Abstract: French is a strategically and economically important language in the regions where the African language Twi is spoken. However, only a very small proportion of Twi speakers in Ghana speak French. The development of a Twi–French parallel corpus and corresponding machine translation applications would provide various advantages, including stimulating trade and job creation, supporting the Ghanaian diaspora in French-speaking nations, assisting French-speaking tourists and immigrants seeking medical care in Ghana, and facilitating numerous downstream natural language processing tasks. Since there are hardly any machine translation systems or parallel corpora between Twi and French that cover a modern and versatile vocabulary, our goal was to extend a modern Twi–English corpus with French and develop machine translation systems between Twi and French: Consequently, in this paper, we present our Twi–French corpus of 10,708 parallel sentences. Furthermore, we describe our machine translation experiments with this corpus. We investigated direct machine translation and cascading systems that use English as a pivot language. Our best Twi–French system is a direct state-of-the-art transformer-based machine translation system that achieves a BLEU score of 0.76. Our best French–Twi system, which is a cascading system that uses English as a pivot language, results in a BLEU score of 0.81. Both systems are fine tuned with our corpus, and our French–Twi system even slightly outperforms Google Translate on our test set by 7% relative.

Keywords: Akuapem Twi; machine translation; low-resource language; Twi–French

1. Introduction

In recent years, machine translation systems have played a key role in communication by removing language barriers [1]. Google’s Neural Machine Translation System [2], for instance, is a multilingual machine translation system, which handles translations of over 100 language pairs. The need for machine translation services has expanded in recent years due to the massive interchange of information across different regions using multiple regional languages [3]. Companies operating in numerous countries throughout the world use machine translation services for a variety of purposes, including internal and external communication, client interaction on a global scale, and more [4]. Moreover, people all around the world are now able to communicate in a variety of languages on social media platforms because of machine translation systems [5]. Furthermore, machine translation has shown considerable potential in terms of revolutionizing foreign language teaching and other applications in the field of education [6–9], and research also demonstrates that machine translation has increased international trading [10].

Despite these numerous benefits, machine translation is not available or has not produced desirable results in several indigenous African languages [11], as compared to the state-of-the-art results achieved with high-resource languages, such as English, Spanish, and French [12–15]. High-resource languages are those that have a large volume of digitalized text [16]. One major reason for the shortfall of machine translation in most African languages is the lack of parallel corpora for these languages [17]. Parallel corpora are required for training machine translation systems, and the performance of statistical machine translation and neural machine translation systems is directly impacted by the number of parallel sentence
pairs available for training [18]. Languages with a large volume of parallel corpora available, such as English and French, are spoken globally and have established themselves in many regions as a mode of education and other activities related to communication [19]. Having machine translation services from an indigenous African language into one of these globally recognized languages has become an essential tool not only for improving communication between the rest of the world and language speakers, but also for assisting the rest of the world in learning about the people’s culture [20].

Ghana is a country in West Africa with over 75 indigenous languages [21]. The Akan Twi language is the most widely spoken language, with about 80% of the country speaking it as their first or second language [21]. Despite Ghana sharing common borders with Togo, Burkina Faso and Côte d’Ivoire, whose official languages are French [22], only an estimated 1–5% of Ghanaians can speak or comprehend some French [23]. French is a global language with an estimated 300 million speakers and “official language of 32 states and governments” [24]. With 59% of global French speakers in Africa and as a major global language for trading [25], the demand to learn French among Ghanaians is increasing rapidly. The demand even led the Ghanaian parliament to approve French as a second official language [26]. A machine translation system between Ghana’s most widely spoken language and French is certainly a valuable resource for bridging the gap.

However, most of the Twi natural language processing resources that can be used to build machine translation systems are classified as noisy and religiously biased [21]. There are already machine translation systems that allow translations from local Ghanaian languages to other languages. For example, Khaya, a neural machine translation system by the GhanaNLP group and Algorine, allows machine translation from Ghanaian languages, such as Twi, Ewe, and Ga, to English [27]. A recent addition of Twi to Google’s Neural Machine Translation System enables translations from Twi to over 100 languages [28]. Furthermore, pre-trained Twi machine translation models are provided by the Language Technology Research Group at the University of Helsinki, including a model for Twi and French machine translation [29], which, however, require parallel corpora to be fine tuned.

Consequently, since there are still hardly any open-source machine translation systems or parallel corpora between Twi and French that cover a modern and versatile vocabulary, our goal was to extend the modern English–Akuapem Twi corpus of [21] and develop machine translation systems between Twi and French. A Twi–French parallel corpus and the corresponding machine translation applications will offer various advantages, including stimulating trade and job creation. Moreover, it will support the global Ghanaian diaspora in French-speaking nations, as they will be able to acquire Twi and Ghanian culture [21]. The system will also help French-speaking tourists and immigrants seeking medical care in Ghana [30]. Additionally, the French–Twi parallel corpus can be used for numerous downstream natural language processing tasks, including named entity recognition and part-of-speech tagging with the appropriate annotations.

Our contributions are as follows:

• We are the first to introduce non-commercial machine translation systems for Twi–French and French–Twi.
• We created a parallel Twi–French corpus by extending an existing Twi–English corpus.
• For our language pairs, we investigated direct machine translation and cascading systems that use English as a pivot language.
• We compared our systems with the commercial system of Google Translate and managed to slightly outperform Google Translate with our best French–Twi system.
• To contribute to the improvement of low-resource languages, we share our code and our corpus with the research community (https://github.com/gyasifred/TW-FR-MT).

In the following section, we will give a brief insight into the linguistic categorization and the peculiarities of Akuapem Twi. In Section 2, we will describe related work regarding existing parallel text corpora and machine translation systems for Twi. Section 4 will present our parallel Twi–French–English corpus. Our Twi–French and French–Twi machine translation experiments will be described in Section 5. In Section 6, we will briefly discuss
the performance of our machine translation output. We will conclude our work in Section 7 and suggest further steps.

2. The Language Twi

As visualized in Figure 1, Twi is a collection of dialects which belongs to Akan. Akan is the language of the Akan ethnic group in Ghana [31] and the principle language of Ghana [32]. The Akan dialects include Agona, Akuapem, Akwamu, Asante, Akyem, Assin, Bono, Fante, Kwahu, Wassa, Sefwi, Anyi, and Guan [33]. These dialects are divided into two categories: Fante and Twi (which includes all non-Fante dialects) [33]. Since not all dialects understand each other, Akuapem Twi serves as a pivot language and is used for education purposes in schools. The Akan Orthography Committee (AOC) developed a unified Akan orthography in 1978, based mainly on Akuapem Twi [34]. Consequently, those Twi dialects that do not have their own orthography use Akuapem Twi (marked in red) [35]. Only Asante and Bono have separate writing systems (marked in orange). Fante does not belong to the Twi dialects and has its own orthography (marked in orange).

![Figure 1. Akan dialects.](image)

To allow all Twi-speaking people to have access to French and since a unified orthography exists, we decided to collect a corpus and build machine translation systems in Akuapem Twi. For simplicity and since all Twi-speaking people would be able to operate with Akuapem Twi machine translation, we refer to Akuapem Twi as Twi in this paper.

Twi is a tonal language with distinct semantic connotations for its high, mid, and low tones [21]. The Twi alphabet consists of 22 letters made up of 15 consonants and 7 vowels [36]. Additionally, though mostly in loanwords, the letters C, J, V, and Z are used. Twi has ten diphthongs. Many Twi words have multiple meanings and can be used interchangeably in the same context [37]. For example, the word sequence “me papa” means “good mood” or “my dad”, depending on the context. In contrast to other languages, removing stop words may affect the entire meaning of a Twi sentence. For instance, the Twi word “na” could represent the word “and”, the word “then”, the phrase “and then”, or the word “mother”. Consequently, we do not remove Twi stop words in the pre-processing pipeline of our machine translation systems.

3. Related Work

In this section, we will look at existing Twi text corpora and machine translation systems.
3.1. Parallel Corpora for Twi

The authors of [37] analyzed the use of the Twi Bible dataset, Jehovah’s Witness data, Wikipedia, and the JW300 Twi corpus [38]. However, they classified Jehovah’s Witness data, Wikipedia and the JW300 Twi corpus as noisy, i.e., not optimal for the use in machine translation tasks due to spellings, Twi sentences formulated in a non-natural way, mixtures of dialects, etc. Furthermore, [21] report that the JW300 Twi corpus and the Bible are religiously biased datasets.

Despite the “noise” and the religious bias, JW300 is a huge corpus “of over 300 languages with around 100 thousand parallel sentences per language pair on average” covering a wide range of topics [38]. In addition to other language pairs, such as English–French and French–English, parallel JW300 corpora for English–Twi, French–Twi, Twi–French, Finnish–Twi, Twi–Finnish, Swedish–Twi, Twi–Swedish, Spanish–Twi, and Twi–Spanish were provided in the OPUS repository (https://opus.nlpl.eu/Opus-MT (accessed on 16 April 2023)) [29]. Due to copyright issues, the corpus is not accessible at the moment. However, the Language Technology Research Group at the University of Helsinki (https://huggingface.co/Helsinki-NLP (accessed on 16 April 2023)) provides large machine translation Twi models called OPUS-MT models, which were trained on the JW300 data when they were still available in the OPUS repository plus other text data.

The authors of [39] present the Twieng corpus, a small English–Twi parallel corpus with 5,419 sentences. The corpus is based on online news portals, Twi literature, the Ghanaian Parliamentary Hansard, the Twi–English Bible, and Social Media crowdsourcing and has a focus on socio-cultural, educational and legal issues.

The TypeCraft Akan corpus provides 80,000 parallel translations in Twi and English [40,41]. However, only the release 1.0 of the TypeCraft Akan Corpus with 669 monolingual sentences is publicly available (https://typecraft.org/w/index.php?title=The_TypeCraft_Akan_Corpus (accessed on 16 April 2023)).

The LORELEI (Low Resource Languages for Emergent Incidents) Akan Representative Language Pack (https://catalog.ldc.upenn.edu/LDC2021T02 (accessed on 16 April 2023)) contains almost 3.3 million Akan words of monolingual text translated into English, as well as 115,000 Akan words translated from English data [42–44]. However, access to this corpus is not freely available, not even for researchers.

The authors of [21] offer a corpus of 25,421 English–Twi sentence pairs whose English sentences were curated from tatoeba.org. For our work, we selected this dataset since it is large enough for machine translation, and the language is more modern compared to the other corpora.

3.2. Twi Machine Translation Systems

The authors of [45] used an English–Twi parallel Bible corpus to train transformer-based neural machine translation systems [46], statistical machine translation systems developed with the Moses toolkit [47] and sequence-to-sequence recurrent neural network-based machine translation systems. The transformer-based neural machine translation systems outperformed the statistical machine translation system and the recurrent neural network-based system in both directions.

Furthermore, GhanaNLP, an open-source initiative and the company Algorine, introduced the Khaya Android app which executes neural machine translation in Ghanaian languages, such as Twi, Ewe, and Ga to English [27]. There are also other Twi machine translation systems available online, but their technology is not clearly explained.

As mentioned in Section 3.1, the Language Technology Research Group at the University of Helsinki provides the OPUS-MT models (https://opus.nlpl.eu/Opus-MT (accessed on 16 April 2023)) for English–Twi, French–Twi, Twi–French, Finnish–Twi, Twi–Finnish, Swedish–Twi, Twi–Swedish, Spanish–Twi, and Twi–Spanish in addition to other language pairs at their GitHub repository [29]. English–French and French–English models are also available, which can be used in pivot machine translation systems. The OPUS-MT models are transformer models which were pre-trained on datasets at the OPUS repository [29].
with the help of the MarianMT toolkit [48]. The company John Snow Labs uses these pre-trained models and provides fine-tuned machine translation models with the Spark NLP package [49]. Additionally, [21] fine-tuned the English–Twi OPUS-MT model for their English–Twi machine translation experiments.

Since they have proven to be successful as base models, we also used the Twi–French, French–Twi, English–Twi, English–French, and French–English OPUS-MT models provided by [29] as base models in our machine translation systems. We fine-tuned these models with the training and validation sets of our corpus of 10,708 parallel sentences, which we introduce in the next section. Since Google’s Neural Machine Translation System recently included Twi [28], we also compare our results to the Google Translate output.

4. Our Parallel Twi–French–English Corpus

Our parallel Twi–French–English corpus is based on a subset of the open-source English–Akuapem Twi corpus created by [21]. The benefit of this corpus is that it is large enough for machine translation, the language is more modern compared to the other corpora, and it was used to build a machine translation system by [21].

We randomly extracted 10,708 sentence pairs from this corpus and let professional translators create the French translation of these sentences. Table 1 summarizes the number of sentences, number of running words (word tokens) and number of unique words (unique words) in the resulting corpus. To contribute to the improvement of low-resource languages, we share the corpus with the research community on GitHub.

Table 1. Statistics of our parallel Twi–French–English corpus.

<table>
<thead>
<tr>
<th></th>
<th>#Sentences</th>
<th>#Word Tokens</th>
<th>#Unique Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twi</td>
<td>10,708</td>
<td>70,400</td>
<td>7,966</td>
</tr>
<tr>
<td>French</td>
<td>10,708</td>
<td>70,257</td>
<td>10,160</td>
</tr>
<tr>
<td>English</td>
<td>10,708</td>
<td>67,677</td>
<td>8,239</td>
</tr>
</tbody>
</table>

To train, tune and evaluate our machine translation systems, we split the corpus into a training set (80%), validation set (10%) and test set (10%) as shown in Table 2. Whereas the validation set was used to find the optimal model parameters in each training epoch with high BLEU scores, we used the test set to evaluate the final system.

Table 2. Distribution of training set, validation set and test set.

<table>
<thead>
<tr>
<th></th>
<th>#Sentences</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>8,566</td>
<td>80</td>
</tr>
<tr>
<td>Validation</td>
<td>1,071</td>
<td>10</td>
</tr>
<tr>
<td>Test</td>
<td>1,071</td>
<td>10</td>
</tr>
</tbody>
</table>

5. Twi Machine Translation

In this section, we will first present our three evaluation metrics. Then, we will describe the setup and the training procedure of our systems. Finally, we will report the systems’ performances.

5.1. Evaluation Metrics

As stated in [50], “although people refer to “the” BLEU score, BLEU is in fact a parameterized metric whose values can vary wildly with changes to these parameters”. Since in the related literature different implementations of BLEU are used, we report our results with three metrics. In all metrics, there are one or more references, i.e., human-translated versions of a sentence, as well as a hypothesis, i.e., a translation generated by the machine translation system. The hypothesis is compared to the reference. In all cases, higher scores reflect better translations.
5.1.1. BLEU

BLEU (Bilingual Evaluation Understudy) [51] is a precision metric which usually compares word-level n-grams [50] of the hypothesis sentence and one or more reference sentences. To obtain sentences for a useful and comparable BLEU evaluation, often text processing (e.g., normalization, tokenization, compound splitting, and the removal of case) is required for both the reference and hypothesis. BLEU scores are dependent on the translated language pair and on the settings of the parameters used to compute the BLEU score. Since no results are yet reported for Twi–French and French–Twi machine translation, we expect scores in a similar range to English–Twi BLEU scores, e.g., in [21].

Compared to the accuracy—which is used for evaluating many other machine learning tasks—BLEU is designed to be a more nuanced and sophisticated measure of translation quality. By using n-grams, it takes into account not only word-by-word matches but also higher-level aspects of the translation and sentence structure. Thus, BLEU can better capture the overall quality of a machine translation system, whereas accuracy provides only a binary assessment of correctness.

We benefited from the corpus-bleu function (https://www.nltk.org/_modules/nltk/translate/bleu_score.html (accessed on 16 April 2023)) in the Natural Language Toolkit (NLTK) [52] to compute the BLEU scores. To be comparable to related work, we used a regular 4-gram BLEU implementation with the parameters smoothing_function = 7, auto_reweigh = False, and weights = (0.25, 0.25, 0.25, 0.25).

5.1.2. AzunreBLEU

Since our corpus is based on [21]’s English–Akuapem Twi corpus, our goal was to investigate if our Twi–French and French–Twi machine translation systems are in the same performance range as their English–Twi machine translation system.

Consequently, we adapted the corpus-bleu function in the NLTK [52] with the parameter values reported by [21]. Indicating that the focus is on “adequacy” instead on “fluency” in the translations, they use the following parameters: smoothing_function = 7, auto_reweigh = True, and weights = (0.58, 0.0, 0.0, 0.0). We refer to this BLEU score variant setting as AzunreBLEU.

5.1.3. SacreBLEU

To have a score that is comparable to related work despite different tokenization and normalization schemes, [50] proposed SacreBLEU. Compared to BLEU, in SacreBLEU, the hypothesis is compared to one or more references as well. However, SacreBLEU expects detokenized hypotheses, applying its own metric-internal pre-processing [50].

To compute the SacreBLEU scores, we benefited from the sacrebleu library (https://github.com/mjpost/sacrebleu (accessed on 16 April 2023)) provided by [50]. We used the regular 4-gram implementation with the following parameters: smooth-method = add-k, force = False, lowercase = True, tokenize = intl, and bleu_effective_order = True.

5.2. Systems’ Setup

Some researchers propose cascading cross-lingual natural language processing approaches to solve the problems of low-resource languages by benefiting from models of rich-resource languages, such as English [53–57]. Consequently, as visualized in Figure 2, we investigate direct machine translation and cascading systems that use English (EN) as a pivot language for the Twi–French (TW–FR) and French–Twi (FR–TW) translations:

- In the direct machine translation systems, the texts in the source language is directly translated to the target language using a <source language>–<target language> model.
- In the cascading systems, the source language text is first translated to EN and then translated into the target language using a <source language>–EN model and an EN–<target language> model.
Figure 2. Systems’ overview.

As mentioned in Section 3.2, the pre-trained transformer-based OPUS-MT models provided by [29] proved to be successful. Therefore, we used the Twi–French, French–Twi, English–Twi, English–French, and French–English models (1) directly for the downstream task. (2) We fine-tuned these models using our 8566 parallel training sentences and 1071 parallel sentences from the validation set. Since a Twi–English OPUS-MT model is not available in the OPUS repository, we used the weights of the Twi–French OPUS-MT model as the initial weights.

To process our texts with the OPUS-MT models, we applied the tokenization provided with the OPUS-MT pre- and post-processing scripts. This tokenization is based on SentencePiece [29,58], a language-independent sub-word tokenization algorithm developed by Google. This tokenization is typically employed in neural network-based text generation systems, where the size of the vocabulary is predetermined prior to the neural model training.

All six models (OPUS_tw-fr, OPUS_tw-en, OPUS_en-fr, OPUS_fr-tw, OPUS_fr-en, OPUS_en-tw) were fine-tuned with a batch size of 8 using the Adam optimizer [59] with a learning rate of $2 \times 10^{-5}$. For each machine translation system, we determined the number of epochs which gave the best SacreBLEU scores (as defined in [50]) on the validation set. While for OPUS_tw-fr, OPUS_en-fr, OPUS_fr-tw, OPUS_en-tw 16 epochs were sufficient, 32 epochs were required for OPUS_tw-en and 24 epochs for OPUS_fr-en.

Since Google’s Neural Machine Translation System recently included Twi [28], we additionally compare our results to the Google Translate output. For all implementations, we used Google Colab (https://colab.research.google.com (accessed on 16 April 2023)).
5.3. Results

Table 3 summarizes the systems’ BLEU, AzunreBLEU, and SacreBLEU scores. We see that systems which are fine-tuned with the particular language pair in our collected corpus—indicated with the suffix “_tuned” and the entry “our corpus”—significantly outperform the OPUS-MT models which are not fine-tuned.


<table>
<thead>
<tr>
<th>Model</th>
<th>Fine-Tuning</th>
<th>BLEU</th>
<th>AzunreBLEU</th>
<th>SacreBLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google_tw-fr</td>
<td>—</td>
<td>0.41</td>
<td>0.79</td>
<td>0.44</td>
</tr>
<tr>
<td>Google_tw-en → Google_en-fr</td>
<td>—</td>
<td>0.41</td>
<td>0.79</td>
<td>0.44</td>
</tr>
<tr>
<td>OPUS_tw-fr</td>
<td>—</td>
<td>0.22</td>
<td>0.60</td>
<td>0.21</td>
</tr>
<tr>
<td>OPUS_tw-fr_tuned</td>
<td>our corpus</td>
<td>0.36</td>
<td>0.76</td>
<td>0.37</td>
</tr>
<tr>
<td>OPUS_tw-en_tuned → OPUS_en-fr_tuned</td>
<td>our corpus</td>
<td>0.30</td>
<td>0.67</td>
<td>0.31</td>
</tr>
<tr>
<td>Google_fr-tw</td>
<td>—</td>
<td>0.37</td>
<td>0.76</td>
<td>0.34</td>
</tr>
<tr>
<td>Google_fr-en → Google_en-tw</td>
<td>—</td>
<td>0.37</td>
<td>0.76</td>
<td>0.34</td>
</tr>
<tr>
<td>OPUS_fr-tw</td>
<td>—</td>
<td>0.21</td>
<td>0.62</td>
<td>0.14</td>
</tr>
<tr>
<td>OPUS_fr-tw_tuned</td>
<td>our corpus</td>
<td>0.40</td>
<td>0.77</td>
<td>0.39</td>
</tr>
<tr>
<td>OPUS_fr-en_tuned → OPUS_en-tw_tuned</td>
<td>our corpus</td>
<td>0.44</td>
<td>0.81</td>
<td>0.42</td>
</tr>
</tbody>
</table>

As AzunreBLEU was used as an evaluation metric in related work [21], Figure 3 visualizes a direct comparison between the AzunreBLEU scores of the Google Translate systems and our systems. Our best Twi–French system is OPUS_tw-fr_tuned, the direct state-of-the-art transformer-based machine translation system, which achieves an AzunreBLEU score of 0.76. Our best French–Twi system OPUS_fr-en_tuned → OPUS_en-tw_tuned, which is a cascading system that uses English as a pivot language, results in an AzunreBLEU score of 0.81. This demonstrates that we have results comparable to [21], who reported an AzunreBLEU of 0.72 in their English–Twi machine translation system.

![Figure 3. AzunreBLEU scores of Google Translate and our machine translation systems.](image)

Table 4 lists the relative improvement of our best Twi–French and French–Twi systems that are fine-tuned with our corpus compared to the pre-trained OPUS-MT models and Google Translate on our test set: We see that leveraging our corpus to fine-tune the models has a huge impact on the systems’ performance in all three evaluation metrics.
Table 4. Difference of our best systems compared to pre-trained OPUS models and Google Translate.

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>AzunreBLEU</th>
<th>SacreBLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Δ to Google_tw_fr</td>
<td>−12.2%</td>
<td>−3.8%</td>
<td>−15.9%</td>
</tr>
<tr>
<td>Δ to OPUS_tw_fr</td>
<td>63.6%</td>
<td>26.7%</td>
<td>76.2%</td>
</tr>
<tr>
<td>Δ to Google_fr_tw</td>
<td>18.9%</td>
<td>6.6%</td>
<td>23.5%</td>
</tr>
<tr>
<td>Δ to OPUS_tw_fr</td>
<td>125.0%</td>
<td>30.6%</td>
<td>200.0%</td>
</tr>
</tbody>
</table>

While we are not able to obtain better results than Google Translate in our Twi–French machine translation task—we suspect that the Google models were trained with significantly more data—we are able to outperform Google Translate slightly in the other direction with 6.6% relatively higher AzunreBLEU scores, 23.5% relatively higher SacreBLEU scores, and 18.9% relatively higher BLEU scores.

6. Discussion

As explained in Section 4, our parallel Twi–French–English corpus is based on a subset of the open-source English–Akuapem Twi corpus created by [21]. The benefit of this corpus is that it is large enough for machine translation, and the language is more modern compared to the other corpora. It contains vocabulary that is useful in everyday conversations. Therefore, our systems perform well in such general conversations, which was our goal in this work. However, since the systems were only fine-tuned with this corpus, they have difficulty with translating technical sentences perfectly.

Figure 4 shows an example where our best French–Twi system OPUS_fr-en_tuned → OPUS_en-tw_tuned is not able to translate the English words for “artificial intelligence” and “machine translation” into Twi. As we can see, Google Translate has no problems with these technical terms, even though in this example the Google Translate translation is not perfect either. To cover technical and other domains better, our system would have to be fine-tuned with text data of the corresponding domains.

Figure 4. Example of our systems’ shortcomings in technical domains.

7. Conclusions and Future Work

French is a strategically and economically important language in the regions where the African language Twi is spoken. However, only a very small proportion of Twi speakers in Ghana speak French. Since there are hardly any machine translation systems or parallel corpora between Twi and French that cover a modern and versatile vocabulary, our goal was to extend the modern Twi–English corpus of [21]. We randomly extracted 10,708 sentence pairs from this corpus and let professional translators create the French translation of these sentences.
Furthermore, we developed machine translation systems between Twi and French. Since in the related literature, different implementations of machine translation evaluation metrics are used, we reported our results with the three metrics BLEU, AzunreBLEU and SacreBLEU. We investigated direct machine translation and cascading systems that use English as a pivot language for Twi–French and French–Twi translations. Our best Twi–French system is OPUS_tw-fr_tuned, the direct state-of-the-art transformer-based machine translation system which achieves an AzunreBLEU score of 0.76. Our best French–Twi system OPUS_fr-en_tuned → OPUS_en-tw_tuned, which is a cascading system that uses English as a pivot language, results in an AzunreBLEU score of 0.81.

After we collect a corpus with rather simple and short sentences, we plan to extend our corpus with more complex sentences of further domains. Since, to date, no language-specific algorithms for pre-processing have been investigated, future work may deal with the peculiarities of the Twi language. Additionally, to allow everyone to access our machine translation systems, our goal is to provide a web interface. Furthermore, we plan to extend the machine translation corpus with more African low-resourced languages.

Author Contributions: Conceptualization, methodology, software, validation, resources, writing, visualization: F.G. and T.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: To contribute to the improvement of low-resource languages, we share our code and our corpus with the research community: https://github.com/gyasifred/TW-FR-MT

Conflicts of Interest: The authors declare no conflict of interest.

References


53. Lin, Z.; Jin, X.; Xu, X.; Wang, Y.; Tan, S.; Cheng, X. Make It Possible: Multilingual Sentiment Analysis Without Much Prior Knowledge. In Proceedings of the IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Warsaw, Poland, 11–14 August 2014; Volume 2, pp. 79–86. [CrossRef]


Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.