Research Article

Journal of Mathematical Techniques and Computational Mathematics

Twieng: A Multi-Domain Twi-English Parallel Corpus for Machine Translation of the Twi Language, A Low-Resource African Language

Gabriel Kwadwo Afram *, Asubam Wejori Benjamin, Adekoya Felix Adebayo

Gabriel Kwadwo Afram University of Energy and Natural Resources, Sunyani, Ghana

***Corresponding author**
Gabriel Kwadwo Afram University of Energy and Natural Resources, Sunyani, Ghana.

Submitted: 20 Jul 2022; **Accepted:** 22 Sep 2022; **Published:** 29 Oct 2022

Citation: Afram, G. K., Benjamin, A. W., Adebayo, A. F. (2022). Twieng: A Multi-Domain Twi-English Parallel Corpus for Machine Translation of the Twi Language, A Low-Resource African Language. *J Math Techniques Comput Math,* 1(1), 48-57.

Abstract

A Twi-English parallel corpus is certainly an important resource for Machine Translation of Twi (ISO 639-3), a Low- Resource Language (LRL) which is mainly spoken in Ghana and Ivory Coast. Currently large-scale multi-domain Twi- English parallel corpus is still unavailable partly due to the difficulties and the arduous efforts required in its design. A digital Twi lexicon curated purposely for linguistic research is also not available. In this paper, we present TWIENG – Twi English corpus, a large-scale multi-domain Twi-English parallel corpus and Twi lexicon, a digital Twi Dictionary. We discuss the data collection methodology, translation, alignment and compilation of the Twi-English parallel sentences and the technology we used to compile and host the corpus. Today's parallel corpora are crawled from the web using web crawlers, the sentence pairs are processed, aligned, tokenized and compiled to create the corpus. We crawled English sentences from Ghanaian indigenous electronic news portals, Ghanaian Parliamentary Hansards, standard literature and also used crowdsourcing. The sentences are translated by professional translators and linguists, then aligned, tokenized and compiled. The corpus is curated using the sketch engine, a corpus manager and analysis software developed by Lexical Computing Limited. The corpus is manually evaluated by Twi professional linguists. The Corpus has 5,419 parallel sentences which were curled from local news portals, Ghana Parliament Hansard, The New Testament of the Twi Bible and through crowdsourcing via social media sites.

CCS CONCEPTS • Computing Methodologies • Artificial Intelligence • Natural Language Processing

Additional Keywords and Phrases: Twi, Parallel Corpus, Tokens, Sketch Engine, Word Sketch, Parallel Concordance.

ACM Reference Format

Afram K. Gabriel, TWIENG: A Multi-Domain Twi-English Parallel Corpus and E-Dictionary for Machine Translation of Twi Language, a Low-Resource African Language. ACM Trans. Asian Low-Resource Language Information Processing.

Introduction

Parallel Corpora which consist of text and their translation aligned side by side are undeniably the fundamental resources for Natural Language Processing (NLP), especially Machine Translation (MT) [2,12,16]. The world's advanced Languages have numerous Parallel, bilingual and Monolingual Corpora which makes it easy to develop state-of-the art (SOTA) MT systems [5].

Africa is home to about 2144 distinct living languages out of the 7111 living languages of the world [31]. Among the numerous African Languages, Yoruba, Swahili, Zulu and Igbo are the languages with a well-developed digital corpus [13,26,30]. The High Resourced Languages (HRLs) and some few African Languages have monolingual, bilingual and parallel corpora freely available online as open-source [33,34,37]. However, there are no standard corpora for Twi apart from the JW300 corpus which is ideologically skewed due to its maximal reliance on the Bible [2]. Twi rather has some standard literature like dictionaries, the Holy Bible, story books and published articles on aspects of the language freely available online and offline but it has no standard curated, annotated and POS tagged corpus. It is refreshing however, to note that some researchers like initiated a bold move to create a language dataset and resources for the Low Resource African Languages (LRALs) but the project has been stalled [8, 29]. This perhaps is the reason why there is no much research into MT of Twi. Considering the fact that any MT model would need enormous data in the form of parallel corpora to train the model which is conspicuously missing for Twi.

The aim of this paper is to build a digital Twi-English parallel corpus of about 5k sentences including a Bible, parliamentary Hansard, medical, news and social media crowdsourced sub corpora which are searchable, scalable and could also be used for Natural Language Processing, especially Machine Translation (MT).

We used the Sketch Engine a corpus curation and analysis software to build our corpus [18,20]. We took advantage of the One-Click dictionary feature and API in Sketch engine to create a lexicon via Lexonomy a free dictionary curation software by Lexical computing1 [23].

The parallel corpus is intended to be used in NLP such as MT in order to give substantial support to computational linguistic research related to Twi. TWIENG would be open sourced which can be used freely by the MT research community. The corpus would be automatically and manually aligned.

Twi Language

Twi, which is also known as *Akan kasa* has about 18 million native speakers and about 45% of Ghanaian speak Twi as their first language. However, about 80% of Ghanaian speak Twi as their first and second language [28]. Around 41% of people in southern Ivory Coast also speak Twi. Other countries, like Jamaica and Suriname, also have people who know and speak Twi. Twi is one of the dialects of the *Akan kasa*. Akuapem Twi, Ashanti or Asante Twi, Fantse (Mfante, Fante, Fanti) and Bono are the various dialects of Akan [11]. The research focuses on the Asante Twi but there is a high level of mutual intelligibility between the various dialects [56].

Twi is under the Kwa subdivision of the Niger-Congo group of the languages of Africa. Twi is a tonal language that involves high, mid and low tones. The meaning of each word changes when you change the tone of the syllables Twi can be written through a common script that was created by the Bureau of Ghana Languages [11]. It has 22 alphabets which consist of 7 vowels and 15 consonants.

The Twi Orthography

[1] https://www.lexicalcomputing.com/

There are 22 alphabets of the Twi Language. (a, b, d, e, ɛ, f, g, h, i, k, l, m, n, o, ɔ, p, r, s, t, u, w, y) of which 7 are vowels (a, e, i, o, u, ɔ, ɛ) and the remaining 15 namely (b, d, f, g, h, k, l, m, n, p, r, s, t, w, y) are consonants[11,56].
C, J, V and Z are used, but only in loanwords. There are 7 vowels, 10 diphthongs and 15 consonants in Twi2.

Related Work

Machine Translation (MT) has achieved SOTA performance in recent years for a few High Resource Languages (HRLs) probably due to the readily availability of parallel corpora and efficient machine translation models such as the Transformer Architecture and its variants [11,14,15,17,24,27,32,36]. HRLs like English, French, Spanish etc. make up about only 2.5% of the world living languages [25]. These languages are highly studied, researched, funded and used for NLP especially MT primarily due to the availability of datasets and tools such as language corpora coupled with efficient NMT models [8,11,32,36]. For example, Opus and sketch engine [6,27] has a large database of parallel corpora for the HRLs but a handful for the LRLs like Twi.

Low Resource Languages (LRLs) on the contrary, can be implied as less studied, resource scarce, less computerized, less favored, less commonly studied, or lower accessed [22,35]. These languages, lack sufficient parallel sentence pairs in order to effectively train the language models for machine translation. This is as a result of the difficulty in obtaining resources and funding for building tools and datasets for these LRLs [25].

A parallel corpus consists of text placed alongside its translation or translations. Parallel corpora are used to train MT models. There are several parallel corpora such as MIZAN: A Large Persian-English Parallel Corpus, a parallel Corpora for Indian Languages, Arabic-English Parallel Corpus, Bianet: A Parallel News Corpus in Turkish, Kurdish and English [3,4,16,31]. The JW300 parallel corpus which consists of about 300 langage pairs among others [2].

The Crubadan project attempted to build an Akan (Twi) parallel corpus by gathering 547,909 Twi words from 176 documents crawled from the web [44]. Facebook uses the Translate Facebook App3 to crowdsource from various translators around the world to translate Facebook to their languages. They translate text ranging from Facebook features and words relating to the language under focus. This helps Facebook to build a corpus which would be used in building a translator for the language. These tells us that there are not many Twi words on the web that can be crawled for the purposes of MT. The LORELEI (Low Resource Languages for Emergent Incidents) program established by the Defense Advanced Research Projects Agency (DARPA) under the auspices of Linguistic Data Consortium (LDC) of the University of Pennsylvania was designed to pursue research and development of more effective language technology, while eliminating the current reliance on manually-translated, manually-transcribed, or manually-annotated corpora [13]. The LORELEI program selected 32 representative languages and 12 incident languages for the study out of the over 6600 LRLs of the world. These languages included Hausa, Yoruba, Twi, Wolof, Somali, Swahili and Zulu which are all African languages [13]. African based researchers have also taken the initiative to bring African LRLs into the limelight. Deep Learning Indaba4 is a research group that aims at building machine learning tools for African Languages. MASAKHANE5 group is a research effort for natural language processing targeting African languages [25].

It is open source, spans across the African continent and distributed with online repository of various resources. MASAKHANE has developed 38 unique language pairs and 45 benchmarks[6]. However, we find it intriguing that there is no single language pair for the numerous Ghanaians.

2 https://www.omniglot.com/writing/akan.htm
3 https://www.facebook.com/translations
4 https://deeplearningindaba.com/2020/
5 https://www.masakhane.io/
6 https://github.com/masakhane-io/masakhane-mt/blob/master/language_pairs.md

Languages apart from the JW300 English–Twi pairs which can be accessed from the Opus repository [2, 5]. The JW300 corpora are however religiously skewed and hence biased ideologically due to its reliance on the Bible Text [2]. Another Akan/Twi cor-

pus worth noting is the Typecraft Akan corpus which has only 1,906 phrases [9].

Objectives

To the best of our knowledge there is no readily available Twi parallel open-source general purpose heterogeneous corpus ever developed for the purposes of Natural Language Processing (NLP), especially Machine Translation (MT) that spans all genre of the Ghanaian Twi speaking society and culture. We therefore present;

1. TWIENG: A Twi-English parallel corpus as our contribution to Twi-English Machine Translation Research. The TWIENG Corpus, is a manually aligned corpus of 30k sentence pairs with 1.5 million tokens which is freely available on sketch Engine and our GitHub repository7 for non-commercial use based on the CC BY-NC-SA 4.08 Licence.
2. Four sub corpora; namely TwiEng web sub corpus, TwiEng news sub corpus, TwiEng Ghana Parliamentary Hansard sub corpus, TwiEng New Testament Bible sub corpus and TwiEng crowdsourced social media sub corpus.
3. Twi-English Bilingual Lexicon; a parallel dictionary of Twi and English.
4. Analysis of various features of the TWIENG corpus which include; Word sketch, Parallel concordance, N-grams, and Word list.

Methodology

The aim of our paper is to create a novel Twi-English (TWIENG) parallel corpus using a multi-domain data source from online news portals, Twi literature, Ghanaian Parliamentary Hansard, Twi-English Bible, Social Media crowdsourcing etc. These sources are chosen because the contents cut across the Ghanaian culture and social life and are open sourced. We downloaded the news archives in English from the major digital news hubs, excerpts of the parliamentary Hansard were also crawled from the official Ghana Parliament website9, Twi articles were also downloaded together with the various literature. Apart from the Twi Medical Glossary and the Twi-English Bible which have already been translated, the rest of the text were only in English [38]. We therefore used the methodology suggested by [23,29] to create the TWIENG corpus.

Corpus Preparation

The parallel text required for building a modern digital parallel corpus are usually crawled from publicly available data online using web crawlers [25, 28]. Despite our thorough search for Twi-English parallel text online, our search could not gather enough text to build the large parallel corpus we intended. The lack of enough Twi-English parallel text on the web is as a result of the fact that, English is the lingua franca in Ghana and the official language used by Ghanaians for communication online. Twi is mainly used unofficially even though it is written and studied in schools. It was not feasible to crawl our parallel data from the web, even though we had a few of the Twi texts from the web, these texts were not aligned with English.

We therefore decided to use two main approaches to collect our data; 1. auto crawling of the few Twi-English parallel sentences we came across on the web and 2. manually gathering our own English sentences and translating them into Twi by professional translators based on standard literature, online digital media portals, Twi text books and story books as

7 https://github.com/gkafram/TwEng-corpus
8 https://creativecommons.org/licenses/by-sa/4.0/
9 https://www.parliament.gh/docs?type=HS

alluded to and previously used by [23,33]. Crowdsourcing for Twi-English sentence pairs via social media was also used. A Google form was designed and the link shared among language enthusiast on social media and students studying Twi. Their responses were collected and analyzed and aligned using MS excel.

Parallel corpora are gathered from the web by crawling sentence pairs. This is true for the HRLs, contrary many LRLs are deficient in this regard. Nevertheless, we still needed to crawl the monolingual data from the web. There are various tools for crawling data from the web, these include Spider Ling, which focus the crawling of the text rich parts of the web and maximize the number of words in the final corpus per megabyte downloaded [7, 49]. Beautiful Soup[10]- a Python package for parsing HTML and XML documents. It creates a parse tree for parsed pages which is then used to mine data from HTML, which is handy for web scraping. Web Scraper11 is a free to use Google chrome extension that is used to scrap web contents. These resources were used to crawl the text from the web since they are free to use.

Majority of the texts were only in English. The obstacle therefore was that we could not get a pre-aligned Twi-English text pair. The onus therefore lied on us to translate the English text to Twi text and align them.

We crawled some free article from the web including Ghanaian parliamentary Hansard, Myjoyonline news archives, adomfmonline news archives, peacefmonline news archives, Ghanaweb news archives and Citinews news archives, books, dictionaries, TwiWiiki, Twi Medical Glossary, the Universal Declaration of Human rights as indicated in table 3 below [34, 52, 53]. The Twi-English New Testament Bible was also of immerse use due to its freely availability in pdf format which was downloaded from the JW website12. The Ghanaian parliamentary Hansard gave the corpus a heterogeneous character due to its focus on socio-cultural, educational and legal issues.

10 https://www.crummy.com/software/BeautifulSoup/
11 https://www.webscraper.io/
12 www.Jw.org

Table 1: Overview of the sources of data for the TWIENG Corpus and number of sentence pairs.

Document Name	Source	No. of docs	Sentence pairs
Twi Medical Glossary	[34]	1	420
Ghana Parliament Hansard	[49]	10	200
Myjoyonline Archives	[50]	10	155
Adomonline Archives	[51]	10	250
Peacefmonline Archives	[52]	10	140
Citinewsroom Archives	[53]	10	250
Ghanaweb archives	[51]	10	300
Daily Graphic Archives	[54]	10	358
English-Twi NT Bible	[55]	1	1330
UDHR	[52,53]	2	164
English-Twi Dictionary	[57]	1	385
Crowdsourced Twi-English sentence pairs	[58]	1	1,325
Total		76	5,419

Text Crawling, Translation and Alignment

Raw text crawling: The raw text was extracted from HTML files from the various websites as indicated in table 1 above with the BeautifulSoup script that makes use of the HTML:Parser module. Spiderling and web scrapper were also used. Sentence Translation: The monolingual texts were translated into Twi by Professional Twi Translators. Ten Professional linguists and translators were tasked to do the translation and alignment. This was a humongous task due to the large number of sentences we were working with 1.3k Twi-English sentence pairs were crowd-sourced via a google form, this added a lot of diversity to the corpus since these sentences covered various themes.

Sentence Alignment

A well-developed corpus is the one that has proper alignment of the HRLs and LRLs sentence pairs. The Twi sentences were manually aligned with the English sentences using a spreadsheet program, MS excel was the best choice due to its availability and cost free. The holy Bible was aligned at the verse level. For the documents downloaded from the web, they were aligned at the paragraph level. The TWIENG corpus consist of 5,419 sentence pairs and over 144k tokens.

Conceptual Framework of the Twieng Corpus

The data crawled from the web was prepared and fed into the Sketch Engine as shown in figure 3.1 below.

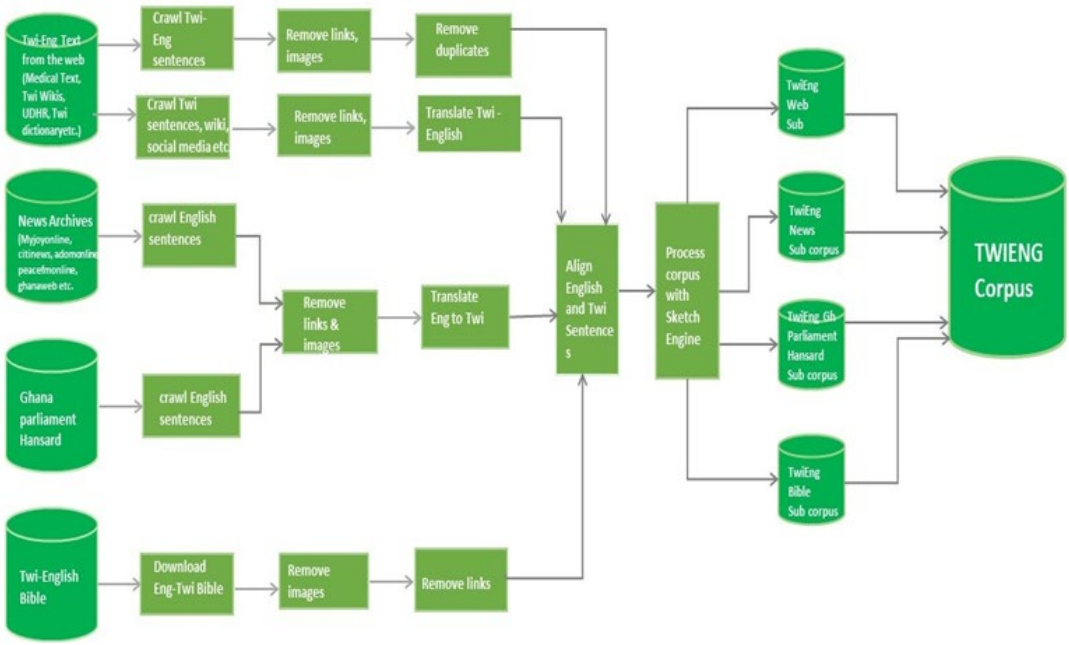

Figure 3.1: Conceptual framework of the TWIENG Corpus

Data Cleaning

The boilerplate removal tool, just text was implemented to remove any unwanted portions of the text, such as hyperlinks and menus, advertising, legal text, tabular data, icons, social media handles and any other types of text unsuitable for linguistic analysis so that they can be included in the corpus [17, 32, 41]. To clean the data, we removed all double white spaces and special characters apart from the accepted Twi orthographic characters. The articles titles, URLs, images, tables and hyperlinks were deleted from the crawled sentences.

Deduplication

Sketch engine has a build in tool that de-duplicate the whole corpus content. Both perfect duplicates as well as near duplicates are removed so that only one instance of each text is maintained. We adjusted the parameters of the tool to our preference.

Tokenization, lemmatization and Tagging

The text was then tokenized using a tokenizer which is built into the Sketch engine. There is specific tokenizer for the supported languages and also a universal tokenizer for unsupported languages. The universal tokenizer only recognizes whitespace characters as token boundaries ignoring any language specific rules. The corpus was further lemmatized by assigning the base form to each word form. In future we expect to POS tag the TWIENG corpus.

Experimental Setup and Results

The corpus was designed based on the methodology by [25, 28]. The sketch engine, a corpus curation and analysis tools were used to create and host the TWIENG corpus.

Algorithm to Prepare the Twieng Corpus with Sketch Engine

The algorithm to prepare a text corpus with sketch engine is outlined below.

ALGORITHM: preparing text corpus with sketch engine.

1. Prepare the source data.
2. Prepare the corpus configuration file if required.
3. Prepare the sub corpus configuration file, if you need to compile a sub corpus.
4. Prepare or reuse a word sketch definition file if you require word sketches or thesaurus.
5. Compile (index) the corpus.
6. Verify corpus consistency, integrity and completeness.

Statistics of the Twieng Corpus

The TWIENG corpus statistics are shown in table 4 below.

Table 4: TWIENG Corpus Statistics and size.

Language	Tokens	Words	Sentences
English	60,187	48,220	5,419
Twi	63,873	50,664	5,419
Total	124,060	98,884	10,838

The Twieng Lexicon

The OneClick dictionary feature in sketch Engine which is linked to Lexonomy via an API is capable of generating a lexicon automatically for the HRLs but unfortunately for the LRLs like Twi, this is not fully possible because POS tag- set and other features have not been developed for Twi [35]. This is indicative in the table below. Nevertheless, we created the Twi-English lexicon using lexonomy even though we could not capture the totality of the lemma in Twi for the constraint of time and inadequate resources and finance. The statistics of the TWIENG lexicon is indicated in the table below.

Table 5: TWIENG lexicon statistics.

Description	English	Twi
Words	4,903	5,748
Tags	60	37
Lempos	3,721	6239
PoS	9	9
Lemma	4,890	5600

Analysis of Features of the Twieng Corpus

The Twieng Corpus Word Sketch.

A word sketch is a single-page summary of collocational behaviour of a specific word, which is obtained statistically from the corpus data and structured according to grammatical patterns in which they occur [29]. Word sketch of the word '*Jehovah* is shown below.

Figure 4.4: Word sketch of the word Jehovah.

Concordance

The parallel concordance only works with parallel corpora which are aligned. The parallel concordance searches for words, phrases, tags, documents, text types or corpus structures in one language and displays the results together with aligned translated segments in another language. The translated segments usually contain the translation of the search word or phrase but the translation may not be included if the translator decided to use a different way of expressing the idea. The concordance can be sorted, filtered, counted and processed further to obtain the desired result [25,26].

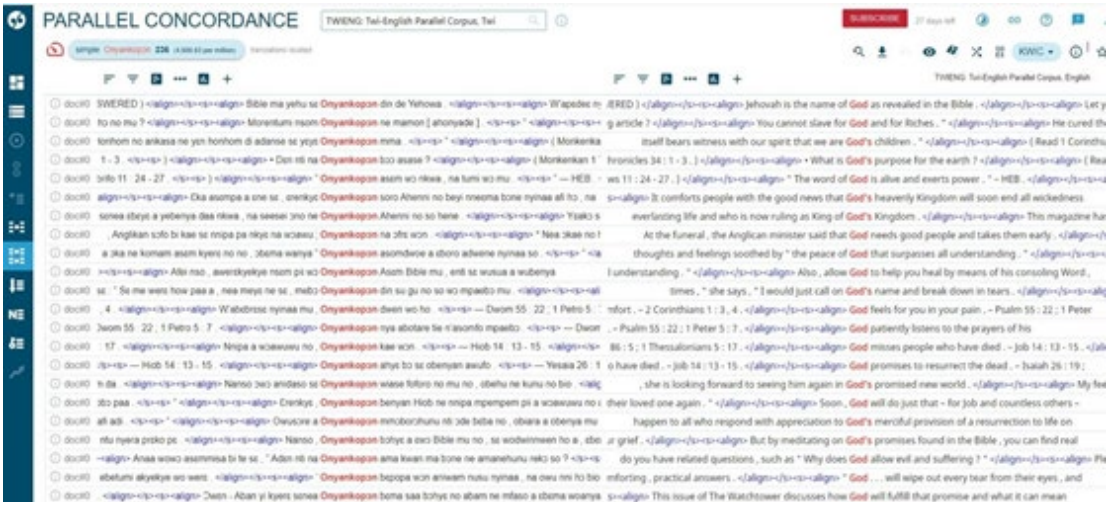

Figure 4.5: Parallel concordance of the words *Onyankopɔn* and *God*

N-Grams

N-grams are also called multi-word expressions or MWEs. The N-gram tool produces frequency lists of sequences of tokens. The user has a choice of filtering options including regular expressions to specify in detail which n-grams should have their frequency generated. N-grams can be generated on any attribute with word and lemma being the most frequently used ones [20,21]. Table 4.6 below reports 3,546 total frequencies of 3-4 grams words.

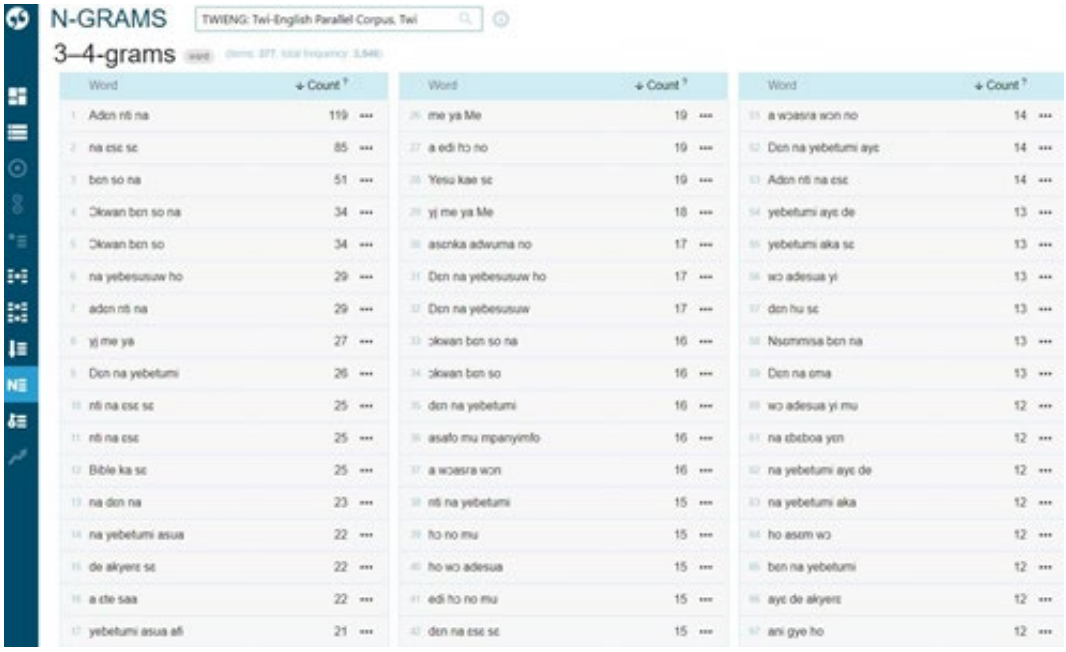

Figure 4.6: 3-4 grams of the TWIENG parallel corpus, Twi.

Wordlist

The wordlist tool is used to generates frequency lists of various kinds: nouns, verbs, adjectives and other parts of speech words beginning, ending, containing certain characters' word forms, tags, lemmas and other attributes or a combination of the three options above. Three different frequency measures can be displayed in the wordlist: frequency, frequency per million and ARF [20,21]. The TWIENG corpus contains 4,932 unique items and total frequencies of 63,873.

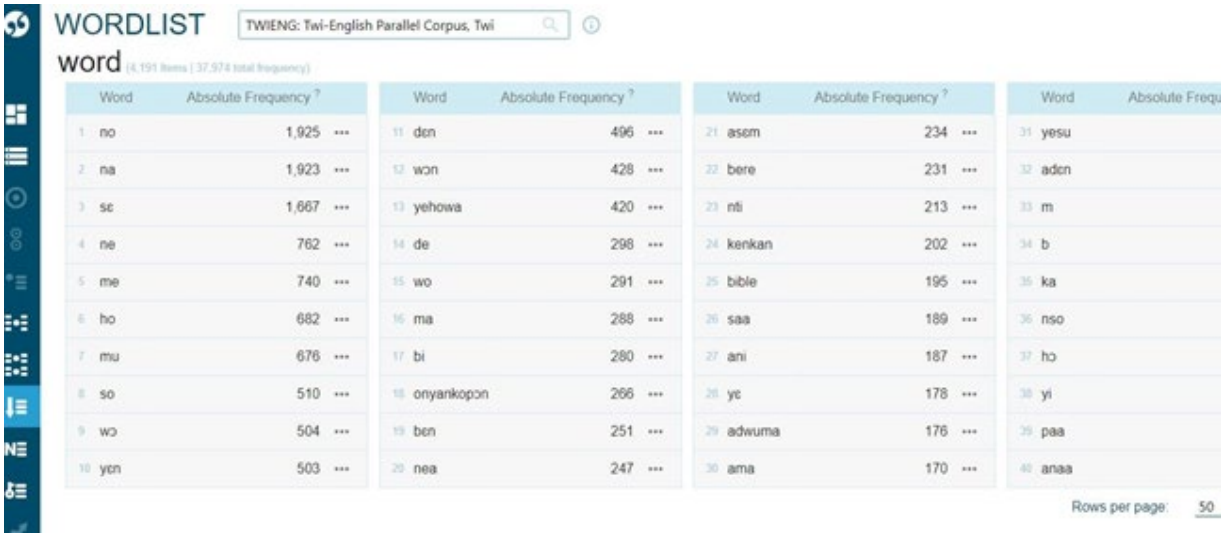

Figure 4.7: TWIENG Corpus wordlist.

Keywords

Keywords and terms assistance us apprehend what the topic of the corpus is or how it differs from the reference corpus. By default, general language corpora are used as reference corpora to represent non-specialized language. Keywords are individual words (tokens) which appear more frequently in the focus corpus than in the reference corpus. Terms on the other hand are multi-word expressions which appear more frequently in the focus corpus than in the reference corpus and, additionally, match the typical format of terminology in the language [20,21].

Figure 4.8: sample keywords of the TWIENG corpus.

Conclusion

In this paper, we presented TWIENG, a novel multi-domain Twi-English parallel corpus of 5,419 sentence pairs and 124,060 tokens. Our corpus is novel for Twi, a low-resource Ghanaian language which is also spoken by a cross-section of the people of Ivory Coast. Our corpus is bigger, better, tokenized, lemmatized and precisely aligned and hence can stand the test of time when used in any MT task that may involve English and Twi. Our corpus has more quality, unbiased and cut across all spheres of life. The TWIENG corpus is open sourced and freely available on the sketch engine website. Finally, professional Twi linguists and translators volunteered to evaluate the corpus manually. Not with standing the efforts we put into this work, Twi is an LRL with many untapped research opportunities. Therefore, a lot of research is needed to bring to light the aspects this paper could not cover.

Acknowledgements

This research was not supported by any grant. We want to thank Mr. Samuel Badu, Miss Ama Achiaa Adams, Miss Serwaa Rita, Mr. Michael Damoah, Nana Kusi Brensian and Mr. Emmanuel Afosah for playing diverse roles in the translation, alignment and evaluation of the TWIENG corpus and all other people who helped this work to get to this level.

References

1. Adomonline. Ghana News, News in Ghana, latest in ghana, Business in Ghana, Entertainment in Ghana, Top Stories in Ghana, Headlines in Ghana, Politics in Ghana, Elections in Ghana, Sports in Ghana, Tourism in Ghana, Health Lifestyle, Radio in Ghana, Celebrations and Advertising HomePage - Adomonline.com.
2. Željko Agic and Ivan Vulic. 2020. JW300: A wide-coverage parallel corpus for low- resource languages. ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (2020), 3204–3210.
3. Alotaibi, H. M. (2017). Arabic-English parallel corpus: A new resource for translation training and language teaching. Arab World English Journal (AWEJ) Volume, 8.
4. Ataman, D. (2018). Bianet: A parallel news corpus in turkish, kurdish and english. arXiv preprint arXiv:1805.05095.
5. Aulamo, M., & Tiedemann, J. (2019). The OPUS resource repository: An open package for creating parallel corpora and machine translation services. In 22nd Nordic Conference on Computational Linguistics (NoDaLiDa). Linköping University Electronic Press.
6. Aulamo, M., Virpioja, S., & Tiedemann, J. (2020, June). OpusFilter: A configurable parallel corpus filtering toolbox. In 58TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL 2020): SYSTEM DEMONSTRATIONS System Demonstrations. The Association for Computational Linguistics.
7. Baisa, V., & Suchomel, V. (2012). Large corpora for Turkic languages and unsupervised morphological analysis. In Proceedings of the Eighth conference on International Language Resources and Evaluation (LREC'12), Istanbul, Turkey. European Language Resources Association (ELRA).
8. Barkarson, S., & Steingrímsson, S. (2019). Compiling and filtering ParIce: an English-icelandic parallel corpus. In Proceedings of the 22nd Nordic Conference on Computational Linguistics (pp. 140-145).
9. Dorothee Beermann. (2013). The TypeCraft (TC). Akan Corpus, (2013), 2016-2018.
10. Beermann, D., Hellan, L., Haugland, T., & Goldhahn, D. (2018). Convergent development of digital resources for West African Languages. Sustaining Knowledge Diversity in the Digital Age, 48.
11. BGL. BGL - The Bureau of Ghana Languages. Retrieved September 1, 2021 from https://www.bgl.gov.gh/language-info/2759436.
12. Bakaric, M. B., & Pacelat, I. L. (2019, September). Parallel corpus of Croatian-Italian administrative texts. In Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019) (pp. 11-18).
13. Christianson, C., Duncan, J., & Onyshkevych, B. (2018). Overview of the DARPA LORELEI Program. Machine Translation, 32(1), 3-9.

14. Citinews. Citinewsroom: Ghana News, Business, Sports, Showbiz, Facts, Opinions.
15. Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. 2019. Universal transformers. 7th International Conference on Learning Representations, ICLR 2019 (2019), 1–23.
16. Doğru, G., Martín Mor, A., & Aguilar-Amat, A. (2018). Parallel corpora preparation for machine translation of low-resource languages: Turkish to English cardiology corpora. In Proceedings of the LREC 2018 Workshop'MultilingualBIO: Multilingual Biomedical Text Processing' (pp. 12-15).
17. Endrédy, I., & Novák, A. (2013). More effective boilerplate removal-the goldminer algorithm. Polibits, (48), 79-83.
18. Fagbolu, O., Ojoawo, A., Ajibade, K., & Alese, B. (2015). Digital yoruba corpus. International Journal of Innovative Science, Engineering and Technology, 2348-7968.
19. Gemmell, C., Rossetto, F., & Dalton, J. (2020, July). Relevance transformer: Generating concise code snippets with relevance feedback. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 2005-2008).
20. Ghaddar, A., & Langlais, P. (2020, May). Sedar: a large scale French-english financial domain parallel corpus. In Proceedings of the 12th Language Resources and Evaluation Conference (pp. 3595-3602).
21. Graphiconline. Ghana news - Top local news in Ghana - Graphic Online. (2021).
22. Jw.org. Kenkan Bible Wɔ Intanɛt So—Wubetumi Atwe Bible Akenkan: PDF.
23. Kashefi, O. (2018). Mizan: A large Persian-English parallel corpus. arXiv preprint arXiv:1801.02107.
24. Ming-wei Chang Kenton, Lee Kristina, and Jacob Devlin. (1953). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Mlm (1953).
25. Kilgarriff, A., & Kosem, I. (2012). Corpus tools for lexicographers (pp. 31-55). na.
26. Kilgarriff, A., Reddy, S., Pomikálek, J., & Avinesh, P. V. S. (2010, May). A corpus factory for many languages. In Proceedings of the seventh international conference on language resources and evaluation (LREC'10).
27. Kovář, V., Baisa, V., & Jakubíček, M. (2016). Sketch engine for bilingual lexicography. International Journal of Lexicography, 29(3), 339-352.
28. Kunilovskaya, M., & Koviazina, M. (2017). Sketch engine: A toolbox for linguistic discovery. Jazykovedny Casopis, 68(3), 503.
29. Living Languages, M David, and Gary F Simons. 2019. Browse the Regions. 2020.
30. Leonhardt, J., Anand, A., & Khosla, M. (2020, April). Boilerplate removal using a neural sequence labeling model. In Companion Proceedings of the Web Conference 2020 (pp. 226-229).
31. Magueresse, A., Carles, V., & Heetderks, E. (2020). Low-resource languages: A review of past work and future challenges. arXiv preprint arXiv:2006.07264.
32. Charles O Marfo and Peter Donkor. (2017). Twi Medical Glossary.
33. Měchura, M. B. (2017, September). Introducing Lexonomy: an open-source dictionary writing and publishing system. In Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference (pp. 19-21).
34. Mehta, S., Ghazvininejad, M., Iyer, S., Zettlemoyer, L., & Hajishirzi, H. (2020). Delight: Very deep and light-weight transformer.
35. Myjoyonline. MyJoyOnline.com - Ghana's most comprehensive website. Credible, fearless and independent journalism. (2021).
36. Orife, I., Kreutzer, J., Sibanda, B., Whitenack, D., Siminyu, K., Martinus, L., ... & Bashir, A. (2020). Masakhane--Machine Translation For Africa. arXiv preprint arXiv:2003.11529.
37. De Pauw, G., Wagacha, P. W., & de Schryver, G. M. (2009, March). The SAWA corpus: a parallel corpus English-Swahili. In Proceedings of the First Workshop on Language Technologies for African Languages (pp. 9-16).
38. Peacefm. Ghana News: Latest News in Ghana | UTV Ghana | Peace FM Online | Ghana Election 2020.
39. Pomikálek, J. (2011). Removing boilerplate and duplicate content from web corpora. Disertacnı práce, Masarykova univerzita, Fakulta informatiky.
40. Rush, A. M. (2018, July). The annotated transformer. In Proceedings of workshop for NLP open source software (NLP-OSS) (pp. 52-60).
41. Rutgers. 2020. Languages Akan (Twi) at Rutgers. (2020), 1–2.
42. Scannell, K. P. (2007). The Crɯbadcn Project: Corpus building for under-resourced languages. Cahiers du Cental, 5, 1.
43. Yuning Shen. (2019). Swahili Media Corpus for research and institutionalized language teaching : Challenges and Opportunities.
44. Siripragada, S., Philip, J., Namboodiri, V. P., & Jawahar, C. V. (2020). A multilingual parallel corpora collection effort for Indian languages. arXiv preprint arXiv:2007.07691.
45. So, D., Le, Q., & Liang, C. (2019, May). The evolved transformer. In International Conference on Machine Learning (pp. 5877-5886). PMLR.
46. Soares, F., Moreira, V. P., & Becker, K. (2019). A large parallel corpus of full-text scientific articles. arXiv preprint arXiv:1905.01852.
47. Suchomel, V., & Pomikálek, J. (2012, April). Efficient web crawling for large text corpora. In Proceedings of the seventh Web as Corpus Workshop (WAC7) (pp. 39-43).
48. Tian, L., Wong, D. F., Chao, L. S., Quaresma, P., Oliveira, F., Lu, Y., ... & Wang, L. (2014, May). Um-corpus: A large english-chinese parallel corpus for statistical machine translation. In Proceedings of the ninth international conference on language resources and evaluation (LREC'14) (pp. 1837-1842).
49. Tracey, J., Strassel, S., Bies, A., Song, Z., Arrigo, M., Griffitt, K., ... & Kuster, N. (2019, August). Corpus building for low resource languages in the DARPA LORELEI program. In Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages (pp. 48-55).
50. United Nations. The Universal Declaration of Human Rights – Stand Ghana. (2021)
51. United Nations. UDHR - Twi (Asante). 2021.
52. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, etal. (2017). Transformer: Attention is

all you need. Advances in Neural Information Processing Systems 2017-Decem, Nips (2017), 5999– 6009.
53. Zhang, Y., Wu, K., Gao, J., & Vines, P. (2006, April). Automatic acquisition of Chinese–English parallel corpus from the web. In European Conference on Information Retrieval (pp. 420-431). Springer, Berlin, Heidelberg.
54. Twi - Wikipedia. Retrieved September 1, 2021.
55. Parliament of Ghana.
56. Full text of "The Tshi Dictionary, 2nd ed."
57. TWIENG Social Media Crowdsourcing data - Google Sheets.

Copyright: *©2022 Gabriel Kwadwo Afram. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.*