 doi:10.20944/preprints202203.0303.v1

Article

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

Gabriel Kwadwo Afram[1*], Benjamin Asubam Weyori[1] and Adekoya Felix Adebayo

[1] University of Energy and Natural Resources, Sunyani, Ghana; felix.adebayor@uenr.edu.gh (A.F.A.); gkafram@susec.edu.gh (B.A.W.)
* Correspondence: benjamin.weyori@uenr.edu.gh

ABSTRACT: A Twi-English parallel corpus is certainly an important resource for Machine Translation of Twi (ISO 639-3), a Low-Resource African Language (LRAL) 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. In this paper, we present TWIENG: a large-scale multi-domain Twi-English parallel corpus. We crawled the sentences from the web using web crawlers, translated, aligned, tokenized and compiled to create the corpus. We crawled English sentences from Ghanaian indigenous electronic news portals, Ghanaian Parliamentary Hansards, Twi Bible and crowdsourcing via google forms. The sentences were translated by professional translators and linguists, they were then aligned, tokenized and compiled. The corpus was curated using the sketch engine, a corpus manager and analysis software developed by Lexical Computing Limited. The corpus was manually evaluated by Twi professional linguists. The Corpus has 5,419 parallel sentences.

Keywords: Twi; Parallel corpus; Tokens; Sketch Engine; Word sketch; Parallel concordance

1. Introduction

Parallel Corpora[2,12,16] 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). 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[2] which is ideologically skewed due to its maximal reliance on the Bible. 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 [8,29] initiated a bold move to create a language dataset and resources for the Low Resource African Languages (LRALs) but the project has been stalled. 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[18,20], a corpus curation and analysis software to build our corpus. 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.

1.1 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2022 doi:10.20944/preprints202203.0303.v1

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

1.2 The Twi Orthography

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 10 diphthongs in the Twi languae[1].

2. RELATED WORK

Machine Translation (MT) has achieved SOTA performance in recent years for a few High Resource Languages (HRLs) [15,24] probably due to the readily availability of parallel corpora and efficient machine translation models such as the Transformer Architecture and its variants[11,14,17,27,32,36]. 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[8] coupled with efficient NMT models[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 [2,3,4,16,31] among others.

The Crubadan project[44] attempted to build an Akan (Twi) parallel corpus by gathering 547,909 Twi words from 176 documents crawled from the web. Facebook uses the Translate Facebook App[2] 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[13] 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. 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[13] which are all African languages. However, the datasets that resulted from this project is not freely available for use by researchers. African based researchers have also taken the initiative to bring African LRLs into the limelight. Deep Learning Indaba[3] is a research group that aims at building machine learning tools for African Languages. MASAKHANE[4][25] group is a research effort for natural language processing targeting African languages. It is open source, spans across the African continent and distributed with online repository of various resources. MASAKHANE has developed 38 unique language

[1] https://www.omniglot.com/writing/akan.htm
[2] https://www.facebook.com/translations
[3] https://deeplearningindaba.com/2020/
[4] https://www.masakhane.io/

 doi:10.20944/preprints202203.0303.v1

pairs and 45 benchmarks[5]. However, we find it intriguing that there is no single language pair for the numerous Ghanaian Languages apart from the JW300[2] English–Twi pairs which can be accessed from the Opus repository[5]. The JW300 corpora are however religiously skewed and hence biased ideologically due to its reliance on the Bible Text[2]. Another Akan/Twi corpus worth noting is the Typecraft Akan corpus[9] which has only 1,906 phrases not significant to train any MT model.

2.1 Objectives

To the best of our knowledge there is no readily available Twi-English 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 5k sentence pairs which is freely available on sketch Engine and our GitHub repository[6] for non-commercial use based on the CC BY-NC-SA 4.0[7] 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. Analysis of various features of the TWIENG corpus which include; Word sketch, Parallel concordance, N-grams, and Word list.

3 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 their contents span 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 website[8], Twi articles were also downloaded together with the various literature. Apart from the Twi Medical Glossary[38] and the Twi-English NT Bible which has already been translated, the rest of the texts were only in English. We therefore used the methodology suggested by [23,29] to create the TWIENG corpus.

3.1 Corpus Preparation

The parallel text required for building a modern digital parallel corpus are usually crawled from publicly available data online[25,28] using web crawlers. 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 standardliterature as 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

[5] https://github.com/masakhane-io/masakhane-mt/blob/master/language_pairs.md

[6] https://github.com/gkafram/TwEng-corpus

[7] https://creativecommons.org/licenses/by-sa/4.0/

[8] https://www.parliament.gh/docs?type=HS

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2022 doi:10.20944/preprints202203.0303.v1

include SpiderLing[7,49], 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. Beautiful Soup[9]- 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 Scraper[10] 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 pairs. 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, Twi Medical Glossary[34], the Universal Declaration of Human rights[52,53] as indicated in table 3 below. 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 website[11]. The Ghanaian parliamentary Hansard gave the corpus a heterogeneous character due to its focus on socio-cultural, educational and legal issues.

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**

3.2 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 crowdsourced 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

[9] https://www.crummy.com/software/BeautifulSoup/

[10] https://www.webscraper.io/

[11] www.Jw.org

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2022 doi:10.20944/preprints202203.0303.v1

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.

1.2 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.

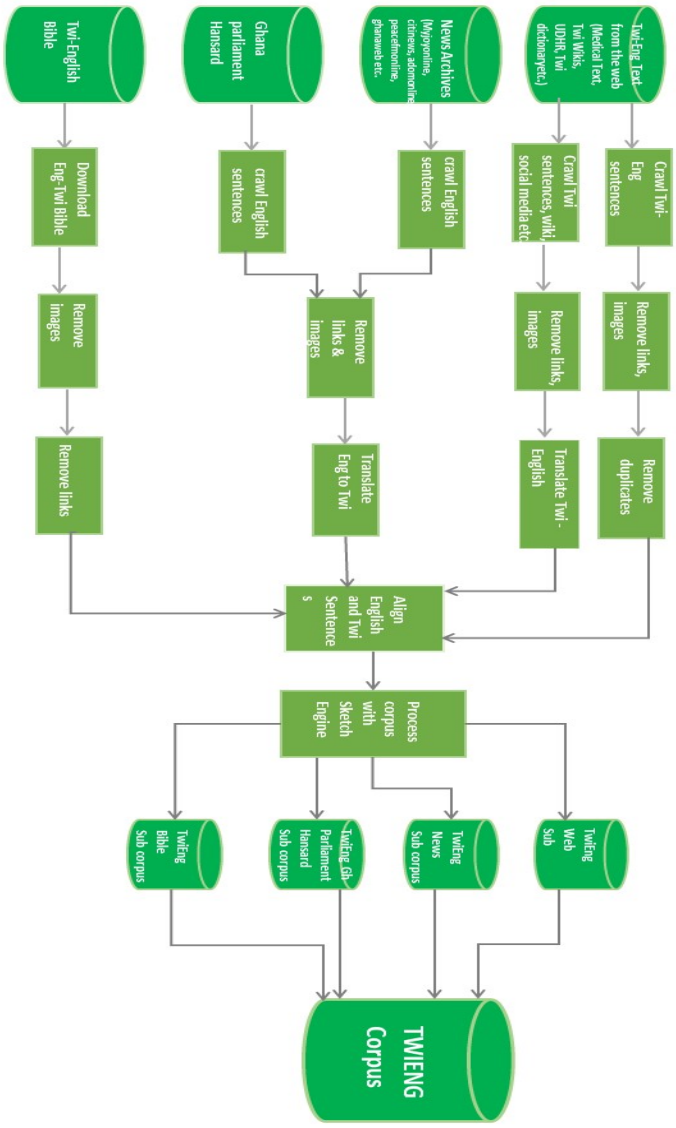

Figure 3. 1 Conceptual framework of the TWIENG Corpus.

3.3 Data Cleaning

The boilerplate removal tool[17,32], justText 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[41].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2022 doi:10.20944/preprints202203.0303.v1

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.

3.4 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.

3.5 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.

4. 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.

4.1 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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2022 doi:10.20944/preprints202203.0303.v1

4.2 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

4.3 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].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2022 doi:10.20944/preprints202203.0303.v1

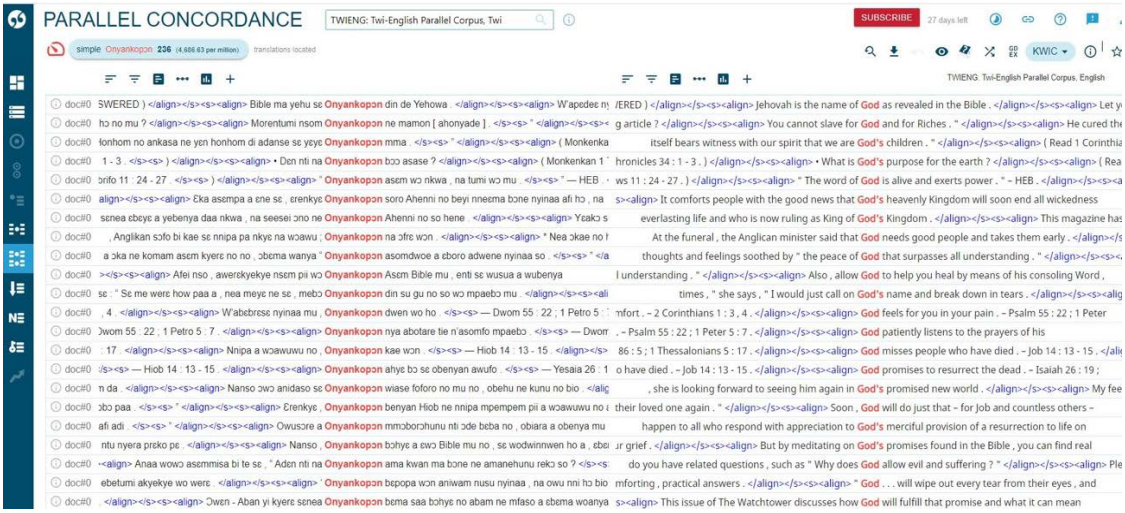

Fig 4.5. Parallel concordance of the words *Onyankopɔn* and *God*

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.

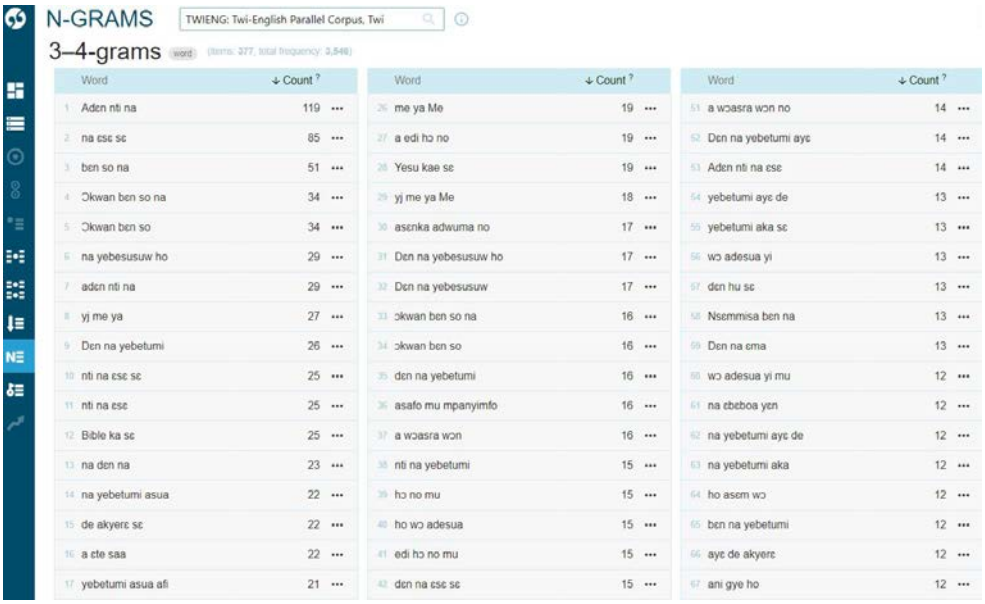

Fig 4.6. 3-4 grams of the TWIENG parallel corpus, Twi.

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.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2022 doi:10.20944/preprints202203.0303.v1

Fig 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].

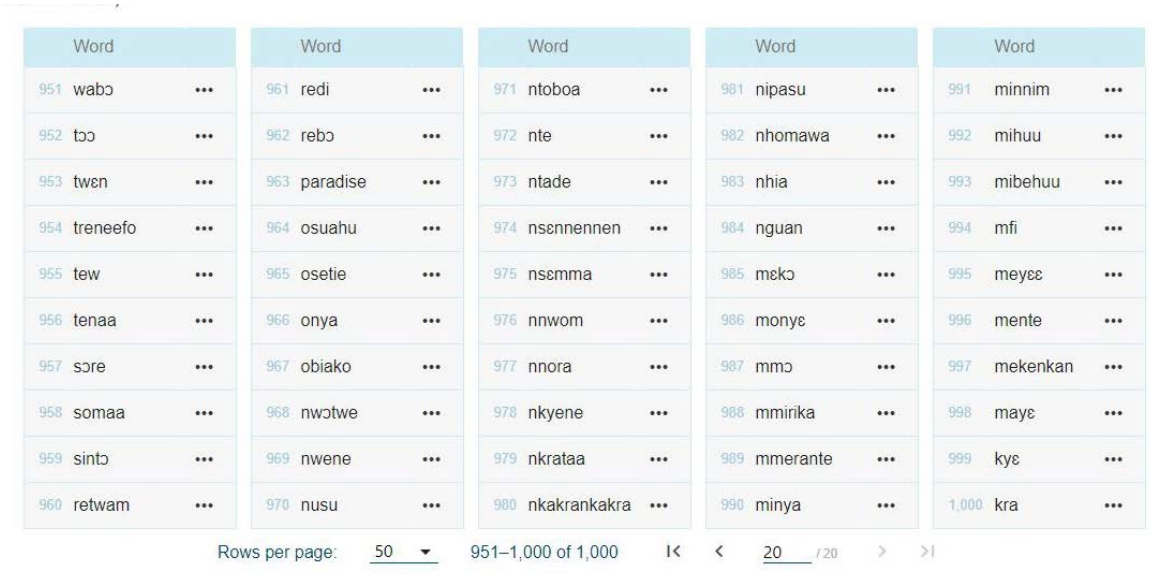

Fig 4.8: sample keywords of the TWIENG corpus.

5 CONCLUSIONS

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.

Notwithstanding the efforts we put into this work, Twi is a LRL with many untapped research opportunities. Therefore, a lot of research is needed to bring to light the aspects this paper could not cover.

6 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, Miss Henrietta Adjei Pokuaa, Mr. Adjei Gyabaah Sylvester, Mr. Yaw Brenya 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. Retrieved August 31, 2021 from https://www.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] Hind Alotaibi. 2017. Arabic-English Parallel Corpus: A New Resource for Translation Training and Language Teaching. *Arab World English Journal* 8, 3 (2017), 319–337. DOI:https://doi.org/10.24093/awej/vol8no3.21

[4] Duygu Ataman. 2018. Bianet: A Parallel News Corpus in Turkish, Kurdish and English. (2018), 1–4.

[5] Mikko Aulamo and Jörg Tiedemann. 2019. The {OPUS} Resource Repository: An Open Package for Creating Parallel Corpora and Machine Translation Services. *Proceedings of the 22nd Nordic Conference on Computational Linguistics* (2019), 389–394.

[6] Mikko Aulamo, Sami Virpioja, and Jörg Tiedemann. 2020. OpusFilter: A Configurable Parallel Corpus Filtering Toolbox. (2020), 150–156. DOI:https://doi.org/10.18653/v1/2020.acl-demos.20

[7] Vít Baisa and Vít Suchomel. 2012. Large Corpora for Turkic Languages and Unsupervised Morphological Analysis. *Proceedings of LREC* (2012).

[8] Starka\dhur Barkarson and Stein\thór Steingr\'\imsson. 2019. Compiling and Filtering {P}ar{I}ce: An {E}nglish-{I}celandic Parallel Corpus. *Proceedings of the 22nd Nordic Conference on Computational Linguistics* 2013 (2019), 140–145.

[9] Dorothee Beermann. 2013. The TypeCraft (TC) Akan Corpus. (2013), 2016–2018.

[10] Dorothee Beermann, Lars Hellan, and Tormod Haugland. 2018. Convergent development of digital resources for West African Languages . Convergent development of digital resources for West African Languages. May (2018).

[11] BGL. BGL - The Bureau of Ghana Languages. Retrieved September 1, 2021 from https://www.bgl.gov.gh/language-info/2759436

[12] Marija Brkic Bakaric and Ivana Lalli Pacelat. 2019. Parallel Corpus of Croatian-Italian Administrative Texts. 0, (2019), 11–18. DOI:https://doi.org/10.26615/issn.2683-0078.2019_002

[13] Caitlin Christianson, Jason Duncan, and Boyan Onyshkevych. 2018. Overview of the DARPA LORELEI Program. *Machine Translation* 32, 1–2 (2018), 3–9. DOI:https://doi.org/10.1007/s10590-017-9212-4

[14] Citinews. Citinewsroom: Ghana News, Business, Sports, Showbiz, Facts, Opinions. Retrieved August 31, 2021 from https://citinewsroom.com/

[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] Gökhan Doğru, Adrià Martín, and Anna Aguilar-amat. 2018. Parallel Corpora Preparation for Machine Translation of Low-Resource Languages : Turkish to English Cardiology Corpora. *Proceedings of the LREC 2018 Workshop'MultilingualBIO: Multilingual Biomedical Text Processing'* (2018), 12–15.

[17] István Endrédy and Attila Novák. 2013. More Effective Boilerplate Removal - the GoldMiner Algorithm. *Polibits* 48, 48 (2013), 79–83. DOI:https://doi.org/10.17562/pb-48-10

[18] Olutola Fagbolu, Akinwale Ojoawo, Kayode Ajibade, and Boniface Alese. 2015. Digital Yorùbá Corpus. 2, 8 (2015), 918–926.

[19] Carlos Gemmell, Federico Rossetto, and Jeffrey Dalton. 2020. Relevance Transformer : Generating Concise Code Snippets with Relevance Feedback. (2020).

[20] Abbas Ghaddar and Philippe Langlais. 2020. SEDAR : a Large Scale French-English Financial Domain Parallel Corpus. May (2020), 3595–3602.

[21] Graphiconline. Ghana news - Top local news in Ghana - Graphic Online. Retrieved August 31, 2021 from https://www.graphic.com.gh/

[22] Jw.org. Kenkan Bible Wɔ Intanɛt So—Wubetumi Atwe Bible Akenkan: PDF. Retrieved August 31, 2021 from https://www.jw.org/tw/nhomakorabea/bible/bi12/nwoma/

[23] Omid Kashefi. 2018. MIZAN: A Large Persian-English Parallel Corpus. (2018).

[24] Ming-wei Chang Kenton, Lee Kristina, and Jacob Devlin. 1953. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Mlm (1953).

[25] Adam Kilgarriff and Iztok Kosem. 2013. Corpus tools for lexicographers. *Electronic Lexicography* (2013), 1–37. DOI:https://doi.org/10.1093/acprof:oso/9780199654864.003.0003

[26] Adam Kilgarriff, Siva Reddy, Jan Pomikálek, and P. V.S. Avinesh. 2010. A corpus factory for many languages. *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010* (2010), 904–910.

[27] Adam Kilgarriff, Siva Reddy, Jan Pomikálek, and P. V.S. Avinesh. 2010. A corpus factory for many languages. *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010* December 2010 (2010), 904–910.

[28] Adam Kilgarriff, Siva Reddy, Jan Pomikálek, and P. V.S. Avinesh. 2010. A corpus factory for many languages. *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010* December 2010 (2010), 904–910.

[29] Vojtěch Kovář, Vít Baisa, and Miloš Jakubíček. 2016. Sketch Engine for Bilingual Lexicography. *International Journal of Lexicography* 29, 3 (2016), 339–352. DOI:https://doi.org/10.1093/ijl/ecw029

[30] Maria Kunilovskaya and Marina Koviazina. 2018. Sketch Engine: A Toolbox for Linguistic Discovery. *Journal of Linguistics/Jazykovedný casopis* 68, 3 (2018), 503–507. DOI:https://doi.org/10.2478/jazcas-2018-0006

[31] Living Languages, M David, and Gary F Simons. 2019. Browse the Regions. 2020.

[32] Jurek Leonhardt, Avishek Anand, and Megha Khosla. 2020. Boilerplate Removal using a Neural Sequence Labeling Model. (2020), 226–229. DOI:https://doi.org/10.1145/3366424.3383547

[33] Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource Languages: A Review of Past Work and Future Challenges. (2020).

[34] Charles O Marfo and Peter Donkor. 2017. Twi Medical Glossary. January (2017).

[35] Michal Měchura. 2017. Introducing Lexonomy: an open-source dictionary writing and publishing system. *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference, Leiden* (2017), 662–679.

[36] Sachin Mehta, Marjan Ghazvininejad, Srinivasan Iyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2020. DeLighT: Very Deep and Light-weight Transformer. (2020).

[37] Myjoyonline. MyJoyOnline.com - Ghana's most comprehensive website. Credible, fearless and independent journalism. Retrieved August 31, 2021 from https://www.myjoyonline.com/

[38] Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, Musie Meressa, Espoir Murhabazi, Orevaoghene Ahia, Elan van Biljon, Arshath Ramkilowan, Adewale Akinfaderin, Alp Öktem, Wole Akin, Ghollah Kioko, Kevin Degila, Herman Kamper, Bonaventure Dossou, Chris Emezue, Kelechi Ogueji, and Abdallah Bashir. 2020. Masakhane -- Machine Translation For Africa. (2020), 1–4.

[39] Guy De Pauw and Peter Waiganjo Wagacha. 2009. The S AWA Corpus : a Parallel Corpus English - Swahili. March (2009), 9–16.

[40] Peacefm. Ghana News: Latest News in Ghana | UTV Ghana | Peace FM Online | Ghana Election 2020. Retrieved August 31, 2021 from https://www.peacefmonline.com/

[41] Jan Pomikálek. 2011. Removing boilerplate and duplicate content from web corpora. *PhD en informatique, Fakulta informatiky* (2011).

[42] Alexander Rush. 2019. The Annotated Transformer. 52–60. DOI:https://doi.org/10.18653/v1/w18-2509

[43] Rutgers. 2020. Languages Akan (Twi) at Rutgers. (2020), 1–2.

[44] Kevin P Scannell. 2007. The Crúbadán Project: Corpus building for under-resourced languages. *Cahiers du Cental* 5, 1 (2007), 5–15.

[45] Yuning Shen. 2019. Swahili Media Corpus for research and institutionalized language teaching : Challenges and Opportunities. August (2019). DOI:https://doi.org/10.13140/RG.2.2.10542.46400

[46] Shashank Siripragada, Jerin Philip, Vinay P. Namboodiri, and C. V. Jawahar. 2020. A Multilingual Parallel Corpora Collection Effort for Indian Languages. *Lrec-2020* May (2020), 3743–3751.

[47] David R. So, Chen Liang, and Quoc V. Le. 2019. The evolved transformer. *36th International Conference on Machine Learning, ICML 2019* 2019-June, (2019), 10315–10328.

[48] Felipe Soares, Viviane Pereira Moreira, and Karin Becker. 2019. A large parallel corpus of full-text scientific articles. *LREC 2018 - 11th International Conference on Language Resources and Evaluation* April (2019), 3459–3463.

[49] Vít Suchomel and Jan Pomikálek. 2012. Efficient Web Crawling for large Text Corpora. *Proceedings of the Seventh Web as Corpus Workshop* (2012), 1–5.

[50] Liang Tian, Derek F. Wong, Lidia S. Chao, Paulo Quaresma, Francisco Oliveira, Yi Lu, Shuo Li, Yiming Wang, and Longyue Wang. 2014. UM-Corpus: A large English-Chinese parallel corpus for statistical machine translation. *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014* (2014), 1837–1842.

[51] Kuster Jennifer Tracy, Stephanie Strassel, Ann Bies, Zhiyi Song, Michael Arrigo, Kira Griffitt, Dana Delgado, Dave Graff, Seth Kulick, Justin Mott, and Neil. 2019. Corpus Building for Low Resource Languages in the {DARPA} {LORELEI} Program. *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages* (2019), 48–55.

[52] United Nations. The Universal Declaration of Human Rights – Stand Ghana. Retrieved September 3, 2021 from http://www.standghana.org/your-rights/the-universal-declaration-of-human-rights/

[53] United Nations. UDHR - Twi (Asante). Retrieved September 2, 2021 from https://unicode.org/udhr/d/udhr_aka_asante.html

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Transformer: Attention is all you need. *Advances in Neural Information Processing Systems* 2017-Decem, Nips (2017), 5999–6009.

[55] Ying Zhang, Ke Wu, Jianfeng Gao, and Phil Vines. 2006. Automatic acquisition of chinese-english parallel corpus from the web. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 3936 LNCS, (2006), 420–431.

[56] Twi - Wikipedia. Retrieved September 1, 2021 from https://en.wikipedia.org/wiki/Twi

[57] Parliament of Ghana. Retrieved August 31, 2021 from https://www.parliament.gh/

[58] Full text of “The Tshi Dictionary, 2nd ed.” Retrieved August 31, 2021 from https://archive.org/stream/rosettaproject_aka_phon-3/rosettaproject_aka_phon-3_djvu.txt

 doi:10.20944/preprints202203.0303.v1

[59] TWIENG Social Media Crowdsourcing data - Google Sheets. Retrieved August 31, 2021 from https://docs.google.com/spreadsheets/d/1CuC1BhrNQy6RI9iV0OZjs5b6X8bpwdgZlW8MHMN_Nl8/edit#gid=2018896676