Towards the Development of Hausa Language Corpus

Dr. Muhammad A. S.[1], Muktar M. Aliyu[2], Dr. Sani I. Zimit[2]

Abstract – Hausa language is the widely spoken language in the west and central Africa. Despite having about 150 million speakers, native and non-native, Hausa language received little attention in work on Natural Language Processing (NLP). Lack of NLP resources for a language can deny its speakers the potential benefits of NLP technology for computer use and information access. This study is an attempt to design and develop corpora for Hausa language as the first desirable step towards addressing the existing gaps for Hausa NLP resources. A bag of words is created from 268 samples of Hausa language text and it consists of a million plus Hausa words in corpus. The words come from a wide range of genres similar to the first widely used English corpora.

Index Terms – Corpus criteria, Hausa NLP, Ngrams, Web scraping and Textual information

1. INTRODUCTION

Corpus is the collection of a language text in accordance to their communicative function without necessary giving regard for the language they contain [2]. It differs from other large collections of machine-readable text such as electronic text in the library or archives in the sense that it is a body of text collected in accordance to explicit design criteria for a specific purpose [1]. Corpora are essential element in developing NLP resources which are used to analyze text, allowing machine to understand how humans speak. Based on this machine-human interactions, real world applications such as text summarization, sentiment analysis, topic extraction, name entity recognition and many more are developed.

Spectator Index in 2018 ranked Hausa language as 11th world's most spoken language with 150 million speakers, native and non-natives. Despite that, Hausa language has very little, isolated NLP resources available. So, this work is an effort to design and develop Hausa language corpus. The language has rich written literature since the first novel (written in Hausa Boko scripts) in 1933. The literature cut across many of the categories of publications under the informative and imaginative prose, similar to the main subdivision of the Brown corpus which is the first broadly available large corpus of English language text.

Designing and developing corpora for a language involve a sequence of activities ranging for planning process to corpus analysis. Planning involves design criteria typically involve decisions such as whether spoken, written language or both should be included, the type of text is to be accounted for, text production period and whether samples of full text are to be included [1]. Other activities include copyright permission, text capturing and corpus processing.

The rest of the section is organized as follows; Section 2 is the literature review of related work from different languages, section 3 is the design consideration for Hausa corpus, then followed by detail implementation of the corpus processing in Section 4. Section 5 is the description of the Data source. Result and discussion is in Section 6 and finally, Conclusion in Section 7.

2. LITERATURE REVIEW

The term corpus means a collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting point of linguistic description or as a means of verifying hypotheses about a language [5]. It is methodically designed to encompass the diversity of a language containing millions of words complied from different running texts across registers. There are many issues related to corpus development such as size of corpus, choice of documents, data selection, collection of text documents (books, newspapers, magazines etc.), text screening, editing and sorting of the collected material according one's requirement, data input, manner of (random, regular, selective etc.) page selection, problems of text screening (omission of foreign words, quotations, dialects etc.), editing, corpus file management etc. [6]. It is essential that a corpus developer for a language, should have a good understanding of the language and it's speakers.

Department of African Languages and Cultures, Ahmadu Bello University (ABU) Zaria issued a communique at the end of one-day colloquium in 2016, that the estimated figures of the native Hausa speakers are more than 70 million and 40 to 50 million non-native speakers that used Hausa as second language in Nigeria on a conservative estimate. Also, Hausa is one of largest spoken language in Africa which is the second largest continents on earth with more than 1.2 billion populations. Despite that, Hausa Language has no major attention in NLP, and there are no existing corpora in Hausa which is available for researchers. Many languages have large corpus for NLP, English in particular, has many large corpora available for NLP. Google has the largest corpus, with one trillion words (tokens) in its N-grams corpus [7]. There are many existing corpora across major languages in the world, but barely you find any existing corpus for many of the African languages, Hausa language inclusive. However, Hausa language is rich in written literatures and

1. *Dr. Muhammad A. S. is currently an Assistant Professor at King Khalid University, Abha, Kingdom of Saudi Arabia. E-mail: muhalisu@mail.com*
2. *Dr. Sani I. Zimit and Muktar M. Aliyu are currently Lecturers at Kano University of Science and Technology, Wudil – Kano, Nigeria. E-mail: sanizimit@gmail.com, muktarmaliyu@gmail.com*

spoken recordings and a lot of archive resources which when utilized will produce much more corpora than the existing corpora in some other languages. Hausa language has been broadcasted in mainstream International Radio stations like BBC, VOA, RFI, Radio DW, and many more, since the launch of BBC Hausa in 1957. In this section, we will review some of these texts that will give more light to our work.

English machine translation using Ruled – based Machine Translation (RBML) which uses parallel corpora design was Developed in [3]. The design of RBML, despite its difficulty and time consuming, is not promising in the recent advancement of NLP as compared with using statistical tools and machine learning. Igbo language, which is one of the major spoken language in sub-Sahara west Africa has about one million tokens in size in its corpus (IgbC) [4], whereas barely you can find any existing Hausa corpus.

The greater number of the native Hausas, speak their mother tongues only, and lack of effective Hausa language machine translation limits their participation in the national discussion and preventing them from reaching out to their global counterparts. Not only that an effective Hausa machine translation is a useful tool to native Hausa speakers, its availability will be a useful tool to both the non-native second language speakers and non-Hausa speakers who wish to study the Hausa language. But effective machine translation can only be developed when there exist NLP resources or tools such as Hausa corpora and language parsers. This study will be limited to development of Hausa language corpus and some basic tools, which are fundamental resources for academic and research in NLP as well as the building block for other NLP applications such as machine translation.

3. DESIGN CONSIDERATIONS FOR HAUSA CORPUS

A. Initial Planning

The long-term objective of this study is to provide a freely available Hausa language corpora repository for future research in Hausa NLP. With this in mind, the current study is designed to collect written Hausa language corpus of various genres. They are categorized into two main division namely Informative prose and Imaginative prose. Informative prose comprises of genre such as Press (reportage, editorial and review), Religion (Book and Periodical), Skills & Hobbies, Popular Lore, Belles-Lettres, Biography, government documents, learned (Sciences, Medicine, Law, Education, Humanities, Technology and Engineering). Fictions (Novels and short stories), Romans and Humor comprises of Imaginative prose.

Following the definition of the design criteria, next is identification of sources of the language texts and a formal request for text acquisition from authors/sources where necessary. Majority of Hausa text are in hardcopy form, many of them require copyright permission which is time-consuming. However, there are growing number of online documents sources in the recent days, some of them are in the public domain.

B. Corpus Collection

This initial study relied on collecting corpus from public webpages using web scraping method. Factors such as size, balance and representativeness need to be considered when collecting corpus. Generally, the bigger the corpus the better most especially if it is language corpora. Corpus balance heavily relied on intuition and feedback from users after the corpus is built. However, reducing skewness factor in corpus comes later and required domain expert participation. Just like corpus balance, representativeness factor comes at a later time. A corpus is said to be representative if user's evaluation of findings drawn from their studies are generalized to the language or particular aspect of language as a whole. Since it is not possible to collect an entire language in order to test for representativeness of corpus. However, it requires pragmatic approach throughout the construction process.

C. Corpus Processing

All corpus must be in electronic form, and simple text (.txt) format is the basic requirement of most corpus investigation packages. Complex embedded formatting with constant changing technology associated with common and popular word processing packages such as Microsoft Word or Portable Document Format (pdf) could not be read by the corpus investigation package. Thus, sequence of methods is required to capture the texts for corpus building. Methods such as data extraction, cleaning, analyzing, and integration to storage are commonly repeated to address issues associated with getting text from various sources. For examples, text acquired from the Internet contains a lot of unwanted formatting and unnecessary textual information, corpus data from hardcopy documents are not readily available in electronic form, they are either keying in when the document is short or scanned them using the OCR (Optical Recognition Software) in the case of a large documents. In either case, there may be inconsistency and required a careful proofreading. Lastly, is issue related to compiling a spoken language is the need for a recording equipment and transcription which is extremely time-consuming process. Just at its simplest, a transcription can simply be an orthographic representation of the words uttered by an informant and probably resemble a play script, it is necessary to transcribe more of the features that are common in unscripted speech such as pause and hesitations, false starts and repetition. And even more importantly, a natural and real data need to be recorded.

D. Corpus Analysis

As essential as the existence of a large corpus for a language, this will not satisfy the demand of linguistic data without accompanied with a set of tools for processing the corpus. These tools are needed to provide useful linguistic and statistical information to make sense of the corpus. While many of these tools exist and are in use for other languages, it is necessary to design a custom tools to meet with specific local needs. Basic tools include Word frequency and Concordance, other advance tools are Lemmatization, Part-of-speech tagging (POS) and Parsers.

The corpus analysis in this study is restricted to Word frequency based on n-gram counts. The corpus is tokenized into Unigram, Bigram and Trigram along with their counts.

4. DETAIL IMPLEMENTATION

In this section we present the detail implementation of corpus building. Figure 1 provides the flow chart of the corpus building. The entire process building process from Data Capture to n-gram counts and visualization are fully automated.

A. Corpus Source Inspection

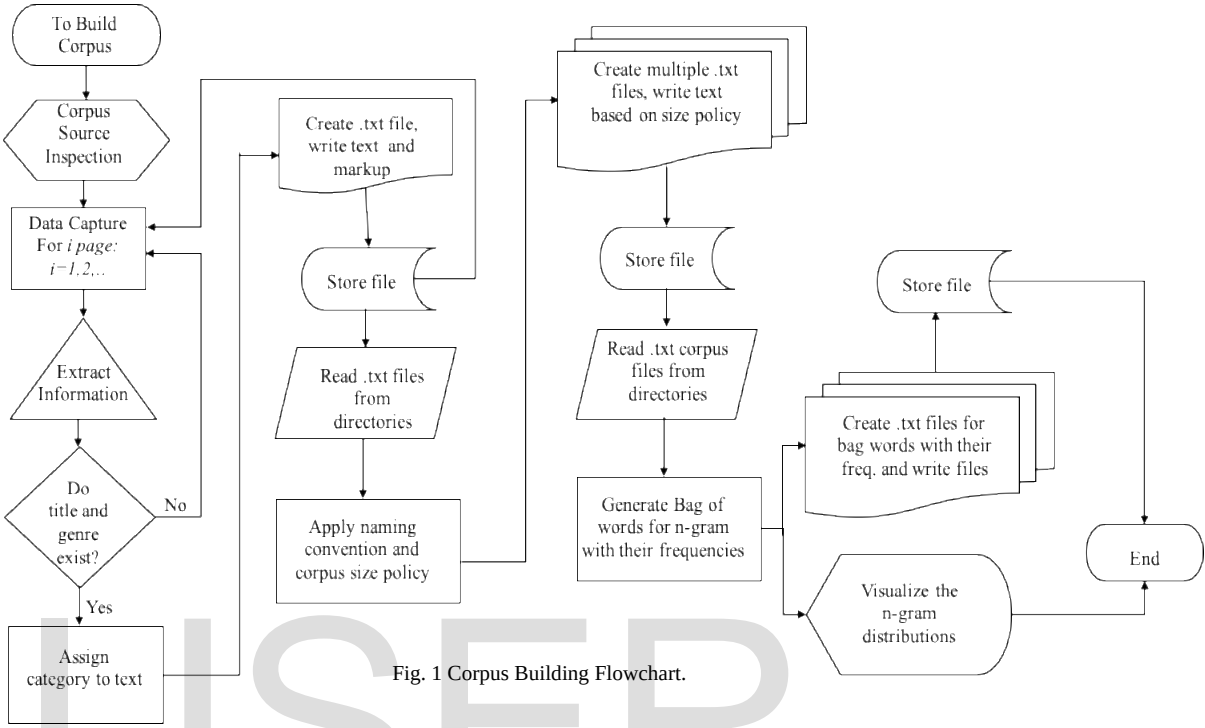

Fig. 1 Corpus Building Flowchart.

Corpus source inspection is the starting point for corpus building process after the design. It involves manual inspection of pages of the data source. This step is important for many reasons; One to skim through contents to ensure its suitability based on the design criteria and to note down the relevant keyword to be used later for categorization of the process, two to collect setup information for data capturing, pending on the data source. In this study, the corpus is built from Internet sources. So, the source inspection is done in two ways, first the visible pages are studied based on the menu and submenu links to identify the category of the content and note down the keyword to associate with the content in the later steps. The second way, was to inspect the source code structural components, this is to identify the relevant html tags to setup in order to capture the relevant information and text body.

B. Data Capture

This process involves using the relevant information like html tag sets from Inspection stage as parameters to capture the require text and markup information such as title, url link, Author name, publication date and category keywords, etc. Pending on the data source, issues unwanted formatting and unnecessary textual information associated with Internet data source is addressed in this stage. Subsequently, relevant information such as article title and category are extracted and used to assign category to a text by creating .txt file with the same name as category with a unique indexing number. The process continue until all Data Capturing is completed.

C. Apply Naming Convention and Corpus Size Policy

To maintain standard, this study adopted naming convention as it was used with a widely used English Language Corpus, Brown Corpus. Table 1 presented the corpus file naming convention, each corpus .txt file naming will have “c” standing for corpus as its first character, then followed by a letter label of the genre or category and a digit indexing number represented as “x” in Table 1. For example, corpus from Press: Reportage which is usually news reporting, will have its first 5 files named ca1.txt, ca2.txt, ..., ca5.txt. Corpus from Learned genre usually scholarly write-ups in science, technology, mathematics, medicine, political science, etc. will have its first 5 files named cj1.txt, cj2.txt, …, cj5.txt.

TABLE I
CORPUS FILE NAMING CONVENTION

SN	Genre	Letter Representation	File Naming
1	Press: Reportage	A	cax
2	Press: Editorial	B	cbx
3	Press: Review	C	ccx
4	Religion	D	cdx
5	Skills and Hobbies	E	cex
6	Popular Lore	F	cex
7	Belles-Lettres	G	cgx
8	Miscellaneous: Government & House Organs	H	chx
9	Learned	J	cjx
10	Fiction: General	K	ckx
11	Fiction: Mystery	L	clx
12	Fiction: Science	M	cmx
13	Fiction: Adventure	N	cnx
14	Fiction: Romance	P	cpx
15	Humor	R	crx

Algorithm I Applying File Naming Convention and sample size policy to Corpus

Input: f_i file; $i = 1, 2, \dots, n$ number of files in the directory
Output: cxy **corpus file**
// *c*, *x*, and *y* in cxy corpus file stands for corpus, letter representing genre and indexing number
// of a corpus file
1**:** $temp_j = [\,],\ Temptxt = \phi$ //$j = 1, 2, \dots, m$ number of genres identified
2: **While** $i < n$
3: **Read** f_i
4: $temp_j = temp_j\ U\ f_i$
5: **Count** the words in $temp_j$ as $len(temp_j)$
6: **Check: if** $len(temp_j) \leq 2000$ words
7: $i = i + 1$
8: **GOTO** Step 2
9 **else** // Reference to Step 6
10: **Check: if** $2000 < len(temp_j) \leq 4000$ words
11: **Open** cxy corpus file according to file naming policy
12: $Temptxt = temp_j$
13: **Write** $Temptxt$ to cxy corpus file
14: **Close** cxy corpus file
15: **Increments** y in cxy by 1 for the identified *j* genre, $Temptxt = \phi, temp_j = [\,], i = i + 1$
16: **else** // Reference to Step 10
17: $sfileLength = 0$
18: $sfileLength = \left\lfloor \frac{len(temp_j)}{2001} \right\rfloor$
19: **If** $sfileLength == 1$
20: $temp = [\,]$
21: $temp = temp_j$
22: **Open** cxy corpus file according to file naming policy
23: $Temptxt = temp[1:2001]$
24: **Write** $Temptxt$ to cxy corpus file
25: **Close** cxy corpus file
26: **Increments** y in cxy by 1 for the identified *j* genre, $Temptxt = \phi, temp_j = temp[2001: end\ ], i = i + 1$
27: **else** // Reference to Step 19
28: $nofile = sfileLength$
29: **While** $nofile \geq 1$
30: $temp = temp_j$
31: $Temptxt = temp\ [1:2001]$
32: **Write** $Temptxt$ to cxy corpus file
33: **Close** cxy corpus file
34: $temp = temp\ [2001: end], nofile = nofile - 1$
35: **Increments** y in cxy by 1 for the identified *j* genre
36: **End while** // Reference to Step 29
37: $Temptxt = \phi, temp_j = temp[2001: end\ ], i = i + 1$
38: **End if** // Reference to Step 19
39: **End if** // Reference to Step 10
40: **End if** // Reference to Step 6
41**: End while** // Reference to Step 2

For sample size policy, on average each corpus file contains more than 2,000 words and no corpus file should have more than 4,000 words. Detail procedure for applying file naming convention and sample size policy is shown in Algorithm I.

The Algorithm begins with reading a preprocessed text file as input and generate a corpus file as output based on the file naming policy and sample size. Each time the Algorithm read a preprocessed text file, appends to a temporary variable, count the number of words in the variable, if they are less than or equal to 2000 words, a new file is read and added to temp variable for the same genre. However, if the number of words are between 2000 to 4000, an output corpus file is created based on the category of the text file in temp variable and naming convention, and the text file is written to the corpus file created and new file is read again into temp variable after it has been emptied. But if the number of words in the temp variable is above 4000, then a number corpus files are created repeatedly pending on the number of words with each file contains 2001 words until when the remaining words in the variable is less than 2001. Then a new file is read and appends to the temp variable and the process continue until all files in the directory are read and corpus file created.

D. Generate Bag of Words for n-gram with their frequencies

To generate a bag of words, corpus files are read from the storage and tokenized, and n-gram consist of unigram, bigram and trigram are generated along with the number of occurrence of each unique instances in the n-gram type. The resulted n-gram counts are save as bag of words. Also, the n-gram are analyze based on their distribution

5. DATA SOURCE

Data source is of great concerns in any corpus design and development processes. Particularly, if a plan is to make a corpus available for future research in the area of natural language processing for the language. Despite the existence of many data source, many of them required obtaining copyright permission to use, and the process required a lot of time. So, in this study, we limited our study to some few private own, public web pages and mostly less active blog sites with contents in Hausa language.

6. RESULTS AND DISCUSSION

This section presents and discusses the results of this research work. Table II presents the distribution of corpus and tokens based on each category or genre and the collection of different categories corpus that formed the bag of words. Despite that the total tokens in the bag of words is 1.1 million, the unique tokens are 33,380 words only. Though the unique words are very low, this may indicate higher number of repeated word's usages in Hausa language, many of which could be stop words.

TABLE II
DISTRIBUTION OF CORPUS SAMPLES AND TOKENS

SN	Corpus Category	No. of Samples in each Category	Total Tokens
1	Press: Reportage	125	813,429
2	Press: Editorial	25	65,751
3	Press: Review	48	120,506
4	Religion	5	11,777
5	Skills and Hobbies	2	4,350
6	Popular Lore	32	78,974
7	Belles-Lettres	7	16,898
8	Learned	21	59,051
9	Fiction: General	3	7,307
	All Corpus	268	1,178,043

Figure II and Figure III provides the distribution of n-gram and their counts for bag of words and each genre corpus respectively for top 20 words. As one expects, the number of occurrences of the top 20 words decreases from unigram to trigram in all categories. Also, the top 20 words unigram is dominated by words such as ‘da’, ‘a’, ‘ya’ ‘ta’, etc. that could be tagged as stop words.

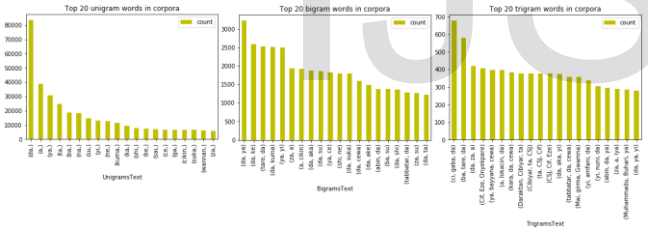

Figure III: Distribution of N-gram Counts based on Bag of Words for Top 20 Words

Figure III: Distribution of N-gram Counts based on Categories for Top 20 Words

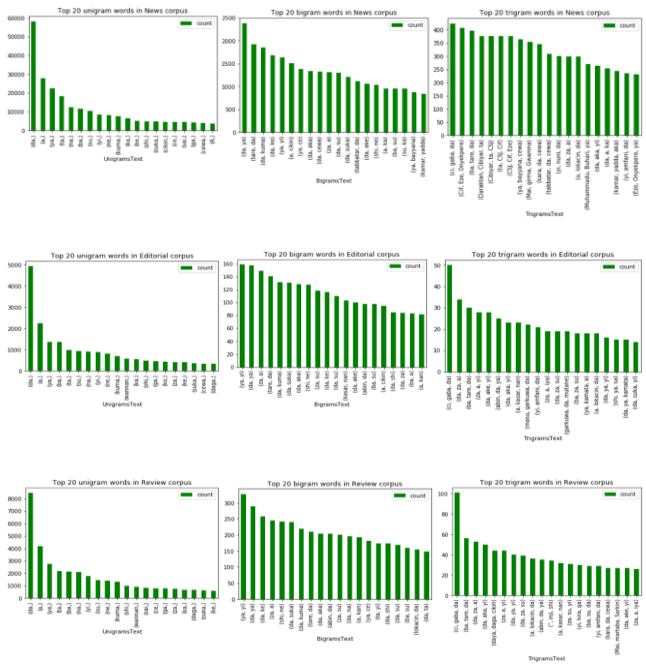

Figure III: Distribution of N-gram Counts based on Categories for Top 20 Words

7. CONCLUSION

This is a pilot study that demonstrate the design and development of Hausa Language Corpus. With this pilot study and basic custom tools developed for Hausa corpus, the future work is to exploit more data source with proper copyright permission and develop more advanced NLP resources such as POS, Lemmatization, and parsers, etc.

References

[1] Atkins, S., Clear J. and Ostler, N. (1991) Corpus Design Criteria.

[2] AHDS (2005) Developing Linguistic Corpora: a Guide to Good Practice.

[3] Tim Schlippe, Edy Guevara Komgang Djomgang, Ngoc Thang Vu, Sebastian Ochs, Tanja Schultz. (2014).

[4] Ikechukwu Onyenwe, Mark Hepple, Uchechukwu Chinedu. (2016). Improving Accuracy of Igbo Corpus Annotation Using Morphological Reconstruction and Transformation-BasedLearning. *JEP-TALN-RECITAL 2016.*

[5] Crystal, David. 1980.*A first dictionary of linguistics and phonetics.* Boulder, CO: Westview.

[6] Shodhganga Chapter 2: Corpus Generation, Retrieved from the Internet 10/6/2019.

[7] Daniel Jurafsky, James H. Martin, 2018 Speech and Language Processing, An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Third Edition draft, September 23, 2018.