an example, for the English adjective *touching*, we find as Dutch translations: *ontroerend*, *aangrijpend*, *emotioneel*, *treffend*, *roerend* and *aandoenlijk*.

Approach	TTR ↑	Yule's I ↑	MTLD ↑	PTF ↓	CDU ↓	SynTTR ↑	BLEU ↑	COMET ↑
HT	0.098	1.226	96.05	0.817	0.549	0.042	-	-
Vanilla MT	0.089	0.951	90.21	0.832	**0.550**	0.040	32.32	0.824
APE	0.092	0.985	90.59	0.827	0.554	**0.041**	30.39	0.808
Tagging	**0.095**	1.111	**94.08**	0.829	**0.550**	**0.041**	31.33	0.807
Tailored RR *(n=5)*	0.091	1.002	92.46	0.829	0.552	**0.041**	30.92	0.815
Tailored RR *(n=10)*	0.091	1.013	93.26	0.829	0.547	**0.041**	30.07	0.810
Tailored RR *(n=20)*	0.092	1.010	93.27	0.830	0.558	**0.041**	28.98	0.802
Tailored RR *(Top-k)*	**0.101**	**1.286**	104.25	**0.815**	0.559	**0.043**	21.21	0.745
Tailored RR *(Top-p)*	0.092	1.017	91.21	0.828	0.552	**0.041**	29.97	0.808
Tailored RR *(DBS)*	0.092	1.010	92.70	0.828	0.553	0.040	29.36	0.805

Table 3: Scores averaged across books, where RR stands for reranking. We provide results for multiple decoding strategies. Beam size is 20. Scores closest to HT are in bold font.

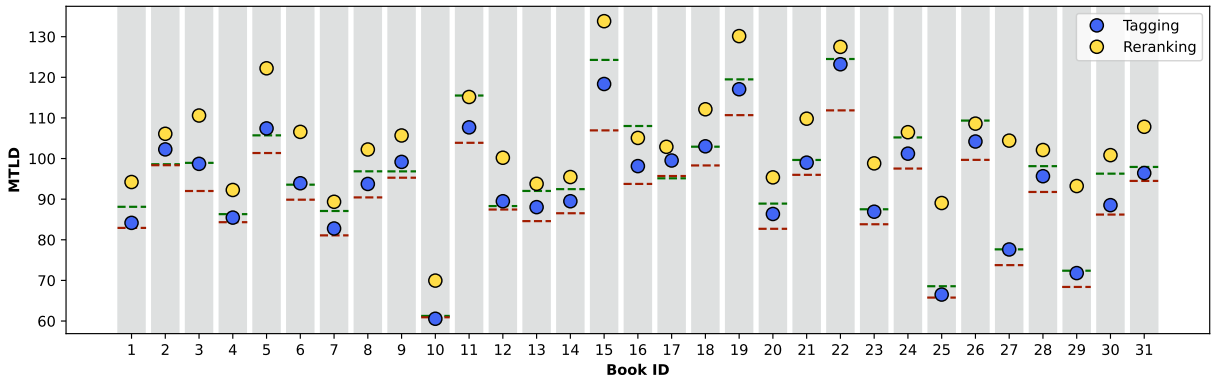

Figure 3: Per-book comparison of MTLD between the (rigid) tagging baseline and (tailored) reranking method, where green dotted lines are HT scores, and red dotted lines represent vanilla MT.

7 Results and Analysis

7.1 Quantitative Results

We first discuss the results over all books. Table 3 shows the average results of measuring lexical diversity and general translation quality across the various approaches. We find that vanilla MT indeed produces lexically poorer translations than HT, according to all our metrics. While the scores of the APE baseline remain close to vanilla MT, our tailored reranking approach retrieves a lexical diversity that is closer to HT. This suggests that our method is a suitable alternative for post-hoc editing, given that one has access to the MT model for generating translation hypotheses. The tagging baseline, which cannot be applied post-hoc, retrieves and MTLD and CDU that is on average closest to HT. Importantly though, it should be noted that reranking and tagging are not mutually exclusive: one could apply reranking to the tagging baseline to increase or decrease lexical diversity further, where desired. When we compare decoding strategies of the tailored reranking method, we first observe that using diverse beams search and choosing a larger n retrieves at most slightly more diversity. Especially top-k decoding retrieves a much higher lexical diversity. However, tailored reranking comes with a compromise in terms of translation quality metrics.

Next, we demonstrate that these averages omit a more fine-grained view. Figure 3 shows the difference in MTLD per book between vanilla MT, HT, tagging and our most diverse reranking system, based on top-k sampling, which is tailored to the LexDiv score of the original English book.[7] Our method renders almost every single book more lexically diverse than the tagging baseline. In some cases, this makes the results closer to HT in terms of lexical diversity (e.g. 7, 13, 14, 16). However, especially in cases where vanilla MT and HT are close already, this is not always true (e.g. 1, 3, 5).

[7] A similar figure with the posthoc baseline APE instead of tagging is shown in Appendix D.

Ex. #	Approach	Text
1	Source	The kid had no mother.
	HT	Dat joch heeft geen moeder gehad.
	Vanilla MT	Het kind had geen moeder.
	Tagging	De jongen had geen moeder.
	Tailored RR	Het joch had geen moeder.
2	Source	He shipped his oars and brought a small line from under the bow.
	HT	Hij haalde de riemen in en pakte een kleine lijn die voor in de boot lag.
	Vanilla MT	Hij trok zijn riemen aan en haalde een klein lijntje onder de boeg vandaan.
	Tagging	Hij verscheurde zijn riemen en haalde een klein streepje onder de boeg vandaan.
	Tailored RR	Hij haalde zijn riemen en trok er een kleine lijn voor onder de boot vandaan.
3	Source	In long shaky strokes Sargent copied the data.
	HT	In lange beverige halen kopieerde Sargent de gegevenheden.
	Vanilla MT	Met lange, bevende slagen kopieerde Sargent de gegevens.
	Tagging	Met lange bevende halen kopieerde Sargent de gegevens.
	Tailored RR	Met lange beverige halen schreef Sargent de data over.

Table 4: Examples to highlight surface-level differences between the systems' output translations, where Tailored RR uses top-k sampling.

7.2 Surface-level Inspection

The output translations were inspected by a native speaker. Table 4 shows three examples of how translations differ between vanilla MT, tagging and tailored reranking (with top-k sampling). In Example 1 (from book 1, *Sunset Park*), we see that the English noun 'kid' is translated as *joch* ('boy') in the human translation, which is less common than the vanilla MT's *kind* ('child') and tagging's *jongen* ('boy'). This is recovered by our tailored reranking system, which uses *joch* too.

Example 2 is taken from book 10, *The Old Man and the Sea*, which has low lexical diversity by default (see Section 3). This is not taken into account by the tagging baseline: the English 'shipped' is translated as a less common (and wrong) *verscheurde* ('shredded'). The tailored reranking system (*haalde*, 'brought') is closest to HT (*haalde in*, 'brought in'). Additionally, the tagging baseline wrongly translates the English 'line' as *streepje* ('small stripe'), while tailored reranking (*lijn*, 'line') is again identical to HT. This case illustrates that choosing a more common translation synonym, which may for instance results in a lower PTF, may for some books be closer to HT.

By contrast, in Example 3 from the more lexically diverse *Ulysses* (book 15), the tagging baseline stays closer to vanilla MT: both translate 'shaky' as *bevend* ('trembling'). Tailored reranking outputs *beverig* ('shaky'), which is again recovering the HT. Furthermore, tailored reranking deviates from all other systems (and HT) by translating 'copied' into the translation synonym *schreef over* (copying something by writing). This case may illustrate why the tailored reranking based on top-k sampling surpasses the other systems in the overall metrics.

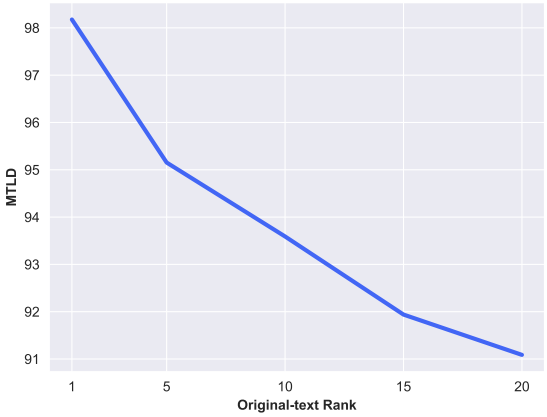

Figure 4: Change in MTLD for choosing different ranks, where beam size is 20 and $n = 20$.

7.3 Ranks and Lexical Diversity

So far, we have assumed that reranking based on the probability of a candidate being original text leads to more lexically diverse output translations.

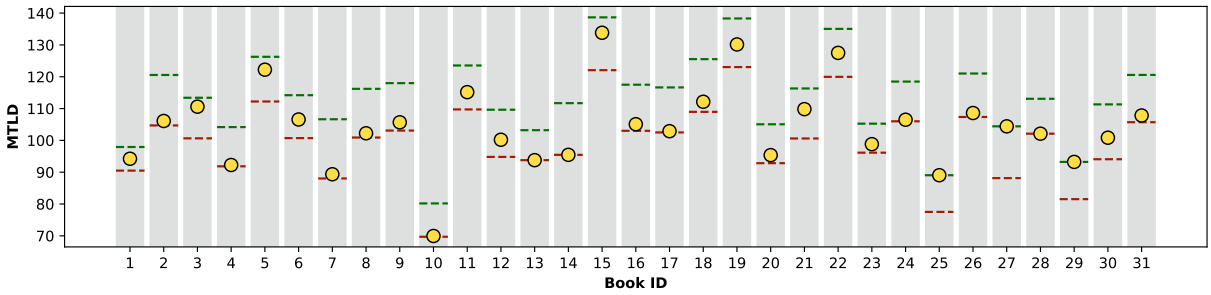

Figure 5: MTLD for highest (green), lowest (red) and tailored (yellow) original-text rank.

Here, we verify whether choosing a lower probability of a candidate being original, actually implies lexically poorer output translations (Figure 4). For the vanilla MT system with beam size 20 and $n = 20$, we first calculate the original-text probability for each translation hypothesis. Similar to reranking, we sort the hypotheses according to this probability. Then, instead of binning, we choose the n^{th} rank, and calculate lexical diversity of the output. Figure 4 shows the change in MTLD scores for choosing a lower diversity rank. We observe that indeed, choosing a lower rank retrieves lower diversity (note that there, a higher rank represents a smaller original-text probability). This trend holds for TTR and Yule's I as well (see Appendix E).

7.4 Tailoring and Lexical Diversity

To further demonstrate the effect of a *tailored* approach in lexical diversity, we compare MTLD scores of a top-k reranking system that always outputs the highest original-text probability, with the same system that always outputs the lowest, and a tailored version. Figure 5 shows the results. Firstly, we observe that, in every case, choosing a rank that represents lower original-text probability retrieves a lower MTLD score than choosing the opposite. This corroborates the findings from the previous section. Next, we look into how the tailored reranking affects the output lexical diversity. In Section 3, we used *The Old Man and the Sea* (book 10) as an example of a book with a low default lexical diversity. We observe that our tailored reranking system outputs the lowest original-text probability rank for this book, resulting in a lower MTLD score. For the example from Section 3 of a lexically rich book, *Ulysses* (book 15), our tailored system outputs a rank with a original-text probability higher than the minimum, thus retrieving an MTLD score that is higher. This shows that tailoring is at least somewhat intuitive.

8 Conclusion

We have argued for flexible recovery of lexical diversity in literary MT. We showed that default diversity varies per book in our dataset, and that this lexical diversity is partially lost through MT. We presented the first approach towards tailored rescoring of translation candidates, which matches HT more closely than previous baselines for some books. Future work could explore how our method can be combined with previous work, as it is in principle model-agnostic. Investigations with document-level translation, instead of sentence-level translation only, could provide additional insights. Furthermore, it may be useful to address this task at an even finer-grained level, by exploring diversity reranking on a sequence-level, instead of a book-level.

Limitations

In this paper, we addressed the increase of lexical diversity in literary MT. However, it should be noted that this is does not encompass writing style as a whole. We evaluated our approach on one high-resource language pair that consist of relatively similar languages, in one translation direction. For the domain of literary translation, we find this to be difficult to avoid. Still, experiments with more languages and resource-scenarios may retrieve interesting results. Moreover, while our data is transparent in the sense that we know and can explain exactly what it contains, we cannot distribute the data ourselves because of copyright. Lastly, we acknowledge that large-scale human evaluation could give useful insights into the differences between the systems.

Acknowledgements

This work was supported by a *Semper Ardens: Accelerate research grant (CF21-0454)* from the

Carlsberg Foundation. We thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine an Hábrók high performance computing cluster.

References

Arcadinho, Samuel David, David Aparicio, Hugo Veiga, and Antonio Alegria. 2022. T5QL: Taming language models for SQL generation. In Bosselut, Antoine, Khyathi Chandu, Kaustubh Dhole, Varun Gangal, Sebastian Gehrmann, Yacine Jernite, Jekaterina Novikova, and Laura Perez-Beltrachini, editors, *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 276–286, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Baker, Mona. 1993. Corpus linguistics and translation studies—implications and applications. In *Text and Technology*, page 233. John Benjamins.

Baroni, Marco and Silvia Bernardini. 2005. A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text. *Literary and Linguistic Computing*, 21(3):259–274, 08.

Boase-Beier, Jean. 2011. *A critical introduction to translation studies*. Bloomsbury Publishing.

Collins, Michael and Terry Koo. 2005. Discriminative reranking for natural language parsing. *Computational Linguistics*, 31(1):25–70.

de Vries, Wietse, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT Model. arXiv:1912.09582, December.

Delabastita, Dirk. 2011. Literary translation. *Handbook of translation studies*, 2:69–78.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

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, July. Association for Computational Linguistics.

Fischer, Lukas and Samuel Läubli. 2020. What's the difference between professional human and machine translation? a blind multi-language study on domain-specific MT. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 215–224, Lisboa, Portugal, November. European Association for Machine Translation.

Freitag, Markus, Isaac Caswell, and Scott Roy. 2019. APE at scale and its implications on MT evaluation biases. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44, Florence, Italy, August. Association for Computational Linguistics.

Freitag, Markus, David Vilar, David Grangier, Colin Cherry, and George Foster. 2022. A natural diet: Towards improving naturalness of machine translation output. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3340–3353, Dublin, Ireland, May. Association for Computational Linguistics.

Heaton, CP. 1970. Style in the old man and the sea. *Style*, pages 11–27.

Jalota, Rricha, Koel Chowdhury, Cristina España-Bonet, and Josef van Genabith. 2023. Translating away translationese without parallel data. In Bouamor, Houda, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7086–7100, Singapore, December. Association for Computational Linguistics.

Jimenez-Crespo, Miguel A. 2023. "translationese" (and "post-editese"?) no more: on importing fuzzy conceptual tools from translation studies in MT research. In Nurminen, Mary, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 261–268, Tampere, Finland, June. European Association for Machine Translation.

Kingma, Diederik P. and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In Bengio, Yoshua and Yann LeCun, editors, *3rd International Conference on Learning Representations (ICLR 2015)*.

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

Kudo, Taku and John Richardson. 2018. SentencePiece: A simple and language independent subword

tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November. Association for Computational Linguistics.

Lakew, Surafel Melaku, Mauro Cettolo, and Marcello Federico. 2018. A comparison of transformer and recurrent neural networks on multilingual neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 641–652, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Lee, Ann, Michael Auli, and Marc'Aurelio Ranzato. 2021. Discriminative reranking for neural machine translation. In Zong, Chengqing, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7250–7264, Online, August. Association for Computational Linguistics.

Liu, Yixin and Pengfei Liu. 2021. SimCLS: A simple framework for contrastive learning of abstractive summarization. In Zong, Chengqing, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072, Online, August. Association for Computational Linguistics.

Matusov, Evgeny. 2019. The challenges of using neural machine translation for literature. In Hadley, James, Maja Popović, Haithem Afli, and Andy Way, editors, *Proceedings of the Qualities of Literary Machine Translation*, pages 10–19, Dublin, Ireland, August. European Association for Machine Translation.

McCarthy, Philip M. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. thesis, The University of Memphis.

Ott, Myle, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota, June. Association for Computational Linguistics.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Popel, Martin, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtskỳ. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications*, 11(1):1–15.

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, Brussels, Belgium, October. Association for Computational Linguistics.

Pylypenko, Daria, Kwabena Amponsah-Kaakyire, Koel Dutta Chowdhury, Josef van Genabith, and Cristina España-Bonet. 2021. Comparing feature-engineering and feature-learning approaches for multilingual translationese classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8596–8611, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

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

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November. Association for Computational Linguistics.

Riera, Jorge Braga. 2022. Literatura-traducción. *Enciclopedia de Traducción e Interpretación*.

Shen, Libin, Anoop Sarkar, and Franz Josef Och. 2004. Discriminative reranking for machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 177–184.

Shen, Lucas. 2022. LexicalRichness: A small module to compute textual lexical richness.

Thompson, Brian and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China, November. Association for Computational Linguistics.

Toral, Antonio and Andy Way. 2015. Machine-assisted translation of literary text: A case study. *Translation Spaces*, 4(2):240–267.

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.

Toral, Antonio, Andreas van Cranenburgh, and Tia Nutters, 2024. *Literary-adapted machine translation in a well-resourced language pair: Explorations with More Data and Wider Contexts*, pages 27–52. Routledge.

Trotta, Joe. 2014. Creativity, playfulness and linguistic carnivalization in james joyce's ulysses.

van der Werff, Tobias, Rik van Noord, and Antonio Toral. 2022. Automatic discrimination of human and neural machine translation: A study with multiple pre-trained models and longer context. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 161–170, Ghent, Belgium, June. 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: Research Track*, pages 222–232.

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.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vijayakumar, Ashwin K, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

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

Wright, Chantal. 2016. *Literary translation*. Routledge.

Yin, Pengcheng and Graham Neubig. 2019. Reranking for neural semantic parsing. In Korhonen, Anna, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4553–4559, Florence, Italy, July. Association for Computational Linguistics.

Yule, C Udny. 1944. *The statistical study of literary vocabulary*. Cambridge University Press.

A Test set novels

ID	Author	Title	Year Published	Genre
1	Paul Auster	Sunset Park	2010	Literary fiction
2	David Baldacci	Divine Justice	2008	Thriller, suspense
3	Julian Barnes	The Sense of an Ending	2011	Literary fiction
4	John Boyne	The Boy in the Striped Pyjamas	2006	Historical fiction
5	John le Carré	Our Kind of Traitor	2010	Thriller, spy fiction
6	Jonathan Franzen	The Corrections	2001	Literary fiction
7	Nicci French	Blue Monday: A Frieda Klein Mystery	2011	Thriller, suspense
8	William Golding	Lord of the Flies	1954	Literary fiction
9	John Grisham	The Confession	2010	Thriller, suspense
10	Ernest Hemingway	The Old Man and the Sea	1952	Literary fiction
11	Patricia Highsmith	Ripley Under Water	1991	Thriller, suspense
12	Khaled Hosseini	A Thousand Splendid Suns	2007	Literary fiction
13	John Irving	Last Night in Twisted River	2009	Literary fiction
14	E.L. James	Fifty Shades of Grey	2011	Erotic thriller
15	James Joyce	Ulysses	1922	Literary fiction
16	Jack Kerouac	On the Road	1957	Literary fiction
17	Stephen King	11/22/63	2011	Science-fiction
18	Sophie Kinsella	Shopaholic and Baby	2007	Popular literature
19	David Mitchell	The Thousand Autumns of Jacob de Zoet	2010	Historical fiction
20	George Orwell	1984	1949	Literary fiction
21	James Patterson	The Quickie	2007	Thriller, suspense
22	Thomas Pynchon	Gravity's Rainbow	1973	Historical fiction
23	Philip Roth	The Plot Against America	2004	Political fiction
24	J.K. Rowling	Harry Potter and the Deathly Hallows	2007	Fantasy
25	J.D. Salinger	The Catcher in the Rye	1951	Literary fiction
26	Karin Slaughter	Fractured	2008	Thriller, suspense
27	John Steinbeck	The Grapes of Wrath	1939	Literary fiction
28	J.R.R Tolkien	The Return of the King	1955	Fantasy
29	Mark Twain	Adventures of Huckleberry Finn	1884	Literary fiction
30	Oscar Wilde	The Picture of Dorian Gray	1890	Literary fiction
31	Irvin D. Yalom	The Spinoza Problem	2012	Historical fiction

Table 5: Information on test set books.

B Regression plots for human translation vs. original text lexical diversity

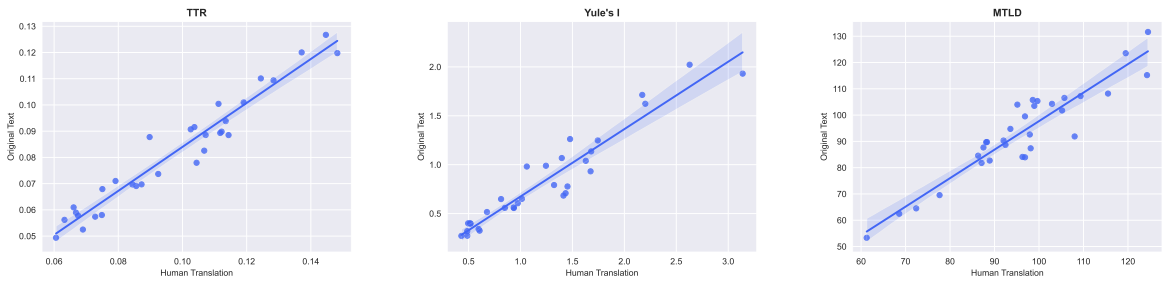

Figure 6: Regression plots for TTR, Yule's I and MTLD, with on the y-axis the scores for the original (English) versions, and on the x-axis those for human translations.

C Annotation workflow for monolingual Dutch books

1. Check whether the book is prose: we generally discard other forms of literature such as poetry and plays and annotate this in category 3 (no label).

2. Check whether the original language of the book is listed on the website of the National Dutch Library.[8] If this is not the case:

   (a) Check whether the language of the book is listed on the website of a Dutch reading community website.[9]

   (b) If step (a) is also not conclusive: check whether more information on the author is available, for instance on a personal website where we can find the original titles.

   (c) In case there is no reliable information available on the original language of a book, we discard the book (category 3: no label)

3. Book titles with Dutch as their original language are annotated with the label '1' (category 1). Books that were written in a language other than Dutch were annotated with the label '0' (category 2).

Special cases An interesting annotation case regards books from bilingual authors who learned Dutch at a later age, such as Kader Abdolah. In our current guidelines, we do not take this into account specifically; if originally written in Dutch, these books are annotated with category 1. We note that books that were translated to Dutch were not all originally written in English: other source languages in the data set include German, French and Spanish.

D Book-level MTLD comparison of APE and tailored reranking (top-k sampling)

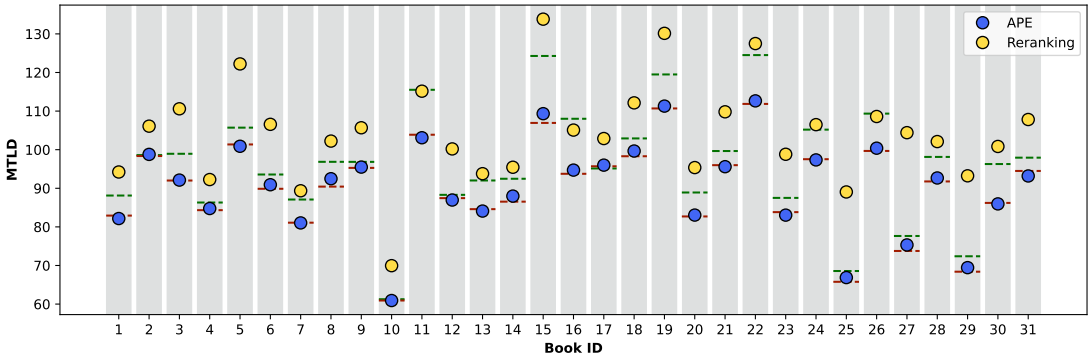

Figure 7: MTLD scores for APE and tailored reranking with top-k sampling, with on the y-axis the MTLD score for each book in our test set (x-axis).

E Lexical diversity according to ranks

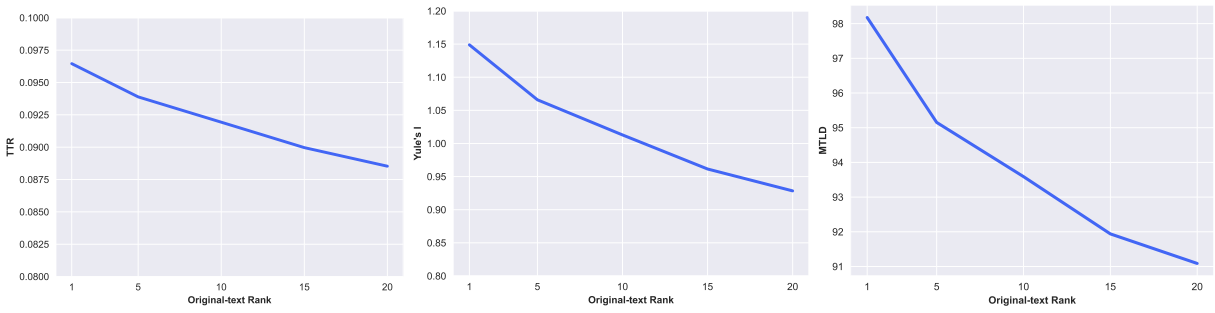

Figure 8: TTR, MTLD and Yule's I according to original-text rank, where a higher rank represents smaller original-text probability.

[8] https://www.bibliotheek.nl/
[9] https://www.hebban.nl/