Development and Validation of Parallel Corpus: A Framework for Building Bikol-Filipino Linguistic Resource

Rosel Oida Onesa, Melvin A. Ballera

Graduate Studies, Technological Institute of the Philippines, Manila, Philippines. Email: mronesa1@tip.edu.ph

The Philippines boasts a rich array of languages, each with its unique cultural significance and history. Despite their importance as a vital component of the country's linguistic heritage, the growth and development of these low-resource languages has been impeded by the lack of parallel corpus – a valuable tool in machine translation. This article presents a comprehensive process of constructing a Bikol-Filipino parallel corpus, commencing from web scraping to sentence and word alignment. The study underlines the significance of evaluating and validating the corpus to ensure its accuracy and reliability. The collected data was utilized to fine-tune a T5-Base transformer model for machine translation and subsequently assessed with the BLEU metric. The resulting score of 73.71 highlights the significance of the generated Bikol-Filipino parallel corpora, making it an invaluable asset for research and development in both languages.

Keywords: Bikol-Filipino translation, language resources, Natural language processing, parallel corpus, transfer learning.

Introduction

In today's interconnected world, data has become more accessible than ever, with the vast majority available online. However, with data coming in different formats, forms, and languages, it can be challenging to understand and analyze it all. Fortunately, machine translation has made it possible to overcome this language barrier and gain insights from data in any language. Machine translation (MT) is a process that uses computer algorithms to translate text from one human language to another automatically. The quality of MT largely depends on the use of parallel corpora, a collection of text pairs with a strict translation relationship between two languages. These text pairs consist of texts in the source language and their translations in the target language, allowing for more accurate and reliable machine translations [1]. It comprises significant bilingual knowledge, a fundamental resource for cross-language information [2] and other NLP tasks.

Parallel corpora, which are collections of texts in two or more languages aligned at the sentence or phrase level, are considered the gold standard resource for many multilingual processing applications [3]. They are particularly useful for machine translation, where the goal is to automatically translate text from one language to another. However, despite their usefulness, parallel corpora are still lacking in many Asian languages [4], which poses a challenge for researchers and developers working on improving multilingual processing. Oco and Roxas have highlighted the issue of insufficient resources as a major setback in research on Philippine languages [5]. The shortage of resources has resulted in a lag in the field, creating significant hurdles for researchers to develop innovative ideas.

This research endeavor is focused on creating a parallel corpus for two low-resource languages: Bikol and Filipino. Despite previous efforts, a formal study exploring this topic has yet to be established. The primary objective of this paper is to bridge this gap by detailing the meticulous process of constructing a Bikol-Filipino parallel corpus - from web scraping to sentence alignment and filtering. The study highlights the significance of evaluating and validating the corpus to ensure its accuracy and reliability. Ultimately, the resulting corpus will provide a crucial resource for advancing research and development in these two languages.

Related Works

The Philippines is home to a diverse range of languages, with 185 distinct languages spoken throughout the country [6]. While Filipino is the official language, several other major languages are spoken, including Cebuano, Hiligaynon, Kapampangan, and Bikolano. Each of these languages has its unique history and cultural significance, contributing to the vibrant tapestry of the Philippines' linguistic landscape. These languages are incredibly rich in their unique ways, reflecting the diverse cultural heritage of the Philippines. However, due to the scarcity of annotated data, these languages are considered low resource [7], which limits their potential for further development. In the field of Machine Translation, language pairs that have limited or insufficient parallel corpora are referred to as low-resource language pairs [8]. Conversely, high-resource language pairs, such as English and French, typically do not face any dataset size issues as researchers have created ample parallel corpora for these languages over the years.

Numerous studies have been conducted to develop corpora for Philippine languages, but the scope of these studies has been relatively limited. One study developed a Cebuano-Filipino parallel corpus by utilizing biblical and web text, using a Recurrent Neural Network (RNN) with OpenNMT sequence modeling tool in TensorFlow [9][10]. The evaluation was done based on the BLEU metric. Another study developed a Winaray language model using a Winaray monolingual corpus scraped from various websites [11]. They trained an encoder-decoder RNN model with the text corpus, and the prediction generated by the model was manually evaluated by a group of selected Winaray-speakers. Additionally, several English-Filipino parallel corpora were developed for specific purposes. For instance, a tourism-related parallel corpus was created and later used in the ASEAN Language Translation System [12]. Moreover, the Korean-Filipino Parallel Corpus Development project aims to create a vast Korean-Foreign Language Parallel Corpus to enhance AI-based Natural Language Processing (NLP) technology and improve the quality of intercultural communication and machine translation [13].

Methodology

The study comprises several crucial steps that are presented in detail in the succeeding subsections.

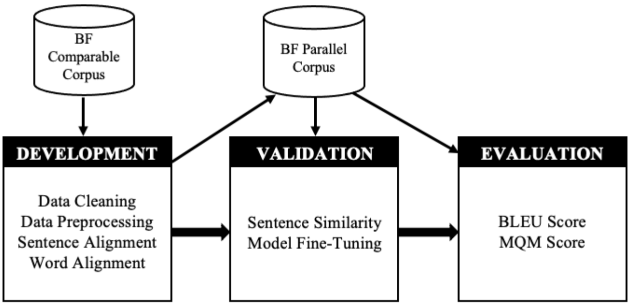


Fig. 1. Bikol-Flipino Corpus Development

Building the Bikol-Filipino Corpus

Building a machine translation requires a large amount of parallel corpus. The process of building a corpus is a time-consuming process involving significant human effort. This section discusses the meticulous process of building the Bikol-Filipino corpus.

Data Sourcing

The Bible and the internet are two significant sources of parallel data [10]. Web crawling is a crucial step in collecting parallel corpora, which involves gathering multilingual data from websites in various languages and extracting their contents [14]. There are several free tools available, such as HTTrack [15] and justtext [16], that can be used to integrate and streamline web crawling.

In this research, we utilized two distinct corpora. The first corpus was sourced from bible.com [17] and comprised of the Bikol and Filipino versions of the Bible. To extract verses, we employed the Python package Beautiful Soup [18] to scrape passages from the books of Genesis, Exodus, Leviticus, and Numbers in the Central Bikol-Marahay na Bareta Biblia for Bikol text and the Filipino-Magandang Balita Biblia (Revised) for the Filipino text. The second corpus was derived from the Bikol Wikipedia [19], containing various Wikipedia articles covering topics such as History, Typhoon, Pollution and Climate Change, and COVID-19 Pandemic. However, we discovered that only a few articles contained parallel material. To expand this corpus, we included sentences from Bikol folk songs translated into Filipino, which we gathered from various sources such as the internet and books.

Data Cleaning and Normalization

Once the dataset has been gathered, the subsequent step is to clean it up by removing any errors, inconsistencies, or irrelevant data. Gathering data from the internet can be a daunting task since irrelevant or noisy information can cloud the dataset. To ensure precision and relevance, the information gathered from the website was meticulously cleaned and preprocessed. This involved various tasks such as eliminating HTML markups, segmenting lines, and converting ASCII characters. During this process, some words or characters were converted to ASCII characters, which were either eliminated from the dataset or manually translated into their equivalent words or characters. Furthermore, words that were missing from the webpage due to being in special tags or the form of images were completed. While comparing two versions of the data, inconsistencies in terms of numeric values, like dates and measurements, were noted. To address these disparities, adjustments were made to the values, and the English version was utilized as the baseline. This ensured that the data was both accurate and uniform across all versions.

During this phase, it was noted that Bikol language names differed from their Filipino counterparts in the absence of the letter "h". Table 1 provides several examples of names.

Table I Names in the Biblical Text

Bikol	Filipino	Bikol	Filipino
Aser	Asher	Ros	Rosh
Asbel	Ashbel	Simron	Shimron
Gerson	Gershon	Suni	Shuni
Nabi	Nahabi	Tera	Terah
Palu	Phallu	Uz	Huz

This linguistic characteristic of Bikol is interesting and worthy of further exploration. In Rinconada Bikol, words often use the letter "h". To preserve Bikol names, they are left in their original form.

Sentence Alignment

This stage of developing the parallel corpus can be quite challenging and time-consuming. The collected data consists of various states, where some sentences are accurately aligned or translated, while others are not, especially in the case of biblical text. Some verses contain a different number of sentences and context, which might be because the translation of the Central Bikol-Marahay na Bareta Biblia and the Filipino-Magandang Balita Biblia (Revised), is not a direct one and could be based on the other language.

The text comprises of sentences of diverse lengths, spanning from two-worded ones to approximately sixty-worded ones. Additionally, certain verses in the Filipino edition were rendered into Bikol utilizing more sentences than their corresponding verses, and vice versa. To enhance brevity and coherence, these sentences were segmented further and aligned manually, leading to the creation of a more streamlined and reliable corpus [1].

The language editor team put in a lot of effort to ensure that every sentence was meticulously reviewed, aligned, and translated. The team comprised of native Bikol speakers, including one of the researchers, and research students, who were all also native speakers of the Bikol and Filipino language.

Word Alignment

To improve sentence alignment, we utilize a technique called word alignment, which involves aligning the corresponding words in each sentence [10]. We also make use of subword unit translation, which involves breaking down words into smaller units or not translating them at all. This process helps improve the accuracy of the alignment and ultimately leads to better translation results. The table below presents sample word alignment performed on the raw Filipino translation.

Table II Sample Word Alignment

Bikol	Filipino	
	Raw	Aligned
Imposibleng gibohon mo iyan	Hindi ninyo magagawa iyon	Imposibleng gawin mo yan
An kabilogan na bilang kan dibisyon ni Ruben minaabot sa syento singkwenta y uno mil kwatro syentos singkwenta	Ang kabuuan ng pangkat na ito ay 151,450	Ang kabuuang bilang ng pangkat ni Ruben ay umabot sa isang daan at limamput isang libo apat na raan at limampu
Ipinamidbid niya bilang tugang an saiyang agom na si Sara kaya ipinakua si Sara ni Abimelek na hade nin Gerar.	Kapatid ang pakilala niya kay Sara kaya itoy ipinakuha ni Abimelec hari ng Gerar	Ipinakilala niya bilang kapatid ang kanyang asawang si Sara
		Ipinakuha si Sara ni Abinelek na hari ng Gerar

The original text underwent multiple alignment methods, which included various translation techniques. For instance, in the first example, "hindi" was substituted with "impossible," which is a direct translation of the same term. Although the raw sentence is acceptable translation, these changes aimed to achieve a better alignment between the two languages and ensure that the intended meaning was accurately conveyed in the translated text. Furthermore, the tense of the aligned text was altered to match that of the Bikol sentence. In the second example, the numerical value was replaced with its corresponding equivalent in words to match the source sentence. Lastly, in the third example, the Bikol sentence was divided into two sentences and then translated into Filipino.

Validation of Bikol-Filipino Corpus

Corpus Similarity

The process of alignment involves comparing linguistic units of two languages to establish a translation relationship. This study seeks to assess the reliability and accuracy of parallel texts gathered from Bikol and Filipino corpora. This was achieved by comparing the original texts to the word-aligned and corrected translation datasets, followed by calculating the cosine similarity to identify any substantial variations between the raw dataset A and the word-aligned dataset B.

$$cos(A, B) = \frac{A \cdot B}{||A|| \cdot ||B||}$$

The Cosine similarity is a valuable mathematical metric that helps to gauge the similarity between two vectors. The score varies between 0 and 1, with a score of 1 signifying a high degree of similarity between the vectors. A score of 1 implies that no further adjustments or processing were made to the raw text when collected. By evaluating this score, we can acquire a more precise understanding of the parallel texts' accuracy and dependability that were collated.

Model Fine-Tuning

A parallel corpus is developed for various Natural Language Processing (NLP) applications like machine translation. The Bikol-Filipino corpus is evaluated by the researchers, by training and fine-tuning a machine translation model. Transfer learning has become a highly effective approach in the field of NLP in recent years. This technique involves training a model on a task that involves a large amount of data, before subsequently fine-tuning it for use in a separate, related task. By leveraging existing knowledge and pre-trained models, this approach can significantly improve the accuracy and efficiency of NLP systems [20].

Refining a model through fine-tuning involves leveraging a pre-existing model to train a new one. This technique, known as transfer learning, can enhance the original model's efficacy by maximizing its previous knowledge and learning from the data it has already processed. Recently, pre-training a model on a job with ample data has gained popularity. The pre-training should imbue the model with multifaceted knowledge so that it can be employed for various tasks in the future [20].

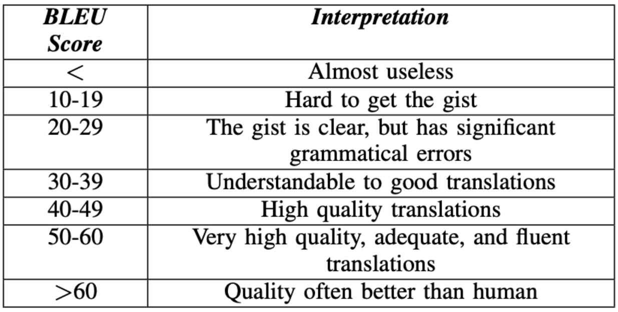
In conducting this study, we utilized a pre-trained transformer model available on the Hugging Face Hub [21]. Our team fine-tuned the model using the collected corpora and thoroughly evaluated its performance. Known as the Text-to-Text Transfer Transformer or T5 [22], this particular encoder-decoder language model is highly adept at handling a wide range of natural language processing tasks, including but not limited to machine translation, document summarization, question answering, and classification tasks [20].

Evaluation of Bikol-Filipino Corpus

BLEU Metric

The BLEU score is a commonly used metric to assess the quality of natural language processing (NLP) models. It is a numerical value that represents how closely the generated text matches the reference texts. BLEU stands for Bilingual Evaluation Understudy, and it provides a way to measure the effectiveness of NLP models [22]. The score ranges from zero to one, where a score closer to one indicates a better match between the generated text and the reference texts. In the evaluation of the fine-tuned model, the BLEU metric was employed, and the resulting score was interpreted using a specific table to determine the model's quality.

Table III BLEU Score Interpretation [22]



BLEU Score	***Interpretation***
<	Almost useless
10-19	Hard to get the gist
20-29	The gist is clear, but has significant grammatical errors
30-39	Understandable to good translations
40-49	High quality translations
50-60	Very high quality, adequate, and fluent translations
>60	Quality often better than human

Result and Discussion

Bikol-Filipino Corpus

A significant effort was undertaken to compile and merge 15,476 sentence pairs in Bikol and Filipino languages, resulting in 7,738 sentence pairs of impeccable quality. The process involved a meticulous curation and alignment of texts in both languages, which was carried out by native speakers of Bikol and Filipino languages who served as language editors.

Table IV Bikol-Filipino Corpus

Corpus Source	Sentence Pair	Bikol-Filipino
Biblical Text	7,385	14,770
Wikipedia Text	192	384
Folk Songs	161	322
Total	7,738	15,476

Additionally, the Bikol and Filipino corpora have produced 5,603 and 6,653 unique tokens, respectively. The top ten most frequently occurring words in both corpora are presented in Table 5.

Table V Top Occurring Words Common in the Corpora

word	Bikol	Filipino
na	5485	3586
sa	5229	5690
mga	3134	2757
si	1364	1087
ni	1154	1584
niya	915	830
siya	806	565
ko	676	608
mo	640	586
ako	523	403

This table presents a comparison of Bikol and Filipino languages, highlighting their shared vocabulary and parallel grammatical structures. Notably, both languages employ identical pronouns and connectors such as "na," which can signify "already" or serve as a conjunction. Another frequently used term is "sa," a versatile preposition in Filipino that encompasses many of the prepositions found in English. In the Bikol language, the word "asin" is commonly used to mean "and.” This conjunction serves to connect two or more words or phrases together. On the other hand, in the Filipino language, the word "ang" is frequently used as an article to introduce a particular noun.

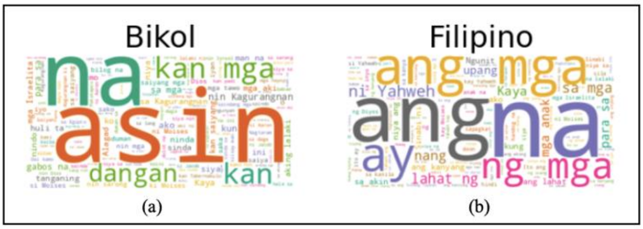


Fig. 2. Bikol-Flipino Corpus Development

Figure 2 displays a Word Cloud highlighting the words most commonly used in both the Bikol and Filipino corpora. Notably, there is a significant overlap of frequently used words between the two languages. For example, the words “na” and “mga” are used in both languages. The Bikol word "kan" has the same meaning as the Filipino word "ng", while "kan mga" in Bikol is equivalent to "ng mga" in Filipino.

Bikol language distinguishes itself from Filipino by the usage of shortened words. The Bikol language lacks numerous abbreviated words that are frequently used in Filipino. For example, the words "siya ay" and "kanya at" are commonly shortened to "siya'y" and "kanya't" in Filipino, while these abbreviations are not utilized in Bikol. Filipino and Bikol are two languages belonging to the Austronesian language family. However, while Filipino has a more flexible sentence structure, Bikol stands out for its verb-initial or predicate-initial sentence structure. This means that the verb or predicate comes at the beginning of the sentence, which is a prominent feature of Bikol grammar. Additionally, certain Bikol words such as "purpura", which translates to "ube" in Filipino or "purple" in English, are no longer in frequent use.

Validating Corpora

Corpus Similarity

Utilizing Cosine Similarity, we performed a similarity analysis to ascertain any notable distinctions between the raw and aligned subword translations in our datasets. However, given that there were varying quantities of sentences in our biblical text, Wikipedia, and Bikol folk song datasets, we opted to randomly select 150 sentences from each corpus for examination.

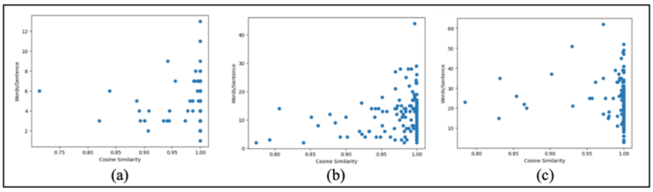


Fig. 3. Similarity value between raw and cleaned corpora

In order to get a better understanding of the similarity between different datasets, Figure 3 provides a visualization of the Cosine Similarity scores for three distinct corpora. These corpora include (a) Bikol folk song, (b) Biblical text, and (c) Wikipedia source. The y-axis of the figure represents the number of words per phrase, while the x-axis denotes the similarity score between the corpora. This score is calculated based on the similarity between raw and word-aligned sentences, and is measured on a scale from 0 to 1.

The sentences are rated on a scale from 0 to 1, with 0 indicating no similarity and 1 indicating high similarity. Figure 3b displays the varying levels of similarity found within the Biblical texts. The raw biblical corpus is a comparable corpus, which means that significant corrections and alignments are necessary to obtain acceptable sentence pairs. In contrast, the Wikipedia source dataset stands out with an impressive similarity score of 98%, indicating that minimal changes were required for near-perfect alignment and translation between sentences. Put simply, the dataset obtained from Wikipedia provides an almost perfect translation of Bikol to Filipino sentences.

Model Fine-Tuning

The dataset was split into three sections - training, validation, and testing. To optimize a pre-trained transformer model T5-Base from Hugging Face, the dataset was tokenized. Throughout the ten epochs of training, the model was fine-tuned using a learning rate of 0.001, a batch size of 128, and a weight decay of 0.1. Essentially, this entailed re-exposing the model to the dataset ten times, with the learning rate regulating the rate at which the model adapts to the data, the batch size determining the number of samples processed in each iteration, and the weight decay preventing overfitting by adding a penalty term to the loss function.

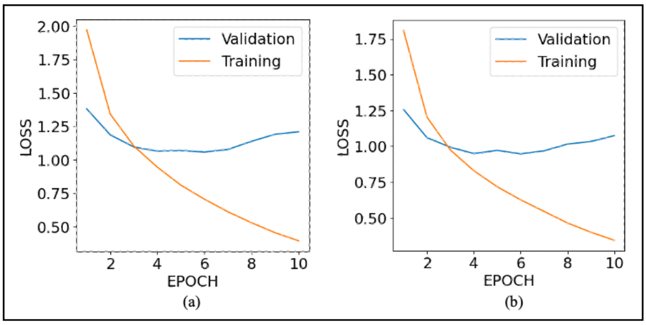


Fig. 4. (a) Bikol-Filipino (b) Filipino-Bikol

The graph depicted in the figure showcases the training process of a model that has been trained to translate from Bikol to Filipino and vice versa. The x-axis displays the number of epochs, and the y-axis presents the loss. Graph (a) represents the Bikol to Filipino translation, and it shows that the training loss consistently decreases, indicating that the model has successfully learned from the training dataset. The validation loss also experiences a gradual decline, with a slight increase observed from epoch 7. Similarly, graph (b), representing the Filipino to Bikol translation, mirrors the same scenario, demonstrating that the model has effectively learned to translate between these two languages.

Evaluation

In translating Bikol to Filipino, the fine-tuned model achieved an average BLEU score of 73.71, and in translating Filipino to Bikol, it scored 72.40. These results indicate that the translations are of high quality. Table 6 presents a sample of the test results, providing further insight into the performance of the model.

Table VI Bikol to Filipino Translation

Input	Target	Prediction	BLEU
Ginibo ni Moises an gabos siring sa ipinagboot nin Kagurangnan	Ginawa ni Moises ang lahat ayon sa iniutos ni Yahweh	Ginawa ni Moises ang lahat ayon sa utos ni Yahweh	91.00
Mantang nagtutukaw siya duman nagpoon siyang maghibi	Habang umuupo siya roon ay nag simula siyang umiyak	Nagsimula siyang umakyat sa kanyang nais na naghimagsik siya	43.64
Takopan nindo an Kaban nin Tipan	Takpan ninyo ang Kaban ng Tipan	Takpan ninyo ang Kaban ng Tipan	100.00

In the first test, a small error was detected in the output where "ipinagboot" was translated to "utos" with a slightly incorrect tense. In the second test, some words in the translation were accurate, including "siyang" and "nagsimula" as the translations for "nagpoon." However, despite these correct words, the translation's overall context and meaning were incorrect. Nonetheless, the last example was accurately translated with a perfect score of 100.

Table VI Filipino to Bikol Translation

Input	Target	Prediction	BLEU
Ikaw lamang ang makakalapit sa akin huwag lalapit ang iba	Ika sana an magdolok sako dai magrani an iba	Ika an madolok ko dai nanggad iyan paglingkod an ibang lalaki	42.36
Tunay na ikaw ay laman ng aking laman at dugo ng aking dugo	Tunay nanggad na laman ka nin sakong laman asin dugo nin sakong dugo	Sa ika sana an aki sagkod an dugo kan sakong dugo	38.44
Takpan ninyo ang Kaban ng Tipan	Takopan nindo an Kaban nin Tipan'	Takopan nindo an Kaban nin Tipan'	100.00

In the first sample test of Filipino to Bikol translation, some words including "ika," "madolok," and "ibang" were translated correctly but had issues with grammar and placementThe second sample test had an incorrect translation, with only the last part accurately translated as "dugo kan sakong dugo." Nevertheless, the third instance was flawlessly translated and achieved a BLEU score of 100, indicating a high level of accuracy in the translation.

Conclusions and Recommendation

This study marks the beginning of an important effort to create a parallel corpus for the Filipino and Bikol dialects, which are both low-resource languages. The researchers have detailed the comprehensive development process for this corpus, which is crucial for advancing machine translation and natural language processing (NLP) applications for these languages. The quality of the resulting corpus was assessed using a transformer model that was fine-tuned and measured using the BLEU metric. In order for the developed corpus to achieve the status of a gold standard, it is highly recommended that it undergoes a thorough manual evaluation by language experts. This evaluation process will not only enhance the precision and accuracy of the corpus but will also help to eliminate any potential errors, making it a reliable and trustworthy source for linguistic analysis.

The researchers are hopeful that this work will pave the way for future advancements in machine translation and other NLP applications for not only the Bikol language but also other Filipino languages.

Acknowledgments

We would like to extend our deepest gratitude to the CSPC AI Research Center for Community Development (AIRCoDe) for their invaluable and extensive resources in providing us with the help we needed in our project. We would also like to express our heartfelt thanks to Jazmine Claire Bernardino and Erika Sheena Monte, our exceptional language editors, for their tireless dedication and unwavering patience in ensuring the quality of our work. Lastly, we would like to acknowledge the remarkable support and significant contribution of the TIP Graduate Studies and CSPC College of Computer Studies, who played an instrumental role in the success of our project.

References

[1] Z. Zong and C. Hong, “Research on alignment in the construction of parallel corpus,” in J. Phys.: Conf. Ser. 1213 042003, 2019. [Online]. Available: https://doi.org/10.1088/1742-6596/1213/4/042003

[2] S. Torres-Ramos and R. E. Garay-Quezada, “A survey on statistical- based parallel corpus alignment,” in Research in Computing Science 90, 2015, pp. pp. 57–76; rec. 2015–01–20; acc. 2015–03–10.

[3] B. Cartoni, S. Zufferey, T. Meyer, and A. Belis, “How Comparable are Parallel Corpora? Measuring the Distribution of General Vocabulary and Connectives,” 2011.Retrieved from https://publications.idiap.ch/downloads/papers/2011/Cartoni_BUCC_2011.pdf

[4] S. Zhu, C. Mi, T. Li, Y. Yang, and C. Xu, “Unsupervised parallel sentences of machine translation for asian language pairs,” ACM Trans. Asian Low-Resour. Lang. Inf. Process., vol. 22, no. 3, mar 2023. [Online]. Available: https://doi.org/10.1145/3486677

[5] N. Oco and R. Roxas, “A survey of machine translation work in the Philippines: From 1998 to 2018,” in Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018). Boston, MA: Association for Machine Translation in the Americas, Mar. 2018, pp. 30–36. [Online]. Available: https://aclanthology.org/W18-2204

[6] D. M. Eberhard, G. F. Simons, and C. D. F. (eds.), “Ethnologue: Languages of the world. Twenty-sixth edition. Dallas, Texas: SIL International,” 2023. [Online]. Available: https://www.ethnologue.com/country/PH

[7] L. Miranda, “Towards a Tagalog NLP pipeline — ljvmiranda921.github.io,” https://ljvmiranda921.github.io/notebook/2023/02/04/tagalog-pipeline/.

[8] C. Zhou, “Building a Catalan-Chinese Parallel Corpus from Wikipedia for Use in Machine Translation,” 2022. Retrieved from https://repositori.upf.edu/bitstream/handle/10230/54140/Zhou_2022.pdf?sequence=1&isAllowed=y

[9] J. L. Fernandez and K. M. M. Adlaon, “Exploring word alignment towards an efficient sentence aligner for filipino and cebuano languages,” in LORESMT, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252818943

[10] K. M. M. Adlaon and N. Marcos, “Building the language resource for a cebuano-filipino neural machine translation system,” in Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval, ser. NLPIR ’19. New York, NY, USA: Association for Computing Machinery, 2019, p. 127–132. [Online]. Available: https://doi.org/10.1145/3342827.3342833

[11] C. B. S. Fernando E. Quiroz, Jr. and J. A. Sabonsolin, “Building the waray-waray neural language model using recurrent neural network,” Mindanao Journal of Science and Technology, vol. 21, no. 1, jun 2023.

[12] C.PonayandC.Cheng,“23.buildinganenglish-filipinotourismcorpus and lexicon for an asean language translation system,” 06 2015.

[13] “Korean-Filipino Parallel Corpus Development - Department of Linguistics - UP Diliman — linguistics.upd.edu.ph,” https://linguistics.upd.edu.ph/korean-filipino-parallel-corpus-development/, [Accessed 11-02-2024].

[14] A. Barro ́n-Ceden ̃o, C. Espan ̃a-Bonet, J. Boldoba, and L. M. i Villodre, “A factory of comparable corpora from wikipedia,” in BUCC@ACL/IJCNLP, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:15575006

[15] G. Dog ̆ru, A. Mart ́ın-Mor, and A. Aguilar-Amat, “Parallel corpora preparation for machine translation of low-resource languages: Turkish to english cardiology corpora,” 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:204762642

[16] J. Pomika ́lek, “Removing boilerplate and duplicate content from web corpora,” Ph.D. dissertation, Masaryk university, Faculty of informatics, Brno, Czech Republic, 2011.

[17] “ReadtheBibleonline.AfreeBibleonyourphone,tablet,orcomputer. — bible.com,” https://www.bible.com/.

[18] “Beautiful Soup Documentation ful Soup 4.12.0 documentation https://www.crummy.com/software/BeautifulSoup/bs4/doc/.

[19] Wikipedia, “Tataramon na Bikol Sentral,” https://bcl.wikipedia.org/wiki/Tataramon na Bikol Sentral.

[20] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” Journal of Machine Learning Research, vol. 21, no. 140, pp. 1–67, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html

[21] “Hugging Face – The AI community building the future. — hugging- face.co,” https://huggingface.co/, [Accessed 19-02-2024].

[22] G. T5, “google-t5/t5-base · Hugging Face — huggingface.co,” https://huggingface.co/google-t5/t5-base, [Accessed 19-02-2024].

[23] “Evaluating models — AutoML Translation Documentation —Google Cloud — cloud.google.com,” https://cloud.google.com/translate/automl/docs/evaluate, 11-02-2024].