Mitigating Translationese in Low-resource Languages: The Storyboard Approach

Garry Kuwanto[1], **Eno-Abasi Urua**[2], **Priscilla Amuok**[†],
Shamsuddeen Hassan Muhammad[3†], **Anuoluwapo Aremu**[†], **Verrah Otiende**[†],
Loice Nanyanga[†], **Teresiah Nyoike**[†], **Aniefon Akpan**[2], **Nsima Udouboh**[2],
Idongesit Archibong[2], **Idara Moses**[2], **Ifeoluwatayo Ige**[4†], **Benjamin Ajibade**[†],
Olumide Awokoya[†], **Idris Abdulmumin**[5†], **Saminu Mohammad Aliyu**[6†],
Ruqayya Iro[†], **Ibrahim Said Ahmad**[7†], **Deontae Smith**[8], **Praise-EL Michaels**[8],
David Ifeoluwa Adelani[9†], **Derry Tanti Wijaya**[1,10], **Anietie Andy**[8]
[1]Department of Computer Science, Boston University
[2]University of Uyo, Nigeria, [3]Imperial College London, [4]Rochester Institute of Technology
[5]Data Science for Social Impact Research Group, University of Pretoria, [6]Bayero University, Kano
[7]Institute For Experiential AI, Northeastern University
[8]Department of Electrical Engineering and Computer Science, Howard University
[9]University College London, [10]Monash University Indonesia, [†]Masakhane

{gkuwanto,wijaya}@bu.edu, {deontae.smith,praise-el.michaels}@bison.howard.edu
anietie.andy@howard.edu

Abstract

Low-resource languages often face challenges in acquiring high-quality language data due to the reliance on translation-based methods, which can introduce the translationese effect. This phenomenon results in translated sentences that lack fluency and naturalness in the target language. In this paper, we propose a novel approach for data collection by leveraging storyboards to elicit more fluent and natural sentences. Our method involves presenting native speakers with visual stimuli in the form of storyboards and collecting their descriptions without direct exposure to the source text. We conducted a comprehensive evaluation comparing our storyboard-based approach with traditional text translation-based methods in terms of accuracy and fluency. Human annotators and quantitative metrics were used to assess translation quality. The results indicate a preference for text translation in terms of accuracy, while our method demonstrates worse accuracy but better fluency in the language focused.

Keywords: Low-resource languages, Translationese, Translation Data

1. Introduction

Low-resource languages pose significant challenges when it comes to acquiring high-quality language data for various applications, including language documentation, linguistic research, and machine translation (Kuwanto et al., 2023). Traditionally, data collection in these languages involves obtaining translations from higher-resource languages, such as English. However, this approach often leads to the introduction of a translationese effect, where the resulting sentences may be less fluent and natural including those produced by professional translators.

Translationese refers to the linguistic phenomenon that occurs when translations exhibit characteristics that are not typical of the target language (Gellerstam, 1986), as well the use of more explicit and standardised constructions (Baker et al., 1993) compared to original text. These characteristics can manifest as unnatural word choices, sentence structures, or even the adoption of foreign syntactic patterns. Translators, while highly skilled in bridging the language gap, often face challenges in recreating the nuanced meaning and linguistic nuances of the original text. As a result, the translated sentences may sound unnatural or stilted to native speakers, detracting from the authenticity and quality of the collected data.

Translationese has been widely recognized for its detrimental impact not only on machine translation tasks but also on other tasks involving cross-lingual transfer learning (Amponsah-Kaakyire et al., 2022; Ni et al., 2022; Artetxe et al., 2020). The presence of translationese introduces biases, diminishes fluency and naturalness, and ultimately affects the overall quality of the output. Previous research efforts have primarily focused on mitigating the translationese effect during the downstream phase, such as treating translationese as a different language (Riley et al., 2020), applying embedding space projection (Yu et al., 2022; Chowdhury et al., 2022), and utilizing paraphrasing techniques (Artetxe et al., 2020; Wein and Schneider, 2023). However, these approaches have limitations, requiring additional annotation or modification of training data. Moreover, these methods primarily address translationese after its occurrence, rather than preventing or minimizing its presence during the data collection phase. To the best of our knowledge, there has been no prior work specifically addressing the reduction of translationese artifacts during the data collection phase itself

In this paper, we address this challenge by introducing the use of storyboards (Burton and Matthewson, 2015), a common field linguist tool, in the data collection process. Our method leverages the power of visual stimuli to elicit more fluent and natural sentences from native speakers without the explicit influence of the source language text. Instead of providing sentences, we present native speakers with a storyboard consisting of images of scenes accompanied by their English sentences an hour before the annotation process, the process where annotators . During the annotation phase, participants are asked to describe the scene in the image, focusing solely on the visual content without access to the English sentences. The primary objective of our research is to examine whether storyboards can be an alternative method to data collection when improving fluency and reducing translationese bias are of interest. By removing direct exposure to source language text during annotation, we hypothesize that the resulting sentences will still hold the meaning of the original sentence while exhibiting improved fluency and naturalness. This approach has the potential to provide a more accurate representation of the target language, facilitating the development of higher quality language resources.

Contributions: We make three major contributions: (1) the collection of data in four typologically-diverse low-resource African languages (Hausa, Ibibio, Swahili, and Yorùbá) in such a way that less translationese artifact arises, (2) the evaluation of the effectiveness of the storyboard approach in generating fluent and more natural sentences, and (3) to our knowledge, the first-ever parallel resource created data for Ibibio in non-religious domain.

In the following sections, we will delve into the details of our data collection method (Section 3), describe the experimental design, evaluation process, and present the comprehensive results of our analysis (Section 4). By combining qualitative evaluation by human annotators and quantitative metrics, we aim to provide an assessment of the effectiveness of our proposed method and its implications for data collection and the mitigation of the translationese effect.

2. Related Work

Data collection for low-resource languages has been a subject of ongoing research and development. Several initiatives have contributed to this field, such as the LORELEI project (Strassel and Tracey, 2016) and the REFLEX-LCTL project (Simpson et al., 2008). These projects, conducted by the Linguistic Data Consortium (LDC), have released annotated corpora for multiple languages, addressing the need for linguistic resources in low-resource settings.

However, the predominant approach to data collection in low-resource languages still involves leveraging monolingual data from higher-resource languages and manually translating it. This is primarily because of the scarcity of monolingual and digital data available in the target language. However, this reliance on translation-based methods introduces challenges, such as the introduction of the translationese effect, where translated sentences may be less fluent and natural in the target language (Chowdhury et al., 2022).

Translationese refers to the phenomenon in which translations exhibit linguistic characteristics that deviate from the typical patterns of the target language (Gellerstam, 1986). The impact of translationese is particularly noticeable in the areas of syntax and grammar (Santos and Redol, 1995). Translations may exhibit unnatural sentence structures, lexical and word order choices that are influenced by the source language (Gellerstam, 1996), adopt foreign syntactic patterns, as well as use more explicit and simpler constructions (Baker et al., 1993). This can result in sentences that sound unnatural or stilted to native speakers.

In the study by (Aranberri, 2020), an analysis of translationese was performed on the Spanish-Basque language pair. The researchers measured various linguistic features, including lexical variety, lexical density, length ratio, and perplexity. These measurements provided insights into the extent of translationese and its impact on the linguistic characteristics of the translated text. Similarly, (Kunilovskaya and Lapshinova-Koltunski, 2019) conducted a similar analysis focusing on English to Russian translation. In a different domain, (Bizzoni et al., 2020) conducted a study that compared translationese across human and machine translations from text and speech.

Multimodal translation tasks, such as the WMT18 multimodal task (Barrault et al., 2018), involve the use of both image and text during the translation process whether manually (Elliott et al., 2016, 2017; Barrault et al., 2018) or automatically (Wijaya et al., 2017; Hewitt et al., 2018; Rasooli et al., 2021; Khani et al., 2021). During the manual annotation for these tasks, annotators are provided with both the source image and the corresponding source text, allowing for direct reference and alignment between the two modalities. However, this direct exposure to the source segment during translation may introduce the translationese effect,

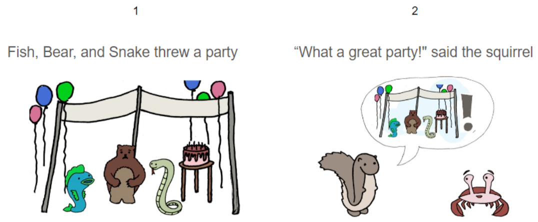
1
Fish, Bear, and Snake threw a party
2
"What a great party!" said the squirrel

Figure 1: Example of English sentence and Image pair

as observed in previous studies (Elliott et al., 2016). For example, it has been observed that the lengths of translations in German are more similar to the lengths of the source English sentences than to the lengths of German image descriptions. This suggests that the presence of the source text can influence the resulting translations and potentially impact the fluency and naturalness of the target language output. In contrast, our storyboard-based data collection approach presents annotators with only the image, without the direct exposure to the source text.

3. Data and Data Collection

In this work, we propose a new method for collecting translations for low-resource languages, with the aim of reducing the influence of source (English) sentences during translation. To achieve this, we utilize a dataset comprising images depicting various scenes and their corresponding English descriptions. For each image and its associated English sentence, we engage two groups of native speakers for each target language. One group translates the English sentences, while the other group writes descriptions (in their respective languages) based on the visual content of the images. In this section, we provide a detailed description of the data and the data collection method employed in our study.

3.1. English Sentence and Image Pair

We obtain our dataset from the Totem Field Storyboards[1], which provides a collection of storyboards consisting of sequential visual representations of stories in specific contexts, accompanied by corresponding English sentences describing the depicted scenes. An example of a sequence of English sentence and image pairs is shown in Figure 1. For our research, we select 26 storyboards from the Totem Field Storyboards, each containing an average of 19 English sentence and image pairs.

3.2. Translators and Focus Languages

Our study focuses on four low-resource African languages: Swahili, Yorùbá, Hausa, and Ibibio. For each target language, we engage a total of four native speakers. Among them, two native speakers are assigned to translate the English sentences, and the other two are shown the images and asked to write descriptions in their respective languages based on the visual content.

3.3. Data Collection

Considering our objective of determining and reducing the translationese effect in low-resource language data collection, we design the data collection process as follows: for each pair of English sentence and image in the storyboards, one group of translators focuses on translating the English sentence, while another group concentrates on translating the image. This separation allows us to examine the impact of English sentences on the resulting translations and evaluate the fluency and naturalness of the sentences generated solely based on visual content. By collecting data through this dual approach, we aim to obtain a more authentic representation of the target languages and mitigate the translationese effect. For each language, the translators were paid a total of 600 US dollars; which they shared equally among themselves.

3.3.1. Control Group: Text Translation

In our study, we include a control group that utilizes the traditional approach of text translation. This control group serves as a baseline for comparison and allows us to evaluate the effectiveness of our storyboard-based method in reducing the translationese effect.

For the control group, native speakers are provided with the English sentences from the storyboards and are instructed to translate them directly into the target languages. These native speakers possess proficiency in both the source and target languages.

The translations generated by the control group represent the typical output obtained through traditional translation-based approaches. These translations are expected to exhibit characteristics of translationese, such as potential deviations from natural language usage, as translators may prioritize fidelity to the source text over fluency in the target language.

3.3.2. Treatment Group: Storyboard-Based Translation

In our storyboard-based translation method, we introduce a preparatory phase before the actual

[1] https://totemfieldstoryboards.org/

annotation process. Before annotating each storyboard, the annotators are grouped together in a meeting where they are given the opportunity to familiarize themselves with the storyboard and the corresponding English sentences. During this meeting, they can read through the storyboard and comprehend the context and content of each image.

After this reading session, we introduce a time gap of ∼1 hour before the annotation process begins. This time gap serves two purposes: (1) it allows the annotators to internalize the visual information from the storyboard, and (2) it minimizes the direct influence of the English sentences on their subsequent annotations.

During the annotation phase, the annotators are provided with the storyboards containing only the images, without any accompanying English sentences. They are instructed to focus solely on the visual content and describe in their respective target languages the scene that is being depicted in each image. This approach ensures that the annotations are driven primarily by the visual stimuli, encouraging the annotators to provide translations that capture the essence of the scenes portrayed in the images.

By removing the explicit exposure to the English sentences during the annotation process, our storyboard-based method aims to reduce the potential influence of the source language and the translationese effect. The annotators are encouraged to rely on their linguistic knowledge and cultural understanding to generate fluent and natural translations that are more aligned with the target language's usage patterns and stylistic conventions.

3.4. Annotation Dataset

The resulting annotation dataset consists of the source English sentences and their corresponding translations in four low-resource African languages: Swahili, Yorùbá, Hausa, and Ibibio. The dataset includes both text translations and translations obtained through our storyboard-based method.

In total, we collected translations for 486 unique English sentences. The distribution of translations is shown in Table 1, indicating the number of translations for each language and data collection method (text or storyboard). The number between text and storyboard translations differs slightly because we gave translators the option to provide alternative translations for each English sentences.

The dataset also includes additional information such as the title of the storyboard and the scene number associated with each translation. We will make the final dataset, along with the source English sentences, publicly available. An anonymized version of the dataset is also attached in the supplementary material for reference.

Language	Text Translation	Storyboard
Hausa	1154	968
Ibibio	887	883
Swahili	1334	1211
Yorùbá	1448	1033

Table 1: Number of translations in the dataset

4. Experimental Design

All computational experiments were conducted on a machine equipped with an Intel Xeon Gold 6226R CPU running at a clock speed of 2.90GHz and an NVIDIA RTX A6000 GPU.

4.1. Human Evaluation

Fluency, for our human evaluation setup, refers to the smoothness and naturalness of the translated sentence in the target language. A fluent sentence should not sound "foreign" or awkward, and should read as if a native speaker had originally written it in that language. Fluency captures the syntactic and grammatical correctness as well as the idiomatic usage of the language.

Accuracy, for our case, refers to the extent to which the translated sentence captures the meaning of the source sentence. An accurate or adequate translation should convey all essential information from the source text without adding, omitting, or distorting any content.

To assess the accuracy and fluency of the translated sentences obtained through the traditional text translation and our proposed storyboard-based data collection processes, we conducted human evaluation with native speakers who are also proficient in English as annotators. Annotators were assigned two tasks: one for accuracy evaluation and another for fluency evaluation. Each task involved comparing a pair of sentences, one from the text translation and the other from the storyboard-based collection process. It is important to note that the sentence pairs provided to the annotators for different languages were taken from the same storyboard scene to ensure consistency and fair comparison across languages. For each task (accuracy and fluency), we randomly select 100 samples from the storyboard scenes and obtain sentence pairs.

To minimize bias and ensure reliable results, three human annotators were assigned to each task. Each annotator independently evaluated the sentence pairs and provided their preference based on accuracy or fluency. The preferences of the annotators were then tallied up to determine the overall preference for each translation approach.

After the evaluation, because we have 3 annotators for each language, the inter-annotator agreement was calculated using the Fleiss' Kappa statistic to measure the consistency of the annotators' preferences. A Fleiss' Kappa value of 1 indicates perfect agreement between the annotators, while a value of 0 suggests agreement equivalent to random chance. For our evaluation, the Fleiss' Kappa value was found to be 0.27 and 0.16 for accuracy and fluency respectively, indicating fair and slight level of agreement among the annotators respectively (Landis and Koch, 1977).

4.1.1. Accuracy Evaluation

For the accuracy evaluation task, annotators were presented with three sentences: the source English sentence, the sentence translated through text translation, and the sentence translated through the storyboard-based method. Their task was to choose which sentence they deemed more accurate for translating the English sentence; they can pick one sentence over the other, or both. The guideline was to "select which sentence is more adequate (i.e., more accurate) for translating the English sentence." An important criterion emphasized in the guideline was that a better translation should include as much content from the English sentence as possible, without adding information not present in the original sentence. The annotators were also asked to disregard the translations of named entities in their accuracy judgment. This was particularly relevant as, during the storyboard data collection process, the translators might not remember the exact names mentioned in the English sentence, while the translators for the text translation could reference the English sentence for accuracy.

To reduce bias, annotators were only provided with the sentences without any indication of which sentence was from the text translation and which was from the storyboard-based method. An example of the accuracy evaluation task for Hausa can be seen in Table 2 where annotators would compare Sentence 1 and Sentence 2, selecting the one they considered more accurate, or both if they deemed the two sentences to be equally accurate.

4.1.2. Fluency Evaluation

For the fluency evaluation task, annotators were not provided with the source English sentence. Instead, they were presented with two sentences from the same storyboard scene: one sentence obtained through text translation and another obtained through the storyboard-based method. Their task was to select the sentence that they deemed to be more fluent and natural. The guideline was to "select which sentence is more fluent (i.e., more natural). A better sentence should be the one that is more natural and grammatical". The annotators were asked to focus on factors such as natural language usage, grammaticality, and overall fluency in making their decision.

To ensure unbiased evaluation, the annotators were only provided with the sentences without any indication of which sentence was from the text translation and which was from the storyboard-based method. An example of the fluency evaluation task for Hausa can be seen in Table 3 where annotators would compare Sentence 1 and Sentence 2, selecting the one they considered more fluent, or both if they deemed the two sentences to be equally fluent.

4.2. Metric-Based Evaluation of Accuracy and Fluency

In addition to human evaluation, we employ several metrics to support and complement the results obtained from the human annotators. These metrics provide quantitative measures of accuracy and fluency, enabling a more comprehensive analysis of the translation quality.

4.2.1. Accuracy

To evaluate the accuracy of translated sentences, we utilize the LASER model (Heffernan et al., 2022), which excels in capturing semantic meaning and supports a wide range of languages except for Ibibio, which is not in the list of 147 languages that the LASER encoder supports. We employ this model to compute the embeddings of the translated sentences and their corresponding English sentences. To compare the embeddings, we compute the cosine distance, which serves as a measure of semantic similarity between the translations and the source English sentences. A higher cosine similarity indicates a stronger alignment in semantic meaning, suggesting higher accuracy in the translation process.

In addition to evaluating the accuracy between the translated sentences and their corresponding English sentences, we also examine the similarity between the translations generated through our storyboard-based method and the text translation method. To measure whether the two methods result in translations with comparable accuracy, we calculate the cosine similarity between embeddings of the translated sentences. The cosine similarity provides an indication of how similar the translations are in terms of their semantic meaning. While we do not expect a perfect similarity between

English Sentence	I bought shoes, a hat, a shirt, and pants
Sentence 1	Na siyo takalma, malafa, riga da kuma wando
English Translation	*I bought shoes, mattress, shirt and pants*
Sentence 2	Na sayo takalma, hula, riga da kuma yan kamfai
English Translation	*I bought shoes, a hat, a shirt and a few shoes*

Table 2: An Example of Accuracy Human Evaluation for Hausa. Annotators were asked to choose the sentence they deemed more accurate for translating the English sentence. They can answer with Sentence 1, Sentence 2, or both if they deemed both sentences to be equally accurate.

Sentence 1	Mary ta fita zuwa shago
English Translation	*Mary goes out to the shop*
Sentence 2	Ta fita siyayya watarana
English Translation	*She went out of the shopping*

Table 3: An Example of Fluency Human Evaluation for Hausa. Annotators were asked to choose the sentence they deemed more fluent. They can answer with Sentence 1, Sentence 2, or both if they deemed both sentences to be equally fluent.

the two methods' translations, we aim for a comparable level of meaning that allows for fluency trade-off.

4.2.2. Fluency

Fluency in translation involves the successful use of vocabulary and sentence structure to convey meaning effectively in the target language. When translationese phenomena occur, patterns emerge that can impact the fluency of translated sentences. Previous studies (Vanmassenhove et al., 2021; Bizzoni et al., 2020) have highlighted two key factors in fluency: lexical diversity and syntactic complexity.

Lexical Diversity To assess the lexical diversity or vocabulary richness of the translated sentences, we employ the Measure of Textual Lexical Diversity (MTLD) metric (McCarthy, 2005). Conceptually, MTLD reflects the average number of words in a row for which a certain TTR (Type-Token Ratio) is maintained, specifically in this analysis, following McCarthy and Jarvis (2010) we use TTR of 0.72. To generate a score, MTLD calculates the TTR for increasingly longer parts of the sample. Every time the TTR drops below the predetermined value, a count (called the factor count) increases by 1, and the TTR evaluations are reset. The algorithm resumes from where it had stopped, and the same process is repeated until the last token of the language sample has been added and the TTR has been estimated. Then, the total number of words in the text is divided by the total factor count. Subsequently, the whole text in the language sample is reversed and another score of MTLD is estimated. The forward and the reversed MTLD scores are averaged to provide the final MTLD estimate.

Intuitively the MTLD value can be seen as the average number of words required for the text to reach a point of stabilization.

Syntactic Complexity: To evaluate syntactic complexity, we leverage an off-the-shelf language model trained on the languages used in the study, specifically AfroXLMR-base (Alabi et al., 2022)—an adaptation of XLM-R to 17 African languages, including Hausa, Swahili and Yorùbá. To estimate the POS perplexity (Bizzoni et al., 2020), we trained a Part-Of-Speech (POS) model by fine-tuning the AfroXLMR language model on Masakha-POS dataset (Dione et al., 2023)— a large-scale POS dataset for 20 African languages, including all focus languages except for Ibibio. POS perplexity measures the difficulty in predicting the sequence of POS tags in a sentence. Higher perplexity values indicate greater syntactic complexity. Perplexity is defined as the exponentiation of the entropy:

$$2^{H(p)} = 2^{-\sum_x p(x)\log_2 p(x)} \qquad (1)$$

Where $p(x)$ is the probability of the predicted POS.

4.3. Results

4.3.1. Accuracy Evaluation

Language	Storyboard	Text	Both
Hausa	21.67%	78.33%	0%
Swahili	14%	64%	22%
Yorùbá	7.67%	68.67%	23.67%
Ibibio	18.33%	79.67%	2%

Table 4: Human Evaluation Results for Accuracy. More annotators choose text translations as more accurate across languages

The accuracy evaluation results in Table 4 show a clear preference among human annotators for text translations in terms of accuracy across all languages. This is consistent with the expectation

Language	Storyboard	Text
Hausa	0.64 ± 0.14	0.74 ± 0.13
Swahili	0.58 ± 0.11	0.69 ± 0.10
Yorùbá	0.64 ± 0.13	0.74 ± 0.12

Table 5: The Average Cosine Similarity Scores between English and Translated Sentences' LASER Embeddings. Across languages, text translations have higher semantic similarity to source English sentences, although the scores still fall within each method's standard deviation

Language	Cosine Similarity
Hausa	0.63 ± 0.19
Swahili	0.62 ± 0.17
Yorùbá	0.59 ± 0.18

Table 6: The Average Cosine Similarity Scores between text translations and storyboard-based translations' LASER embeddings.

that text translations, typically done by professional translators, would be more semantically aligned with the source sentences.

The cosine similarity values between LASER embeddings of the translated sentences and their corresponding English sentences, as shown in Table 5, further support this observation. Text translations generally exhibit higher semantic similarity to the source English sentences. However, it's important to note that while the cosine similarity values for storyboard translations are lower, they are still within the standard deviation range of the text translations, suggesting that the semantic content is not drastically different.

The cosine similarity between translations from the two methods, presented in Table 6, indicates that the semantic content of translations from both methods is relatively comparable, even if they are not identical.

4.3.2. Fluency Evaluation

Language	Storyboard	Text	Both	p-value
Hausa	60%	39.67%	0.33%	0.0002
Swahili	47.67%	41.33%	11%	0.11
Yorùbá	34%	18.6%	47.33%	0.0008
Ibibio	36%	26%	38%	0.01

Table 7: Human Evaluation Results for Fluency. More annotators choose storyboard translations as more fluent across languages. P-Value is for the null hypothesis of the annotators choosing by random.

The fluency evaluation in Table 7 suggests a preference among human annotators for storyboard translations in terms of fluency across all languages. This indicates that while storyboard translations might not always capture the exact semantic content of the source, they tend to produce more naturally flowing sentences in the target languages.

Language	Storyboard	Text
Ibibio	8.08 ± 17.06	7.27 ± 18.35
Hausa	7.8 ± 16.77	11.41 ± 23.48
Yorùbá	16.12 ± 30.05	14.16 ± 29.84
Swahili	6.83 ± 19.64	5.03 ± 16.61

Table 8: MTLD of Translated Sentences. Across languages, except for Hausa, storyboard translations have higher average MTLD scores

Language	Storyboard	Text
Hausa	6.68	49.42
Swahili	55.32	14.6
Yorùbá	6.49×10^2	7.93×10^7

Table 9: POS Perplexity of Translated Sentences. Across languages, except for Swahili, the storyboard translations have lower POS perplexity, indicating more natural translations in terms of sentence structure than the text translations

The MTLD scores in Table 8, despite having a high standard deviation, provide further evidence for this observation. Except for Hausa, storyboard translations generally exhibit greater lexical diversity, which is often associated with more natural and fluent sentences.

The POS perplexity values in Table 9 offer insights into the syntactic complexity of the translations. Lower POS perplexity values typically indicate more natural sentence structures in the target language. The results suggest that, except for Swahili, storyboard translations tend to produce more naturally structured sentences than text translations.

In summary, while text translations are generally more accurate and semantically aligned with the source, storyboard translations offer advantages in terms of fluency and naturalness of the produced sentences.

5. Discussion

5.1. Comparison of Accuracy Evaluation

The human accuracy evaluation results (Table 4) reveal a clear preference for text translation over our proposed method. Human evaluators consistently rated text translation as more accurate compared to our method. This preference aligns with the linguistic intuition that text translations, performed by professional translators, tend to produce more accurate translations. Similarly, looking at the cosine similarities between the source English sentences and the translations (Table 5), we can see that text translations have higher cosine similarities to the English sentences, indicating higher semantic similarities. This observation is consistent with human evaluators' preference for accuracy.

However, when comparing with the cosine similarities between our method's (storyboard) translations and the text translations (Table 6), we found that the cosine similarity scores were not significantly lower. This suggests that the information conveyed through our storyboard translations remains comparable to that of text translations, albeit with a slightly lower degree of semantic similarity. Despite the storyboard translations not matching the accuracy of the text translations perfectly, our storyboard-based data collection method still provides translations that are reasonably accurate and comparable.

5.2. Comparison of Fluency Evaluation

The human fluency evaluation presents a contrast to the accuracy evaluations. Human evaluation results reveal a clear preference for our storyboard translations over text translations in terms of fluency (Table 7).

In analyzing the fluency metrics, looking at the Measure of Textual Lexical Diversity (MTLD), our storyboard translations consistently yielded higher MTLD scores compared to text translations for Ibibio, Yorùbá, and Swahili (Table 8). This suggests that our storyboard-based data collection method can capture a wider range of vocabulary and exhibit greater lexical diversity, enhancing the fluency of the resulting translations in these languages.

We also evaluated the syntactic complexity of the translations using the Part-Of-Speech (POS) perplexity metric, which provide insights into the syntactic intricacy and sentence structure of the translations. Looking at POS perplexity, our storyboard translations exhibited lower POS perplexity values for Hausa and Yorùbá compared to text translations (Table 9). This suggests that our storyboard-based data collection method achieves a more natural sentence structure in these languages.

5.3. Reflections on the Storyboard Approach

The storyboard approach, while pioneering, calls for an in-depth examination of its intrinsic strengths and limitations, especially when contrasted against conventional text translation techniques.

Enhanced Fluency As underscored by our findings, the storyboard technique frequently yields translations perceived as more fluent by human evaluators. This can be ascribed to the more spontaneous elicitation process, where native speakers articulate visual stimuli, culminating in more innate sentence constructs.

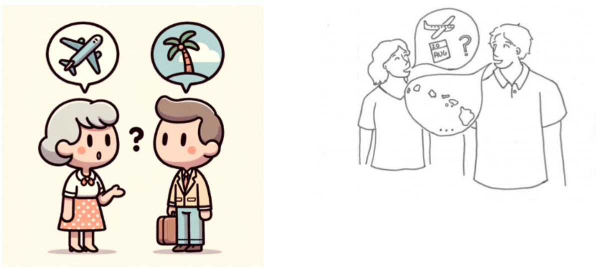

Figure 2: Comparison between the DALLE-3 generated storyboard (left) and the manually designed storyboard (right). The similarities highlight the potential of generative AI in automating the storyboard creation process

Lexical Diversity: The method appears to encapsulate a broader lexicon, potentially enriching the translations with diverse linguistic expressions.

Accuracy Concerns The storyboard approach, while fluent, occasionally compromises on accuracy. This is seen in both human evaluations and cosine similarity metrics. The omission or modification of certain nuances, particularly named entities, can influence the perceived precision of the translations. We have a few suggestions on how to improve this. Firstly, refining the storyboards to provide clearer context, especially regarding named entities, can help translators capture essential details. Secondly, implementing post-processing techniques can further align the translations with the source content. By incorporating correct named entities from the source and removing extraneous information, we can achieve a better balance between fluency and accuracy. Finally, introducing temporal separation between viewing the source and translating can help translators produce more natural translations, less influenced by the source language structure. By integrating these strategies, we aim to enhance the accuracy of the storyboard-based approach without compromising its inherent fluency.

Feasibility for Complex Texts The viability of the method for translating intricate and detailed texts remains a point of contention. Storyboards, by their visual essence, might fall short in capturing the depth and nuances of elaborate narratives or technical documents.

6. Future Work

The storyboard approach, while innovative, faces challenges in terms of the resource-intensive nature of creating detailed storyboards. A promising solution lies in the integration of generative AI models, such as DALLE-3. Our initial exploration with DALLE-3 to recreate a manually designed storyboard showed potential in automating the storyboard creation process, as evidenced by the comparison in Figure 2. However, inconsistencies in DALLE-3's generated characters and the challenge of conveying complex messages visually remain areas for improvement. As generative AI models evolve, there's potential for more sophisticated storyboard generation. Collaborative efforts between linguists, artists, and AI researchers could lead to hybrid methodologies that merge manual design with AI-driven automation. Other future work could consider expanding the complexity of messages captured in storyboard, and improving the accuracy of the storyboard.

7. Conclusion

We introduced an alternative method for gathering translation data in low-resource settings using storyboards and visual-based translation techniques. Our findings, derived from both human assessments and automated metrics for accuracy and fluency, revealed that while traditional text translations excel in accuracy, our method offers a balance of accuracy and heightened fluency across multiple languages. Notably, the use of visual-based translation enriched lexical variety and fostered more organic sentence formations, leading to translations that resonate more naturally. these findings have implications for the development of data collection systems that can mitigate the translationese phenomena. The potential applications and impact of our approach include improvement in performance of machine translation task and other tasks that will benefit from better cross-lingual training data and transfer.

Limitations

While our study focuses on and observes the effectiveness of the storyboard-based data collection method and its potential to mitigate the translationese effect, it is important to acknowledge certain limitations.

Firstly, our study focused on four low-resource African languages, namely Swahili, Yorùbá, Hausa, and Ibibio. While we also contribute to the development of parallel resources for these languages (including the first-ever parallel resource for Ibibio in non-religious domain), our analysis and findings on the translationese phenomena in the resulting dataset may not necessarily generalize to other languages or language families. Therefore, caution should be exercised when applying the results to different linguistic contexts. In addition, our study specifically focused on the use of storyboards to collect data in these low-resource languages. The effectiveness of this method may vary in different data collection scenarios, such as with different types of visual stimuli or in languages with different linguistic characteristics.

Secondly, the number of annotators and the size of the dataset used in our study were limited. Although efforts were made to mitigate bias and ensure reliability through multiple annotators, a larger sample size could provide more robust and representative results. In addition, the human evaluation process involves subjective judgments made by human annotators. Individual preferences and biases may influence the evaluation results. While we attempted to minimize bias by using multiple annotators and consistent sentence pairings, the subjective nature of the evaluation should be considered.

Thirdly, in translating, the fluency and naturalness of the translated sentences can be influenced by various external factors, such as the annotators' language proficiency, cultural background, and familiarity with the subject matter. While efforts were made to select skilled annotators, these factors may still have an impact on the results.

Furthermore, the storyboard-based data collection method, while innovative, is inherently more challenging compared to traditional text translation. This method involves a more complex setup for data collection, as participants need to be effectively oriented to provide translations based on visual stimuli rather than textual source material. These logistical and financial considerations make the method potentially less scalable and more expensive than traditional approaches, especially for larger datasets or multiple languages.

Lastly, our evaluation mainly focused on accuracy and fluency, and utilized metrics such as cosine similarity, lexical diversity, and syntactic complexity. While these metrics provide some measures of translationese, they may not capture all aspects of translation quality. Additional metrics or qualitative assessments could further enhance the evaluation.

Ethics Statement

The data collection process adhered to standard guidelines. Informed consent was obtained from all participants involved in the study, and they were fully informed about the purpose of the research, their rights, and how their data would be used. Participants were assured of the anonymity and confidentiality of their information throughout the data collection process.

Acknowledgements

This work is supported in part by the DARPA HR001118S0044 (the LwLL program) and the

Indonesia-US Research Collaboration in Open Digital Technology grant funded by the Indonesian Ministry of Education, Culture, Research, and Technology. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of DARPA or the U.S. Government.

8. Bibliographical References

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Kwabena Amponsah-Kaakyire, Daria Pylypenko, Josef van Genabith, and Cristina España-Bonet. 2022. Explaining translationese: why are neural classifiers better and what do they learn? In *BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*.

Nora Aranberri. 2020. Can translationese features help users select an mt system for post-editing? *Proces. del Leng. Natural*, 64:93–100.

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

Mona Baker, Gill Francis, and Elena Tognini-Bonelli. 1993. 'corpus linguistics and translation studies: Implications and applications'. In *Text and Technology: In Honour of John Sinclair*. John Benjamins Publishing Company.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels. Association for Computational Linguistics.

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

Strang Burton and Lisa Matthewson. 2015. Targeted construction storyboards in semantic fieldwork. *Methodologies in semantic fieldwork*, pages 135–156.

Koel Dutta Chowdhury, Rricha Jalota, Cristina España-Bonet, and Josef van Genabith. 2022. Towards debiasing translation artifacts.

Cheikh M. Bamba Dione, David Ifeoluwa Adelani, Peter Nabende, Jesujoba Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jonathan Mukiibi, Blessing Sibanda, Bonaventure F. P. Dossou, Andiswa Bukula, Rooweither Mabuya, Allahsera Auguste Tapo, Edwin Munkoh-Buabeng, victoire Memdjokam Koagne, Fatoumata Ouoba Kabore, Amelia Taylor, Godson KALIPE, Tebogo Macucwa, Vukosi Marivate, Tajuddeen Gwadabe, Mboning Tchiaze Elvis, Ikechukwu Onyenwe, Gratien Atindogbe, Tolulope Adelani, Idris Akinade, Olanrewaju Samuel, Marien Nahimana, Théogène Musabeyezu, Emile Niyomutabazi, Ester Chimhenga, Kudzai Gotosa, Patrick Mizha, Apelete Agbolo, Seydou Traore, Chinedu Uchechukwu, Aliyu Yusuf, Muhammad Abdullahi, and Dietrich Klakow. 2023. MasakhaPOS: Part-of-Speech Tagging for Typologically Diverse African languages . In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

D. Elliott, S. Frank, K. Sima'an, and L. Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Martin Gellerstam. 1986. Translationese in swedish novels translated from english.

Martin Gellerstam. 1996. Translations as a source for cross-linguistic studies. *Lund studies in English*, 88:53–62.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

John Hewitt, Daphne Ippolito, Brendan Callahan, Reno Kriz, Derry Tanti Wijaya, and Chris Callison-Burch. 2018. Learning translations via images with a massively multilingual image dataset. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2566–2576, Melbourne, Australia. Association for Computational Linguistics.

Nikzad Khani, Isidora Tourni, Mohammad Sadegh Rasooli, Chris Callison-Burch, and Derry Tanti Wijaya. 2021. Cultural and geographical influences on image translatability of words across languages. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 198–209, Online. Association for Computational Linguistics.

Maria Kunilovskaya and Ekaterina Lapshinova-Koltunski. 2019. Translationese Features as Indicators of Quality in English-Russian Human Translation. In *Proceedings of the Second Workshop Human-Informed Translation and Interpreting Technology associated with RANLP 2019*, pages 47–56. Incoma Ltd., Shoumen, Bulgaria.

Garry Kuwanto, Afra Feyza Akyürek, Isidora Chara Tourni, Siyang Li, Alex Jones, and Derry Wijaya. 2023. Low-resource machine translation training curriculum fit for low-resource languages. In *Pacific Rim International Conference on Artificial Intelligence*, pages 453–458. Springer.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Philip M McCarthy. 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.

Philip M. McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392.

Jingwei Ni, Zhijing Jin, Markus Freitag, Mrinmaya Sachan, and Bernhard Schölkopf. 2022. Original or translated? a causal analysis of the impact of translationese on machine translation performance. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5303–5320, Seattle, United States. Association for Computational Linguistics.

Mohammad Sadegh Rasooli, Chris Callison-Burch, and Derry Tanti Wijaya. 2021. “wikily” supervised neural translation tailored to cross-lingual tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1655–1670, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Diana Santos and Rua Alves Redol. 1995. On grammatical translationese.

Heather Simpson, Christopher Cieri, Kazuaki Maeda, Kathryn Baker, and Boyan Onyshkevych. 2008. Human Language Technology Resources for Less Commonly Taught Languages: Lessons Learned Toward Creation of Basic Language Resources.

Stephanie Strassel and Jennifer Tracey. 2016. LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3273–3280, Portorož, Slovenia. European Language Resources Association (ELRA).

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics.

Shira Wein and Nathan Schneider. 2023. Translationese reduction using abstract meaning representation.

Derry Tanti Wijaya, Brendan Callahan, John Hewitt, Jie Gao, Xiao Ling, Marianna Apidianaki, and Chris Callison-Burch. 2017. Learning translations via matrix completion. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1452–1463, Copenhagen, Denmark. Association for Computational Linguistics.

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