THE INTERNATIONAL JOURNAL OF SCIENCE & TECHNOLEDGE

Automation of the Compilation and Processing of a Hausa Corpus

Eno Ubong Ekpo
Student, University of Ibadan, Nigeria
Ubong Sunday Ekpo
Student, ARCIS University of Ibadan, Nigeria
Dr. Tunde Adegbola
Associate Lecturer (Artificial Intelligence), ARCIS University of Ibadan, Nigeria
CEO, African Languages Technologies Initiative, Ibadan, Nigeria

Abstract:
A spell checker is an indispensable tool for text editing as it can be used to assist the possible poor language skills of writers as well as to identify and correct inevitable typing errors. With a population of over 40 million speakers, the Hausa language is the second most widely spoken language in Africa, yet it is without a standard spell checker.
To create a Hausa spell checker, a Hausa corpus was built by data entry and web crawling. The wordlist was cleaned to remove non-Hausa words as well as to correct typographical and other errors. Also, in order to determine the extent to which the modest corpus used for the spell checker covers the Hausa language, the rate of increase in the size of the wordlist in relation to corpus size was determined. A modest 2 million-word Hausa corpus was realized. The corpus was then tokenized to produce a wordlist of about 30,000 Hausa tokens. After cleaning, the wordlist was reduced to 23,306 tokens. Based on the use of Hausa morphology, the word list was compressed to 12,569 stems and 62 affix rules. This made up the spell checker files. Also, a700,000-word corpus drawn from the Hausa corpus was tokenized in separate files with a successive increment of 20,000 words per file.
Results showed that Hausa morphology proved effective for information compression as expected and a rudimentary spell checker was produced. Furthermore, results of the corpus study showed that a corpus of 20,000 words would produce an average of about 3000 tokens and the number of new tokens produced will decrease with every addition of each new file until it asymptotes to a point that an addition of corpus of any size would produce little or no new tokens at all. The rate of new tokens realised with each addition decreased from 2000 tokens to 1000 tokens and to less than that.
This work is recommended for use by individuals, institutions and organisations to guide in the design of standard spell checkers in Hausa language and other languages that feature agglutination.

1. Background to the Study

1.1. What is a Spell Checker?

A spell checker is defined as an application or program that flags words in a document that may not be spelled correctly and suggests possible corrections. It may either be a stand-alone programthat is capable of operating on its own or may be in form of an add-on that is integrated into larger applications such as text editors, word processors, email clients, electronic dictionaries or search engines (Mossberg 2007, James 1980). Naseem (2004), identifies and distinguishes between the two basic functionalities of a spell checker which includes; error detection and error correction.

Error detection function verifies the validity of a word in a language, while the error correction function suggests corrections. Error correction is further classified into automatic correction and interactive correction.

According to Liang (2008), most spell checkers are designed for first language speakers of a language and not for learners such as second language speakers and children.This makes it difficult for this group of users to select from a list of suggested corrections after the spell checker has detected an error, thus spell checking becomes a futile venture for such users. Thus, meaning that spell checkers can only be effectively used by the highly-learned speakers of a language. This view is however confirmed a false notion, after the result of a study by Galletta et al (2006) reveals that there is no difference between the performance of both learned and not—learned users of spell checkers.

Spell checking has become an intricate part of text editing employed by the extremely learned and students alike.This is because apart from increasing the speed and accuracy of editing, spell checkers repose in the mind of the users a certain confidence that they have done a thorough job. However, worries have been expressed on the fact that spell checkers have also become a hinderance. This is because some users have become so dependent on spell checkers that their grammar and spelling skills have declined and they find it difficult to spell words correctly when writing without the help of a spell checker. This is a disadvantage of spell checkers which could

be taken to mean that spell checkers do not help establish correct spellings in the memory of users but rather makes writing so easy for them that the do not need to know how to spell.

Another disadvantage of the spell checker pointed out by Galletta et al (2006) is the fact that a word may be spelt correctly but may not be the word the user intended to use. For example, the user may mistakenly write the word ***"too"*** while he/she means to write ***"to".*** This suggests that users still need to proof read their work before saving the document or sending it. The poem below is used to illustrate the point that a correctly written word may be an error in the sense that it is not the word the writer had the intention of using. This poem was written by Dr. Jerrold H. Zarin 1991 and assisted by Mark Eckman with an original length of 225 words, and containing 123 incorrectly used words.Though all the words are correct, however, they are incorrectly used.The spell checker couldnot tell this since it passed them all as correct words.

Ode to the Spell Checker

Eye have a spelling chequer,
It came with my Pea Sea.
It plane lee marks four my revue
Miss Steaks I can knot sea.
Eye strike the quays and type a whirred
And weight four it two say
Weather eye am write oar wrong
It tells me straight a weigh…

A grammar checker on the other hand, would detect the errors in the above poem. A grammar checker is defined as a program or part of a program that attempts to verify written text for grammatical correctness. They may be a feature of a larger program just like the spell checker, or they may be stand-alone programs (Hopper, 1987). It is also an embodiment of the rules and structure that we use to make sentences. This makes a grammar checker more capable of handling syntax errors, errors of subject, object, predicate positions and others as indicated by the above poem. The spell checker is therefore limited to its simple function of spell checking.

From the foregoing, it is quite obvious that a spell checker as an information system is a necessity, however it is also important that a spell checker be representative of the vocabulary of the language in question. To acheive this, a large electronic corpus is needed. This is where the problem lies. For languages accorded the status "major" which are characterised by a large number of speakers and due standardisation, a level of development of language resources have been attained and this makes it a bit easier to access corpora. However, for minority languages or languages with much less number of speakers, corpus is usually scarce. Percieved solutions to this scarcity are presented in this work along with the results of experimenting with these solutions.

1.1.1. The Hausa Language

Apart from Swahili, Hausa is the most widely spoken African language. It is one of the three languages in Nigeria accorded the status-major. Though there are small clusters of Hausa speaking communities in Sudan, Cameroun, Ghana, Tripoli, Alexandria , Sudan and a sizeable community of speakers in Niger, Nigeria is undoubtedly the home of Hausa language (Abubakar, 2001). Hausa language is also being taught by many foreign universities and news is broadcasted in Hausa language by many radio stations. Foreign universities that include Hausa language in their curriculum as at 2001 includes the Universities of Leipzig, and London, Indiana, Hamburg, Johann Wolfgang Goethe, Frankfurt, Warsaw, Sabbah, Libya and Ghana Universities. In Nigeria, the Bayero University in Kano, Ahmadu Bello University Zaria in Kaduna state and others also have the Hausa language in their curriculum. Among the Radio stations which broadcast in Hausa are, the B.B.C, London, Voice of America, Radio Beijing, Radio Koln, Radio Moscow, Radio Cairo, Radio Ghana, Radio Cameroun and Radio Niger. (Abubakar 2001, Buhari 2011).

Despite all the above, the Hausa language suffers from an insufficiency of corpora. Access to corpora is a major problem for many African languages and limits research and development in such languages. This is quite disturbing because in the spate of the rapid globalization in the world today, it has become pertinent for any entity of future relevance to be up-to-date in the global sphere, through technology and the internet. Standardisation of a language goes beyond the specification of an orthography. It is a continous process that involves creating a standard and maintaining that standard. A spell checker is an important tool for maintaining the standard of a language to ensure uniformity in its usage among its speakers no matter the dialectal differences. It also serves to foster effective communication, consistency in the translation of works, and is a standard for beginners to adhere to. Though the Hausa language is well documented in books ranging from grammar to literature and culture, it is really tasking to find a large body of text in electronic form, even on the internet.

1.2. The Problem

For the existing large number of Hausa speakers, writers, students, translators, and all those involved in hausa data processing and the design of human language technologies, it is of immense importance that they have extensive corpora that is heterogenous and also reliable sources of reference such as hausa dictionaries, spell checkers, morphological analysers and other language tools for reference. There seem to be a scarcity of large deposits of Hausa corpora in electronic form and also no spell checker available for the hausa language for now in the Hunspell program.

In a nut shell, the problems to be solved in this study can be summarised into the following points;-

Hausa language needs a spell checker.

A large corpus is needed to create the spell checker, and this we do not have.
How do we access corpus or a large body of Hausa text?
How do we know our corpus is representative of Hausa vocabulary?
How can the Hausa morphology be used to achieve a drastic reduction in text file size?

1.3. Objectives of the Study

This study is aimed at automating the compilation and processing of a standard Hausa corpus for the creation of a spell checker
The specific objectives include the following;

1) To determine the relationship between the number of words in a corpus and the
2) number of tokens.
3) To determine the ratio between corpus size and the number of tokens.
4) To investigate the possibility of information compression through the use of Hausa morphology.
5) To determine the size of corpus that will be required to produce a token size that is representative of Hausa vocabulary.

1.4. Research Questions

The following research questions have been formulated for the purpose of this research;

(i) What is the relationship between the number of words in a corpus and the number of tokens?
(ii) What is the ratio between corpus size and the number of tokens?
(iii) How much compression can be achieved through the use of Hausa morphology?
(iv) What size of corpus do we require to cover Hausa spell checker significantly?

1.5. Justification of the Study

Hausa is a language with millions of speakers, students, writers, translators and researchers. Each of the above groups participate in Hausa text editing as much as their task requires, therefore the design of a spell checker is highly important for the Hausa language because it has a standard orthography, and just like the English language, this standardisation will be better enforced and maintained with innovations like the spell checker.

The results of the experiments and solutions of this study will not only add to knowledge, it will also arouse in readers of this work, a high level of intuitive capabilities. Readers and future researchers in this area will be inspired to think "outside the box", ask questions and seek fervently for answers to those questions. The researcher has taken up the challenge of going through with this research despite the sparse resources and limited research in this area to serve as an encouragement to future researchers and academics who have in them a penchant for problem solving, and solution seeking no matter the cost orincoveniences.

2. Methodology

2.1. Overview of Research Design and Methodology

The research design for this study is exploratory and the methodology is dependent on the stages involved in investigating the problem of corpus scarcity and the process of compressing information with morphology which will involve the creation of a spell checker.

2.1.1. Data Collection

The data for this study includes a Hausa corpus of at least over two million words. This was gotten from secondary data sources. To gather as many words as possible, different secondary sources such as online text written in Hausa language were gathered using a web crawler and organised into a list with a lexical analyser. A Hausa glossary in hard copy was also converted to soft copy for the purpose of this study. The data collection method used are itemised as follows;

- Webcrawling hausa content

Webcrawling is a process whereby a program is written which goes into specified internet websites online and gathers content automatically. The websites to be visited by the webcrawler are pre-specified in the source code. The researcher also visited these websites to confirm that they contain predominantly Hausa text (except for proper nouns such as names in Hausa news websites e.g www.bbc.co.uk/Hausa/news) before the webcrawling is done. Content was crawled from the following websites for use in this study;
http//:www.bbc.co.uk/Hausa/news/2012
http//:www.aminiya.com
http//:www.dw.com
http//:www.youversion.com/bible/Hausa

- Typing hard copies

This method describes a process where a list of words in hardcopy material is typed into a text editor thereby converting it to softcopy. For this study the glossary of the Hausa text titled "An Introduction to Hausa Morphology" by Abdulhamid Abubakar (2001), was typed to add to the content gotten from the web. Other available lists featuring the Hausa general orthography were also be typed.
Combining the above methods of data collection was considered a more productive, less expensive and less time consuming approach than adopting any single method The only disadvantage recorded for this method was the rigour that was involved at the editing

stage of the wordlist. However, the editing was made easier by the use of a Excel functions for sorting in alphabetical order and duplicates removal. Only Hausa text that contain the special hooked characters, ƙ, ɓ, ɗ, and ƴ were be used. Using words with the Hausa special characters was to ensure that any body of text that was retrieved was at least written in Hausa general orthography.

2.2. Corpus Analysis

The next stage in the methodology was an analysis of the corpus. As a result of the fact that the study has two dimensions, the corpus were also analysed in accordance with the requirements of the two experimental solutions that make up the study. Thus, corpus analysis was carried out in two parts. These includes the following;

2.2.1. Investigating the Challenges of Corpus

To investigate the challenges of creating a spell checker, the researcher identified that the following tasks had to be done:

Prediction of the corpus size that would be needed to extract tokens or unique words that were representative enough to justify the creation of a spell checker.

Recording the trend of increases and decrease in the number of tokens gotten from each group of corpus and find out the cause of these variations.

Recording of the number of tokens that would result from the combination of two groups of corpus.

Recording of the average number of tokens that may be derived from a certain corpus size. The above steps are taken to ensure representativeness of the corpus necessary for the creation of a spell checker.

A corpus can be said to be representative if the findings from that corpus are generalizable to language or a particular aspect of language as a whole. According to Evans (2007), it may not be possible to collect an entire language to test the representativeness of a corpus. However, we will use the notion of 'saturation' also known as 'closure', (McEnery et al 2006:15-16, as cited in Evans, 2007). Saturation (at the lexical level) can be tested for by taking a corpus and dividing it into equal sections in terms of number of words. If another section of the same size is now added, the number of new items in the new section should be approximately the same as in the other sections. This is done in the second stage analysis of corpus.

2.2.2. Corpus Division and Recording of Tokens

The first stage in this analysis was the division stage. This division was done in two parts. In the first part of the division, the corpus was divided first into separate files of 20,000 words each and saved in a folder. In the second part of the division, the entire corpus was again divided into separate files but this time, there was an escalation of every next file. This means that if the first file contained 20,000 words, the next file would contain 40,000 words, the next 60,000 and so on until the corpus is exhausted (i.e the content of the first file plus an additional 20,000 words made up the second file, and the content of the first and second file with an additional 20,000 words made up the third file, and so on). Next, each file was run through a lexical analyser to extract the frequencies of word occurence in each file as well as the total number of unique words (tokens) that result from each file. The number of tokens in each file were then recorded in an Excel spreadsheet in a colum next to the number of unique words. This was done for all the files in both groups.

The data recorded was used to plot graphs for the two parts of division and the results shown in these graphs formed the basis for hypothesis testing and drawing of conclusions for this phase of the study.

2.3. Methodology for Compression Studies

A lexical analyzer was applied to a Hausa corpus of over 2 million words. This converted the corpus to a list in Excel application showing only unique words and no redundances. Also, columns were automatically created by the lexical analyzer which showed the frequency of occurence of each word that made up the corpus. The wordlist that was typed manually was also added to the analysed corpus and a duplicate check was run using the duplicate function in Excel. The next phase was the editing phase, where the researcher checked the wordlist for loan or non-Hausa words and incorrectly spelled words. These errors included non-Hausa words, duplicates and incompletely spelt words probably due to text editing errors or program errors. The words with the highest frequency were mostly single letter words like 'a', most of which were wrong or non Hausa words. Afterwards qualified Hausa linguist Mr Usman also reviewed the list of words to make necessary corrections.

The processes for the creation of the dictionary and affix file is the same used in the creation of a hunspell spell-checker. This was used to experiment on information compression with the use of morphology;

2.3.1. Creating the Dictionary File

The dictionary file was created from the wordlist that has resulted from the above lexical analysis and editing processes. The dictionary file was created by copying the wordlist to a notepad file and saving it as "Hausa.dic". The total number of words were written at the head of the list and the text file was saved in unicode text format, that is UTF-8. This was to ensure that all the special characters for the language were recognised by the system and not substituted with question marks or little squares.

2.3.2. Creating the Affix File

The affix file is simply a rule file that contains the rules for affixing. This means that each rule in the affix file was meant to define the condition upon which an affix can be applied to a word or not applied to a word.The affix file was created by studying the hunspell manual instructions created by Nemeth Laszlo.

2.3.3. Munching

Before the munching process, some simple program lines were used on the Linux Ubuntu Terminal to get and install the hunspell software from the internet onto the Linux OS. The dictionary file was then sorted automatically by the hunspell program to ensure that there are no duplicates before munching takes place. Munching describes a process whereby the rules in the affix file are applied on the dictionary file thereby compressing it and the result is a reduction in file size.

2.3.4. Testing and Comparison

This stage involved typing in erroneous Hausa text on the Ubuntu terminal where the file was munched, in order to check and confirm that the affix rules have been used to affect the dictionary file. This was followed by a recording of the dictionary file size after compression. Haven completed the above processes, the file size of the dictionary file, before compression (pre-compression size), was compared with the size of the dictionary file after compression (post-compression size). The result produced will be discussed in the next chapter.

Alternatively, we can test the workability of the file by continuing with the following steps.

2.3.5. Creating Anextension

The extension .oxt is a folder consisting of files that describe and define the files that make up the spell checker. This means that it is a folder containing meta-data (data about data) of the spell checker.An English extension was downloaded from the internet and modified with information that describe the production, content, author and contributors to the design of the dictionary and affix files. The dictionary and affix files were then copied into the extension folder and then the folder was compressed and saved according to the description file name of the dictionary and affix file. After compression, the folder was renamed from .zip to oxt. At this point, the spell checker was ready for integration with open office org.

2.3.6. Integration with Open Office Word Processor

To integrate the .oxt folder containing the dictionary and affix files (i.e. the spell checker) to open office word processor, the word processor was launched and on the menu tab, 'tools' was selected. This produced a drop-down list from which extension manager was also selected. An interface for adding or removing spell checker extensions resulted from the above selection and the “Add” button was selected upon. As expected the interface for file selection for the entire system was produced from this so that the researcher was able to access the extension (.oxt) containing the dictionary and affix files from the computer desktop where it was saved and added to the word processor as an add-on. The “OK” button ended the integration process and the word processor was closed and re-lunched for testing.

2.3.7. Testing

To test the spell checker a string of Hausa text with deliberate errors written by some native Hausa speakers were used to test the system on the open office word processor.

3. Results

To demonstrate perceived solutions to the challenge of corpus scarcity firstly, it was necessary to determine the number of tokens or unique words that will be derived from a certain corpus size. To achieve this, the relationship between corpus size and number of tokens was studied and the results are presented in section 1 below. Secondly, practical strategies for corpus gathering are identified and discussed in section 3.1.3.

3.1. The Relationship between Corpus Size and Number of Tokens

To study the relationship between corpus size and the number of tokens, a corpus of 700,000 words was analysed in two stages. In the first analysis, The corpus was divided into separate files of 20,000 words each. Each file was then run through a lexical analyzer to produce the number of unique words or tokens for each file. The graphs in 1-3 illustrate the results.

As is illustrate by the graph in figure 1, there is a high level of inconsistency in the number of tokens derived from the first ten files.

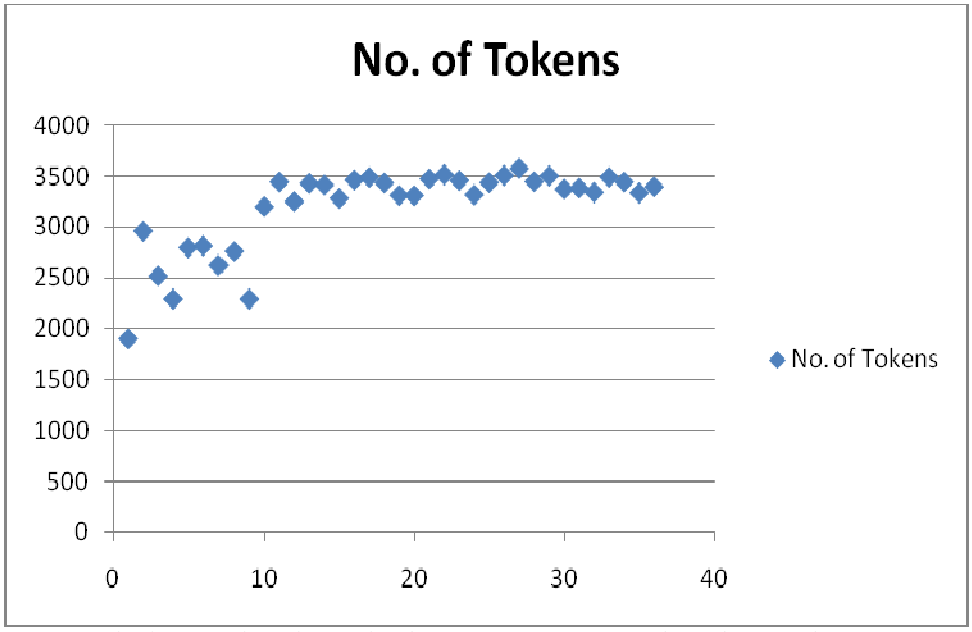

Figure 1:A Graph Showing the relationship between corpus size and number of tokens for first analysis

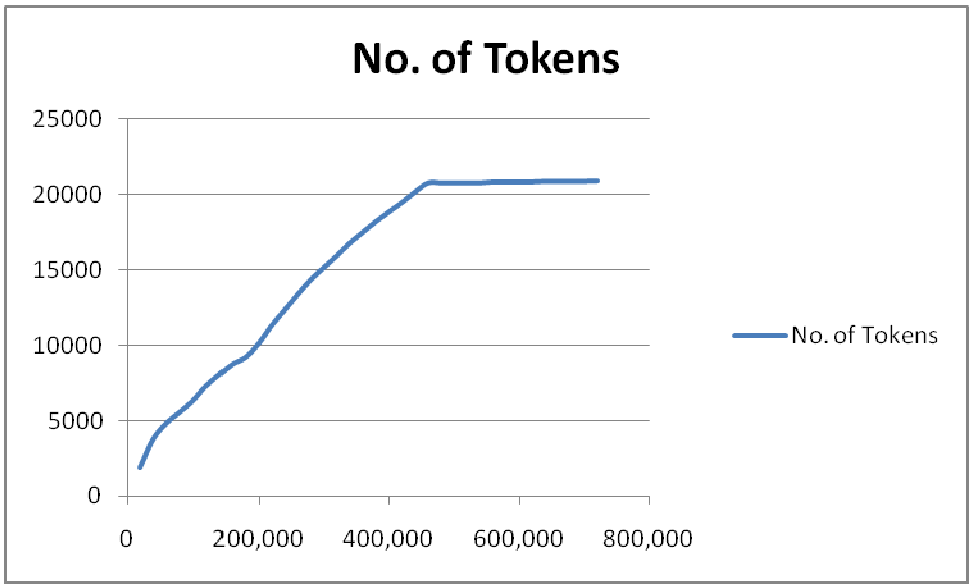

Figure 2: A Graph Showing the relationship between corpus size and number of tokens for second analysis

The graph above is an illustration of the result of the second analysis. The second analysis was carried out using the same corpus of 700,000 words. The corpus was also divided into different files, but this time, there was an escalation of every next file. This means that if the first file contained 20,000 words, the next file would contain 40,000 words, the next 60,000 and so on until the corpus is exhausted (i.e the content of the first file plus an additional 20,000 words made up the second file, and the content of the first and second file with an additional 20,000 words made up the third file, and so on). The first file had 1907 tokens or unique words and with an additional 20,000 words, the second file had 3787 tokens retrieved from it. An additional 20,000 words made up the third file which saw the previous number increased to 4829, and in the next file tokens increased to 5557 words.

It can be observed that in the latter part of the graph where the graph asymtotes we have an almost perfect line. This is as a result of the little or no increase in the number of tokens derived from the last fourteen files. Infact with an addition of every 20,000 words we find that there is no increase at all in token size from the 24^{th} to 27^{th} files. When we searched for the cause of this trend, we discovered the type of information in these four files bordered on a mixture of political essays and social essays. This was probably such a perfect mix that no single additional token was realized. This shows a high consistency in the writings of the author, and a high tendency for redundancy in texts on certain topics.

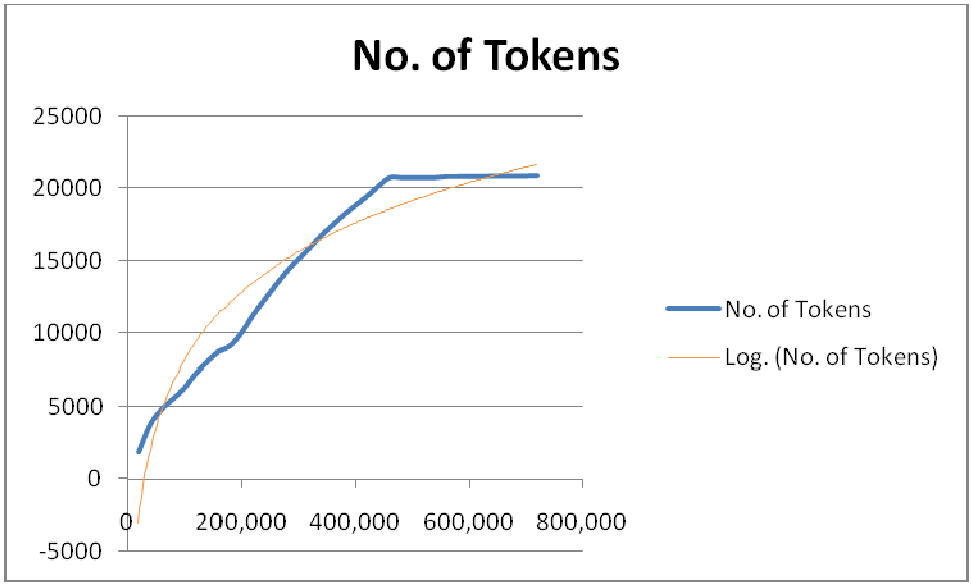

Figure 3: Logarithmic Regression for Second Corpus Analysis

The logarithmic trend for the second analysis result shows that if other corpora of similar information content and type are continually added in the same quantity to the above corpus, the resulting trend line will show an almost perfect level of consistency than what we observe now towards the end of the trend line. Infact it will get to a stage where there will be absolutely no additional tokens realised. To discuss the findings of this study further, it is important that we associate the results with someconcepts of language studies. The concept of heterogeneity of language which is discussed in the next section may lend better understanding to the reason for the characters exhibited in the results presented above. In the meantime, the ratio of corpus size to number of tokens was gotten by finding the mean of all the tokens recorded from each file in the first stage of the corpus analysis. The mean number of tokens recorded for the files in the first corpus analysis was 3176.7.

This therefore means that, for a corpus of size 20000 words, we may derive a token size of approximately 3177. Therefore, the ratio of corpus size to number of tokens = 20000 words: 3177 tokens.

3.2. Discussion and Implications of the Study

3.2.1. Heterogeneity of Language

Heterogeneity of a language is the state of a language having a variety of vocabularies with regards to the various subjects or topics of discourse available to that language. Denoual (2011) uses the term "multidimensional'' to refer to the diversity in vocabulary based on the subjects of discussion, that a single language may possess,which could be responsible for varying values when the corpora of such a language is measured. Also, the vastness of a writer's vocabulary or his/her wealth of knowledge of the language vocabulary determines how expository or conservative expressions will be. In this study,the homogeneity or heterogeneity of corpus was not determined before hand.

In the work done by Denoual (2011), the homogeneity or heterogeneity of corpora used was first determined because the study was basically on corpus analysis and it was important to know the state of the corpus for clearer and uncomplicated results. This is in line with the position of Rose et al (2007) that it is important to determine that a corpus is homogenous before carrying out any similarity measure on such corpus, depending on the focus of the study. For Natural Language Processing systems, Denoual (2012) suggests that data for training such systems should be heterogenous, that is, data should be derived from diverse subjects of discourse in order for the system to function effectively.

Different subjects of discourse were identified in the corpus used for this research work such as social, political, religious, economic and historical writings and essays. Although the foci of this study was not corpus analysis,it was important to discover the reason for any differences in the number of words derived from extracts of the corpus. As can be seen on the illustrations in **figure 1 figure 2** and **figure 3**, there is visible heterogeniety in the corpus used in this study, and even in thecorpora derived fromthe same source/ style of writing.

The study has therefore been able to bring to light the heterogenic nature of the Hausa language which is expressed in the huge differences between the varying sizes of tokens derived from the corpus under topics of social, political, religious and economic essays and expository writings and that of news articles on these subjects.This means that as there exists different subjects of discussion in a language so also there exist different vocabularies and the subjects; sports, cookery, marriage, politics, economy each have their own unique vocabulary. Thus, we can say that 'subject' is a major factor of heterogeneity in Hausa language. From the foregoing we therefore arrive at the conclusion that the number of tokens that can be retrieved or realised from corpora is dependent to a great extent on the following;

→ The author's writing style :- For example the author's style of writing could be *expository* or *reservatory*.

- → The subject of discussion :- The subject of discussion may be based on the economy, education, health etc.
- → The purpose of writing :- For example the purpose of writing could be for information on current happenings e.g news, for information about a particular subject or field e.g essays, for recreation e.g stories, prose, novel, and for the purpose of reporting progress and procedures of projects or tasks.

Finally, The complexity of a language, rules of grammar, and the extent of its vocabulary.

In conclusion, Hausa is a heterogenous language as shown in the results illustrated in **figures 1** and **2** To realise a token size that will be sufficient and also be representative of Hausa language, corpora must be obtained pertaining to different subjects and disciplines, in different forms of writing such as news format, essay format, report format, as much as possible. This can be achieved by either developing corpora manually by typing, or by translation, and also webcrawling. A combination of these three solutions will go a long way to make it a lot easier and less challenging to develop a Hausa spell checker. This may also be applied to other agglutinative languages that show the symptoms identified in Hausa language.

3.2. Proposed Solutions to Corpus Scarcity

The second step towards solving the problem of corpus, is identifying strategies for corpus gathering. we may choose from three possible solutions.

3.2.1. Scanning Written Text

Optical Character Recognition scanners have been used since the 1970s for conducting editable scan on written text. From its price of about £70,000, it may now be purchased for as little as £30. However, in this part of the world, it is not so easily available, and when it is, the cost is rather high. However, if it can be afforded, then the use of editable scan is definitely an option for corpus gathering.

3.2.2. Web Crawling

This may be achieved with the use of a WebCrawler program to crawl content from websites on-line. However, his method also has its challenges. One major challenge is the issue of copyright. Any one obtaining texts online for the purpose of academic research may point out to the publisher of such work what the text is for. The purpose must be non-profit. Again, asking and receiving permission to use the text may take a long process or even involve financial costs. Another issue is the problem of loan words. Web crawling content may bring in not only the Hausa words on a web page but also many English words and even other languages will be crawled together with the language of focus. Many sites that could contain voluminous Hausa text also contain excessive loan words which do not contribute to Hausa morphology.This may make the process of corpus cleaning or editing take longer, and be more tedious than necessary.

3.2.3. Data Entry

This is the process of converting text from hardcopy to soft copy using the typing method. With the availability of the language font, and expert typists to do the work, it might be a good option. However, when we talk of corpus, we refer to hundreds of thousands of words. Typing may prove quite tasking in this regard. Therefore, to solve the problem of availability of corpus, it is recommended that the three options above are combined to solve the problem.

3.3. Ergodicity of Language

The word ergodicity derived from the term ergodic was coined from two greek words 'ergos' meaning 'work' and 'hodos' meaning path (Aarseth, 1997). Therefore the ergodicity of language is the regularity in language which stems from the source of that language and its rules of morphology. Aarseth (1997), defines an ergodic text as one that contains the rules for its own use.Thus, language can be said to contain certain requirements that facilitate the generation of words. This also means that things, both living and non-living such as plants, animals, food crops, seasons, events, are given names based on their availability in the native environment of the speakers of a language and thus other words are formed. This is why for instance the waterleaf plant which grows in areas where the weather features high level of humidity and rainfall is called ***"mmongmmong ikong"*** by the ibibioa and the same plant is also called ***"guree"*** by the Yoruba, whereas the northern parts of Nigeria do not have a known name for the plant because the plant is not a characteristic of that environment.

Also, the Hausa name for rice which is "shinkafa'' does not sound like a derivative of the English name 'rice'. This is because the Hausas have been farmers of the local rice for a long time, especially those in Adamawa and Kebbi states so they had the crop in their environment and thus named it accordingly long before the introduction of the foreign species of rice to Nigeria. Some other languages however, derive the name for this food from the English name "rice" because they had no access to the crop in their environment until it was imported either from the north or from overseas. The Yorubas call it '***iresi'***, the oron people call it "edesi" even the French call it ***riz*** and these names all sound like proper derivatives of the English word, ***rice***. In addition, there is an element of derivation in the name for the animal 'pig' called 'elede' by the yorubas and "alade" by the northern Hausas. Since it is easier to find pigs in northern streets, much more than the western part of Nigeria, we may say that the name for pig in Yoruba was derived from the Hausa name for the same animal. In the same way, the Hausa language has vocabulary for different subjects of discourse.

In view of the foregoing we may say that the regularity exhibited by language is dependent on the source or environment of origin of that language, the type or resources, people, events, seasons, crops, animals, plants found in the environment where the language originated and lastly the morphology of the language. This is why the vocabulary of human language is infinite. As long as the rules of morphology remains constant, such as that of affixing suffixes and prefixes to roots or stems of words, language will constantly,

infinitely birth new words. As the meaning of the word ergodicity implies, we can say that human language has a "work path'' or rules for operation or regularity'' and as long as these rules do not change, and the native environment of the language does not change, the type of people found there do not change, then language will always exhibit regularity.

4. Conclusions, Recommendations and Future Research Directions

4.1. Conclusion

The investigation and study of corpora in this study was done to provide a guide and control the quality of the dictionary file in the process of creating a spell checker.

The wealth of information produced in this study will be relevant to future and ongoing studies on information technology design and theory. In addition to the foregoing, this study will serve the purposes of standardization of texts, consistency of spelling and avoidance of misunderstanding with the attendant advantages they offer to written communication process. Also in the aspect of language, the project demonstrates the do ability of hunspell spell checkers for African languages as a necessary first step in ensuring the presence of African languages in Cyberspace and the survival of African languages in the Information Age.

4.2. Recommendation

The following recommendations are offered for the benefit of further research related to contemporary information age tools such as human language technology tools as well as research on language resources.

(i) Languages interested in creating dictionaries to embed in the Hunspell spell-checker program will find this work very informative and useful.

(ii) As this study is basically a pathfinder, the Hausa speaking community of writers, translators, students and linguists will find it interesting and useful as a basis for developing larger dictionary and affix files for the Hausa language to be embedded in the Hunspell software successfully for public use.

(iii) Recent events have shown that the world has realised the importance of information as the bedrock of human existence. Since language in whatever form is the only means of communication, various government bodies and non-governmental organisations have shown increasing efforts to acquire and/or design tools that will enable them gain access to information that they would not have had otherwise because of language barrier. The issue of language is of high important especially in the aspect of national security. Therefore, this study is a contribution to on-going academic or sponsored researches on human language technologies including assistive languages technologies for the disabled, by organisations such the United States of America's Defence Intelligence Agency in conjunction with the Office of the director of National Intelligence, the African Academy of Languages (ACALAN), Human Language Technology Research Institute and others.

4.3. Future Research Directions

Further research on this study should focus on more than this work has achieved including the following:

1. There is much room for further research on the intricacies of corpora for Hausa language. The corpus analysis done in this study barely scratches the surface of work that abounds in this area.
2. This study was only able to munch the affix file with single character (i.e. A-Z and 0-9) rule-tags directly. Double character-tags munching was achieved through the tedious process of manual replacement of single tags with double tags, in order to accommodate the growing number of affix rules in the affix file. Further study will do well to investigate deeply, why the double character munching was not achieved, whilst it has been stated in the Hunspell manual that double character tags and even mixed tags (alpha-numeric tags) are agreeable to Hunspell.
3. There is still room for research in the aspect of creation of affix rules for Hausa language, compounding of words, and other
4. To produce a spell-checker that is linguistically sound, the dictionary file must be more thoroughly edited with no linguistic errors. Perhaps converting a standard Hausa dictionary from hardcopy to soft copy will produce the required linguistic standard for a spell checker but it would still require editing to remove typographic errors and there will be less opportunity to learn from corpus but the dictionary realised will be more linguistically sound.

5. References

i. Aarseth, E .J. (1997). Cybertexts: Perspectives on Ergodic Literature. Doi:10.4324/9780203935170

ii. Abubakar, A. (2001). An lntroductory Hausa Morphology.Doi:10.3239/9783640076123

iii. Buhari, H. A. (2011). A Comparative Analysis of EWord-Formation Processes In Engish And Hausa. (Master's Thesis) Department of English and Literary Studies, Faculty of Arts, Ahmadu Bello University, Zaria, Nigeria.

iv. Hopper, Paul (1987): Emergent grammar. In: Aske, Jon et al. (ed.) (1987): General session and parasession on grammar and cognition. Proceedings of the thirteenth annual meeting. Doi:10.1017/S0954394502143018.

v. Liang, H.L. (2008). Spell checkers and correctors: a unified treatment. (MSc dissertation) University of Pretoria, Pretoria, <http://up etd. up. ac. za/ thesis/ available/etd-06252009-163007/ >E1 305/ gm.

vi. Mossberg, W. (2007). "Review". Wall Street Journal. Retrieved. Book s.goo gle.com .ng/book s? isbn=113317189.