Statistical Analyses of Myanmar Corpora

Hla Hla Htay, G. Bharadwaja Kumar, Kavi Narayana Murthy
Department of Computer and Information Sciences
University of Hyderabad
hla_hla_htay@yahoo.co.uk, g_vijayabharadwaj@yahoo.com, knmuh@yahoo.com

Abstract

Myanmar(Burmese) is a member of the Burmese-Lolo group of the Sino-Tibetan language spoken by about 21 millions people in Myanmar (Burma).It is also official langauge of Myanmar. Corpus based statistical analyses has not yet taken much focus on Myanmar language. In this paper, we have carried out statistical analysis on Myanmar text corpora of about 1.6M words. In Myanmar language, sentences are clearly delimited by a unique sentence boundary marker but words are not always delimited by spaces. It is therefore non-trivial to segment sentences into words. Syllabification itself is a non-trivial task. Syllabification is done with longest string matching using syllable list. Word Segmentation is performed with longest syllable matching by looking up the dictionary. We have checked manually the outputs of syllabification and word segmentation modules. We have obtained 99% accuracy on syllabification and 80% on word segmentation. We have carried out growth rate analysis and coverage analysis of Myanmar corpus. We have also carried out analyses of word length and sentence length distributions of Myanmar corpus.

1 Introduction

Myanmar(Burmese) is a member of the *Burmese-Lolo* group of the *Sino-Tibetan* language spoken by about 21 millions people in Myanmar (Burma).It is also official langauge of Myanmar. According to history, Myanmar script has originated from *Brahmi* script which flourished in India from about 500 B.C. to over 300 A.D. The script is syllabic in nature, and written from left to right.

Myanmar script is a writing system constructed from consonants, consonant combination symbols (i.e. Medials), vowel symbols related to relevant consonants and diacritic marks indicating tone level (niggahita, visajjaniya). Myanmar language is composed of 33 consonants, 11 consonant combination symbols, 12 basic vowels and extension vowels, vowels symbols, diacritic marks and devowelizing consonants [1, 2]. See in Table 1 and 2. Sentences are clearly delimited by a unique sentence boundary marker "။" which is called ပုဒ်မ[pou ′ ma.] but words are not always delimited by spaces. Although there is a general tendency to insert spaces between phrases, inserting spaces is more of a convenience rather than a rule. Spaces may sometimes be inserted between words and even between a root word and the associated postposition. In fact in the past spaces were rarely used. It is therefore non-trivial to segment sentences into words. Word tokenizing plays a vital role in

most Natural Language Processing applications such as Summarization, Information Retrieval, Machine Translation, Text Categorization, and so on.

Consonants				
က [ka.] **u**	ခ [kha.] **c**	ဂ [ga.] *	ဃ [ga.] **C**	င [nga.] **i**
စ [sa.] **p**	ဆ [hsa.] **c**	ဇ [za.] **Z**	ဈ [za.] **ps**	ည [nja.] **n**
ဋ [ta.] #	ဌ [hta.] **q**	ဍ [da.] **!**	ဎ [da.] ¡	ဏ [na.] **P**
တ [ta.] **w**	ထ [hta.] **x**	ဒ [da.] ’	ဓ [da.] ”	န [na.] **e**
ပ [pa.] **y**	ဖ [hpa.] **z**	ဗ [ba.] **A**	ဘ [ba.] **b**	မ [ma.] **r**
	ယ [ja.] ,	ရ [ja. or la.] **&**	လ [la.] **v**	ဝ [wa.] **0**
	သ [tha.] **o**	ဟ [ha.] [	ဠ [la.] **V**	အ [a.] **t**
Consonants Combination symbols				
ျ **-s**ြ**-** **j-**ွ **-G**ှ **-S**ျွ **-R**ြွ**-** **>-**ှျ **-Q**ြွ**-** **j** **§-**ွ **-T**ျွ ြွ **T̀**				
Devowelizing consonants				
-င်-if ်**-န်-cf** ်**-ဂ်-*f**်**-င်-if** ်**-စ်** **-pf**်**-ဇ်** **-Zf**်**-ဉ်** **-Of**်**-ည်** **-nf**်**-ဋ်** **#f**်				
-ဍ် **-Xf**်**-ဏ်** **-Pf-တ်** **-wf**်**-ထ်** **-xf**်**-ဒ်** **-f́**်**-ö̈f** ်**-န်** **-ef**်**-က်** **-uf**				
-ဖ် **-Af**်**-ဘ်** **-bf**်	**-မ်** **-rf**်**-ယ်** **-,f**်**-ရ်** **-&f**်**-လ်** **-vf**်**-ဝ်** **-0f**်**-သ်** **-of**်**-ဟ်** **-[f**်**-ဠ်** **-Vf**			

Table 1: Consonants

Basic vowels and its extension		
အ [a.]**t**ာ [a]**tm**ံ [i]အ [a.]**t**ီ [i.]**£**ိ [i]ဥ [u.]**O**ဦ [u]**OD** ဧ [ei]**{**ေ [ei:]**tJ**ြ [o:]**Mo**ော် [o]**aMomf**ံ [an]**tH**ိုး [ou]**tdk**		
-ိ **-**	**-**ာ **-m**	**-** ာ း **-m;**
-ု **-d**	**-**ူ **-D**	**-**ူ း **-D;**
-ု **-k**	**-**ူ **-l**	**-** ူ း **-l;**
-ေ **a-**	**-**ေ့ **a-h**	**-**ေ း **a-;**
-ဲ **-J**	**-**ဲ့ **-Jh**	**-**ဲ း **-Jh;**
-ော **a-m**	**-**ော့ **a-mh**	**-**ော် **a-mf**
-ဴ **-H**	**-**ဴ့ **-Hh**	**-**ဴ း **-Hh;**

Table 2: Vowels

Corpus based approaches to language have made significant contributions to linguistic research as also in education and language technology. A corpus is a large and representative collection of language material stored in a computer processable form [3]. Corpora provide realistic, interesting and insightful examples of language use for theory building and for verifying hypotheses[4, 5, 6, 7]. Insights obtained from analysis of corpora have led to fresh and better understanding of how language actually works[8, 9, 10, 11]. Corpora provide the basic language data from which lexical resources such as dictionaries, thesauri, word-nets, etc. can be generated. Language technologies and applications such as Morphological Analyzers,

Stemmers, Syntactic Parsers, Spell Checkers, Information Retrieval systems, Information Extraction systems, Automatic Text Summarization systems, Automatic Text Categorization systems, Machine Translation systems etc. greatly benefit from language corpora. Development of large and representative corpora and annotating them with morphological, syntactic and semantic information is therefore considered to be a priority area. Corpus based statistical approaches have emerged as promising alternatives to traditional linguistic approaches. Hybrid approaches that combine traditional linguistic approaches with corpus based statistical approaches have also become attractive.

Corpus based studies in English date back to 1960s [12, 13, 14, 15, 16, 17, 18]. Corpus linguistics has not yet become a major aspect of education and research in linguistics in Asian languages [19] in general and Myanmar in particular. Even plain text corpora available are inadequate and annotated corpora are hardly available in many languages.

Even, large scale plain text corpora is not available in Myanmar language. There is a greater need to develop large scale Myanmar corpora. In this paper, we will discuss briefly about the development of corpus carried out in University of Hyderabad. Then, we will explain algorithms for Syllabification and word tokenization. Finally, we discuss in detail about the different analyses carried out on Myanmar corpus.

2 Myanmar Text Collection

Development of lexical resources is a very tedious and time consuming task and developing purely with manual effort is very slow. We have downloaded Myanmar texts from various web sites including news sites and on-line magazines. As of now, our Myanmar corpus contains 1.8M sentences. The downloaded corpora need to be cleaned up to remove hypertext markup etc. We have developed the necessary scripts in Perl. Also, different sites use different font formats and character encoding standards are not yet widely followed. We have mapped these various formats into the standard *WinInnwa* font format. We have stored the cleaned up texts in ASCII format. This will enable processing in environments where Unicode is not yet supported. It is easy to switch to Unicode where required.

The corpus includes over 300 full books as well as free and trail text from online book store to include a wide variety of Myanmar writings including a variety of genres, types and styles - modern and ancient, prose and poetry. It also contains from official newspapers in Myanmar. Text are converted to standard *Wininnwa* font using tools developed here. The corpora are ISCII encoded and are seen to be reasonably clean.

A corpus should be constructed in keeping with the principles of corpus linguistics [20, 21, 22]. It must be 'large' and 'representative'. A balanced corpus, however, does not mean nearly equal amounts of material from various genres, types and styles. Application of language in areas other than literature is a relatively recent phenomenon. Newspapers cover a wider variety of topics and styles including sports, science and technology, politics, economics and business, cinema etc. No corpus should be put to use for a given application without a careful analysis of its nature and contents. While there can be no guarantee that our corpus is good

enough for any given use or application, we feel that the corpus is good enough for some kinds of applications we have in mind. Our aim shall be to strive to build larger, more balanced and more representative corpora.

Preliminary studies suggested that Myanmar sentences can be tokenized by eliminating stop words. Hopple [23] also notices that particles ending phrases can be removed to recognize words in a sentence. Stop words are defined as non-information-bearing words. They form closed classes and hence can be listed. Stop words include prepositions/post-positions, conjunctions, particles, inflections etc. These words appear so frequently that their usefulness is limited. In Information Retrieval, for example, search engines ignore stop words at the time of searching a key phrase. In Information Extraction and Text Summarization also, stop words are pushed aside and treated as irrelevant information, in order to extract the most relevant and important information.

We have collected stop words by analyzing official newspapers, Myanmar grammar text books and CD versions of English-English- Myanmar (Students Dictionary) [24], English- Myanmar Dictionary [25], and The Khit Thit English-Myanmar dictionary [26]. We have also looked at stop word lists in English [27] and mapped them to equivalent stop words in Myanmar. See Table 3. As of now, our stop words list contains about 1216 entries. Stop words can be prefixes of other stop words leading to ambiguities. Usually, the longest matching stop word is the right choice.

As we keep analyzing texts, we can identify some words that can appear independently without combining with other words or suffixes. We build a list of such valid words and we keep adding new valid words as we progress through our segmentation process, gradually developing larger and larger lists of valid words. This list of known words can be made use of for hypothesizing candidate words as we go along.

Myanmar language uses a syllabic writing system [28] unlike English and many other western languages which use an alphabetic writing system. Interestingly, almost every syllable has a meaning in Myanmar language. This can also be seen from the work of Hopple [23].

We have developed scripts in Perl to syllabify words using our list of syllables and then generate n-gram statistics using Text::Ngrams which is developed by Vlado Keselj [29]. Example of collected Ngrams are shown in Table 4. We have used "type=word" option treating syllables as words. We had to modify this program a bit since Myanmar uses zero (as " ဝ [wa.] " letter) and the other special characters (",", ";", ">", ".", "&", "[", "]"etc.) which were being ignored in the original Text::Ngrams software. We collect all possible n-grams of syllables upto 5-grams. Almost all monograms are meaningful words. Many bigrams are also valid words and as we move towards longer n-grams, we generally get less and less number of valid words. We have used mutual information for even-syllables words and maximum entropy for odd-syllables to hypothesize possible words. Manual checking is essential to finally choose valid words.

There are lots of valid words which are not described in published dictionaries. The entries of words in the Myanmar-English dictionary which is produced by the Department of the Myanmar Language Commission are mainly words of the common Myanmar vocabulary. Most of the compound words have been omitted in the dictionary [1]. This can be seen in the preface and guide to the dictionary of the Myanmar-English dictionary produced by Department of the Myanmar Language

Commission, Ministry of Education. 4-syllables words like " ထူးထူးဆန်းဆန်း "[htu: htu: zan: zan:] (strange), " ထူးထူးကဲကဲ " [htu: htu: ke: ke:](outstanding) and " ထူးထူးခြားခြား " [htu: htu: gja: gja:](different) are not listed in dictionary although we usually use those words in every day life. Statistical construction of machine readable dictionaries has many advantages. New words which appear from time to time such as internet, names of medicines, can also be detected. Compounds words also can be seen.

With this technique, morphological structure of words also can be analyzed. See in Table 5. The above-mentioned three and four-syllables words are adverbs derived from the verbs " ထူးဆန်း "[htu: zan:], " ထူးကဲ "[htu: ke:], and " ထူးခြား "[htu: gja:]. Statistical dictionaries can be updated much more easily than published printed dictionaries, which need more time, cost and man power to bring out a fresh edition. Common names such as names of persons, cities, committees etc. can be also mined. Length statistics will be a useful hint and many researchers have used longest string matching [30],[31].

Table 3: Stop words of English Vs Myanmar

Prepositions and adverbs	
always	အမြဲ [a̲ mje:], အမြဲတမ်း [a̲ mje: dan:], အမြဲတစေ [a̲ mje: da̲ zei]
Nominative personal pronouns	
I	ကျွန်တော် [kjun do], ကျွန်မ [kja̲ ma.], ငါ [nga], ကျုပ် [kjou ′], ကျနော် [kja̲ no], ကျွန်ုပ် [kja̲nou ′], ကျမ [kja̲ ma.]
Accusative personal pronouns	
me	ကျွန်တော်အား [kjun do a:], ကျွန်တော်ကို [kjun do kou], ကျွန် မကို [kja̲ ma. gou] , ငါ့ကို [nga. gou], ကျုပ်ကို [kjou ′ kou], ကျွန်ုပ်ကို [kja̲nou ′ gou]
Reflexive personal pronouns	
myself	မိမိကိုယ်တိုင် [mi. mi. kou dain], မိမိဘာသာ [mi. mi. hpa dha], မိမိဘာသာ [mi. mi. hpa dha], ကိုယ်ကိုယ်တိုင် [kou kou dain], ကိုယ့်ဘာသာ [kou. hpa dha]
Relative pronouns	
That	သည့် [thi.], မည့် [mji.], တဲ့ [te.]
Possessive pronouns and adjectives	
my	ကျွန်ုပ်၏ [kjou ′ i.], ကျွန်တော်၏ [kjun do i.], ကျွန်မ၏ [kja̲ ma. i.], ကျနော်၏ [kja̲ nou ′ i.], ကျမ၏ [kja̲ ma. i.], ငါ့ရဲ့ [nga i.], ကျုပ်ရဲ့ [kjou ′ i.], ကျွန်ုပ်ရဲ့ [kjou ′ je.], ကျွန်တော့်ရဲ့ [kjun do je.], ကျွန်မရဲ့ [kja̲ ma. je.], ကျနော့်ရဲ့ [kja̲ nou ′ je.], ကျမရဲ့ [kja̲ ma. je.], ငါရဲ့ [nga je.], ကျုပ်ရဲ့ [kjou ′ je.], ကျွန်တော့် [kjun do.], ကျနော့် [kja̲ no.]
Demonstrative pronouns and adjectives	
this	အရာ [i a̲ ja], ဟောဒါ [ho: da], ဟောဒီ [ho: dhi]နီ [ ]
Indefinite pronouns and adjectives	
some	အချို့ [a̲ chou.], အချို့သော [a̲ chou. tho:], တချို့ [ta̲ chou.], တချို့သော [a̲ chou. tho:], တချို့ချို့ [ta̲ chou.ta̲ chou.], တချို့တလေ [ta̲ chou.ta̲ lei]

Continued on next page

Table 3 – continued from previous page

Conjunctions)	
and	နှင့် [hnin.], ပြီးလျှင် [pji: hljin], ၄င်းနောက် [la̱ gaun:]
Questions	
how	အဘယ်ကဲ့သို့ [a̱ be khe. dhou.], မည်ကဲ့သို့ [mji khe. dhou.], မည်သည့်နည်းနှင့် [mji dhi. ni: hnin.] , မည်သည့်နည်းဖြင့် [mji dhi. ni: hpji ′], မည်သို့ [mji dhou.], ဘယ်လိုလဲ [be lou le:], သို့ပေမည့် [dho. bei me.], မည်သည့်နည်းနှင့်မဆို [mji dhi. ni: hnin. ma̱ hsou], ဘယ်နည်းနှင့် [be ni: hnin.], မည်ရွေ့မည်မျှ [mji jwei. mji hmja.], အဘယ်မျှလောက် [a̱ be hmja. lau ′], ဘယ်လောက် [be lau ′]

Table 4: Example of collected Ngrams

No.	Bigram bisyllables	Trigram 3-syllables	4-gram 4-syllables
1.	ဖန်ထည် glassware [hpan de]	ဝါးခနဲ laughing or yawning loudly [wa: ga̱ ne:]	ငယ်ငယ်ကြီးကြီး young or old [nge nge kji: kji:]
2.	ဖန်တုံး glass stone [hpan toun:]	ဝုန်းခနဲ with an uproar [woun: ga̱ ne:]	ခါးခါးသီးသီး bitterly [kha: ga: thi: dhi:]
3.	ဖန်တီး create [hpan di:]	ဝေါခနဲ with an roar [wo: ga̱ ne:]	တောင့်တောင့်တင်းတင်း very stout [taun. taun. din: din:]
4.	ဖန်လာ happen [hpan la]	အားခနဲ a loud voice [a: ga̱ ne:]	နှစ်နှစ်သက်သက် like much [hni ′ hni ′ the ′ the ′]
5.	ဖန်အိမ် lantern [hpan ein]	ဗုန်းခနဲ with a big sound [boun: ga̱ ne:]	နှစ်နှစ်ကာကာ whole-heartedly [hni ′ hni ′ ka ga]
6.	ဖန်ဆင်း create with supernatural power [hpan zin:]	စွေခနဲ with squinted–eye [swei ga̱ ne:]	ပျစ်ပျစ်နှစ်နှစ် thick [pji ′ pji ′ hni ′ hni ′]
7.	ဖန်သား glassware [hpan tha:]	စွေ့ခနဲ effortlessly [swei. ga̱ ne:]	ဝုံ့ဝုံ့စားစား outbravely [wun. wun. sa: za:]
8.	ဖန်ခုန် game's name [hpan khoun]	ဆောင့်ခနဲ with stamping [hsaun. ga̱ ne:]	စွန့်စွန့်စားစား riskly [sun. sun. sa: za:]
9.	ဖန်ခွက် glass [hpan gwe ′]	မှေးခနဲ a short nap [hmei: ga̱ ne:]	စဉ်းစဉ်းစားစား thoughtfully [sin: sin: sa: za:]

Continued on next page

Table 4 – continued from previous page

No.	Bigram bisyllables	Trigram 3-syllables	4-gram 4-syllables
10.	ကန်စောင်း bank of lake [kan saun:]	ထောင်းခနဲ fuming with rage [htaun: ga̲ ne:]	များများစားစား many,much [mja: mja: sa: za:]

A basic unit 1 syllable	B (Verb)= A + သည်	C (Noun)= အ+A	D (Negative)= မ+A+ ဘူး	E (Noun)= A+ မှု
ကောင်း [kaun:] good (Adj)	ကောင်းသည် [kaun: thi] is good	အကောင်း [a̲ kaun:] good	မကောင်းဘူး [ma. kaun: bu:] Not good	ကောင်းမှု [kaun: mhu.] good deeds
ဆိုး [hso:] bad (Adj)	ဆိုးသည် [hso: thi] is bad	အဆိုး [a̲ hso:] bad	မဆိုးဘူး [ma. hso: bu:] Not bad	ဆိုးမှု [hso: mhu.] Bad Deeds
ရောင်း [jaun:] sell(Verb)	ရောင်းသည် [jaun: thi] sell	အရောင်း [a̲ jaun:] sale	မရောင်းဘူး [ma. jaun: bu:] not sell	ရောင်းမှု [jaun: mhu.] sale
ရေး [jei:] write(Verb)	ရေးသည် [jei: thi] write	အရေး [a̲ jei:] writing	မရေးဘူး [ma. jei: bu:] do not write	ရေးမှု [jei: mhu.]
ပြော [pjo:] talk,speak(Verb)	ပြောသည် [pjo: thi] talk,speak	အပြော [a̲ pjo:] talk,speech	မပြောဘူး [ma. pjo: bu:] not talk,speak	ပြောမှု [pjo: mhu.] a talk

Table 5: An example patterns of Myanmar Morphological Analysis

3 Syllabification and Word Segmentation

Since dictionaries and other lexical resources are not yet widely available in electronic form for Myanmar language, we have collected possible syllables (including ဓမ္မတက္က) and 2 lakhs Myanmar word-lists. With the help of these stored syllables and word lists, we have done Syllabification and word segmentation. The first step to build a word hypothesizer is Syllabification of the input text by looking up syllable lists. In second step, we exploit lists of words (n-grams at syllable level) for word segmentation from left to right.

Myanmar Natural Language Processing Group has listed 1894 syllables that can appear in Myanmar texts. We have observed that there are some more syllables, especially in foreign words including Pali and Sanskrit words which are widely used in Myanmar. We have collected other possible syllables using Myanmar-English dictionary. As we collected texts from internet which has lack of standard typing sequences, we also collected different possible typing sequences of syllables which will be seen same appearance. Following is an example of syllables in different typing

sequences. Now we have over 4000 syllables in our list.We have developed scripts in Perl to syllabify words using Longest String Matching and our list of syllables.

On screen	ဦး	ကို
	ဦး	ကို
In ascii	MuD:	udk
	BuD:	ukd

Table 6: Syllable with different typing sequences

In Myanmar Text Syllabification, Longest string matching alone can be handled. The table 6 also shows that different Typing sequences of syllables are detected.Failure caused due to
1. the combination of the writing sequences (typing sequences) of syllables
2. foreign words borrowed from **English** and **Pali**
3. the need of new syllables entries which are rarely used

3.1 Longest String Matching

Naive Method: Pseudo code In this matching, it goes from left-to-right scan in greedy manner.

```
1.  Load the set of syllables from syllable-file
2.  Load the sentences to be processed from sentence-file
3.  Store all syllables of length j in N_j where j = 10..1
4.  for-each sentence do
5.      length ← length of the sentence
6.      pos ← 0
7.      while (length > 0) do
8.          for j = 10..1 do
9.              for-each syllable in N_j do
10.                 if string-match sentence(pos, pos + j) with syllable
11.                     Syllable found. Mark syllable
12.                     pos ← pos + j
13.                     length ← length − j
14.                 End if
15.             End for
16.     End while
17. Print syllabified string
18. End for
```

Similarly, we have done tokenization with longest syllable matching using collected 2 lakhs words list. A example sentence segmentation is given in Table 7. Even though Syllabification can be done with longest string matching, tokenization has needed to be improved with Hidden Marchov Models (HMM) like machine learning techniques in order to get perfect system. Research is still going on. We have achieved about 99% accuracy in Syllabification and 80% accuracy in word segmentation [32].

ကျောင်းအုပ်ဆရာကြီးသည် အကြမ်းဖက်မှုကို စက်ဆုပ်သည်				
ကျောင်းအုပ်ဆရာကြီး	သည်	အကြမ်းဖက်မှု	ကို	စက်ဆုပ်သည်
[kyaung: aop hsa ya kyi:]	[thi]	[a kyan: phak mhu]	[ko]	[sak sop thi]
The headmaster		violence		abhors
N_{subj}	Particle	N_{obj}	Particle	$V_{present}$

Table 7: A sentence being segmented into words

4 Preliminary Analysis of Myanmar Corpora

Here we present a preliminary analysis of Myanmar corpus developed in University of Hyderabad. This includes news, novels, online magazines, and free and trail text of online bookshop. We have approximately 18 lakhs sentences but we have carried out analysis for only 150,000 sentences in this work.

4.1 Type-Token Analysis

The figure 1 shows the results of a type-token growth rate analysis. Each distinct word form is a type and each occurrence of a type counts as a token. If we analyze the entire corpus in one go, we will get the total number of types, total number of tokens and the global type-token ratio. Instead, if we perform type-token analysis incrementally, by starting with a small randomly selected part of the corpus and iteratively adding more texts randomly, we get a type-token growth rate curve that shows how many new types will be found as the corpus size increases.

Note that by types we mean fully inflected word forms, not root forms or citation forms found in dictionaries. Also, compounding will have their effect and the tokens we get do not necessarily correspond to the linguistic definition of a word understood in semantic terms. There is no automatic way to extract words based on meaning. Wide coverage, high performance, robust morphological analyzers are not yet available in most languages under study and here we restrict our analyses to full words. From figure 1, we can see that the curve is not saturated which tells

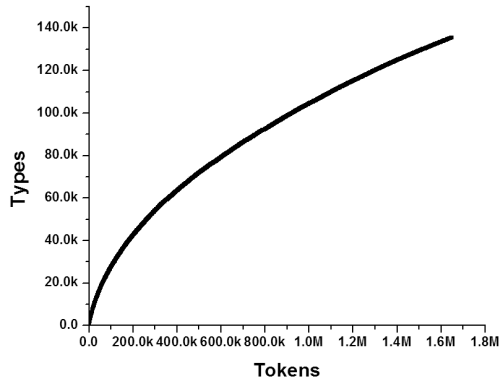

Figure 1: Type-Token Growth Rate Analysis of Myanmar Corpora

us that we need to analyze more corpus to understand the behavior of Myanmar

langauge.

4.2 Coverage Analysis

Table 8: Self Coverage Analysis of Myanmar Corpus

%Coverage	Approx. No. of Types
50	1300
60	2700
70	5700
80	12200
85	18900
90	31300
95	60300
96	69600
97	86100
98	102600
99	119100
100	135518

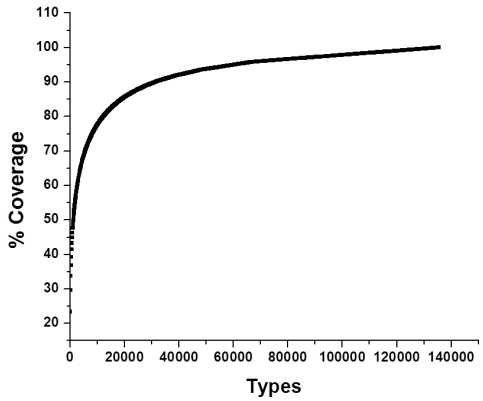

Figure 2: Coverage Analysis of Myanmar Corpora

Coverage analysis deals with the examination of how much of a corpus can be covered by a given set of types. We perform a type-token analysis and prepare a list of types sorted in decreasing order of frequency of occurrence. By thresholding on this list, we can select the most frequent n words in the language, for any given value of n. We then explore what percentage of words in a corpus are found in the list so selected. Here we perform self-coverage analysis - coverage analysis on the same corpus from which the words are extracted. (It would be instructive to perform coverage analysis on other corpora as and when they become available.)

From the figure 2,we can see that about 1300 most frequent words are sufficient to give about 50% coverage of the corpus. 60% coverage can be obtained by just the first 2700 words or so. This being a self-coverage analysis 100% coverage can be obtained by using all the words in the word list.

4.3 Sentence Length Analysis

Figure 3 shows the sentence length distribution for Myanmar corpus.

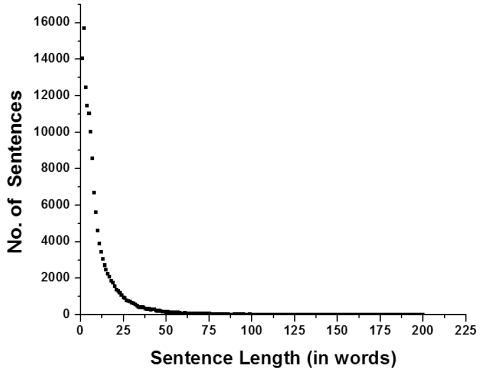

Figure 3: Word Length Analysis in terms of Syllables

4.4 Word Length Variation with Frequency

Table 9: Words and Syllable Structure

No. of syllables	No of words	Example
1	3776	ကောင်း
		Good (Adj)
		[kaun:]
2	45063	လိပ်ပြာ
		Butterfly, Soul (N)
		[lei ′ pja]
3	72795	ပြတင်းပေါက်
		Window (N)
		[b a̲ din: bau ′]
4	48636	ပြည်တွင်းထုတ်ကုန်
		Domestic Product (N)
		[pji dwin: htou ′ koun]
5	28932	လျှပ်စစ်ထမင်းအိုး
		[hlja ′ si ′ ht a̲ min: ou:]
		Rice Cooker(N)

Continued on next page

Table 9 – continued from previous page

No. of syllables	No of words	Example
6	16485	သူနာပြုဆရာမ
		Nurse(female) (N)
		[thu na bju. hs a̲ ja ma.]
7	9013	ရင်းနှီးသွားကြပေတော့သည်
		become friend (V)
		[jin: hni: thwa: kya. pei to. thi]
8	4620	ပြည်ထောင်စုမြန်မာနိုင်ငံတော်
		Union of Myanmar (N)
		[pji daun zu. mj a̲ ma nain gan to]
9	2404	သံယံဇာတအရင်းအမြစ်များ
		Natural Resources (N)
		[than jan za ta. a̲ jin: a̲ mji ′]
10	1209	ခြေမကိုင်မိလက်မကိုင်မိဖြစ်သည်
		be agitated or shaken(V)
		[chei ma̲ kain mi. le ′ ma̲ kain mi. hpji ′ thi]

Figure 4 shows the word length distribution for Myanmar corpus. Words that occur

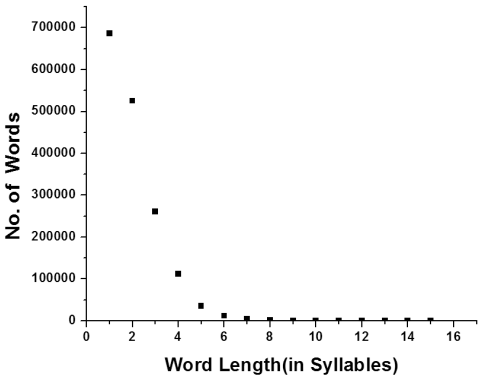

Figure 4: Word Length in relation to Word Frequency

frequently tend to be small words. It is therefore interesting to explore the relation between word frequency and word length. Figure 4 shows the scatter diagram of word length measured in word frequency. Word length is averaged over all words of a given frequency. It can be seen that the least frequent words are larger and word length shows a gradual decrease as we move towards more frequent words. High frequency words show a greater spread in terms of word length. Yet we can see a trend - words tend to become smaller and smaller as we move towards the most frequent words. The streaks that we see are due to clustering effect due to averaging.

5 Entropy, Perplexity

Entropy is a measure of information content. Entropy is related to probability, redundancy and uncertainty and is thus invaluable in language analysis. The more we know about something, the lower the entropy will be because we are less surprised by the outcome of a trial. Entropy can be interpreted as the minimum number of bits required to encode a given piece of information. Entropy can be calculated using the formula

$$H(X) = \sum_{x=1}^{N} -p(x) \log_2 p(x)$$

where N is the number of Word Types in the language.

H-maximum will be obtained when the probabilities of all the words in the corpus are same.

$H_{max} = \log_2 N$

$H_{Relative} = \frac{H_{actual}}{H_{max}}$

$Redundancy = \frac{H_{max} - H_{actual}}{H_{max}}$

Perplexity is useful for evaluating language models. A perplexity of k means that you are as surprised on average as you would have been if you had had to guess between k equiprobable choices at each step. Perplexity of the given model can be evaluated by

$P(x) = 2^{H(x)}$

where H(x) is the entropy of the given model.

The values of the Entropy and Perplexity for the corpus are shown in table 7.

Table 10: Entropy Analysis of Myanmar Corpus

Entropy	13.247
Relative Entropy	0.777
Redundancy	0.222
Perplexity	9725.432

6 Conclusions

In this paper we have described syllabification which is important starting point to identify the words in the text, word tokenization, Myanmar stop-words, introducing

Myanmar morphological formation and statistical analyses of a fairly large text corpus of Myanmar.Word Segmentation is performed with longest syllable matching by looking up the dictionary. We have checked manually the outputs of syllabification and word segmentation modules. We have obtained 99% accuracy on syllabification and 80% on word segmentation. It is perhaps for the first time that a statistical analysis has been carried out on Myanmar language. These analyses point to issues relating to technology development as also to detailed linguistic analysis necessary for a complete understanding of the language. Larger corpora are needed in Myanmar language for meaningful analysis and technology development. Syllabification is done with longest string matching using syllable list.

References

[1] *Myanmar-English Dictionary.* Department of the Myanmar Language Commission,Ministry of Education, Union of Myanmar.

[2] Y. K. Thu and Y. Urano, "Text entry for myanmar language sms: Proposal of 3 possible input methods, simulation and analysis," February 2006.

[3] J. Sinclair, *Corpus, Concordance, Collocation.* Oxford University Press, Oxford, 1991.

[4] M. Barlow, "Corpora for theory and practice," *International journal of Corpus linguistics*, vol. 1, no. 1, pp. 1–38, 1996.

[5] I. Lancashire, C. Percy, and C. Mayer, *Synchronic Corpus linguistics.* Rodopi, Amsterdam, Atlanta, 1996.

[6] W. Teubert, "Corpus linguistics: A partisan view," *International journal of Corpus linguistics*, vol. 4, no. 1, pp. 1–16, 2000.

[7] N. Oostdijk and P. Hann, *Corpus based research into language.* Rodopi, Amsterdam, Atlanta, 1994.

[8] D. Biber, "Investigating language use through corpus-based analyses of association patterns," *International journal of Corpus linguistics*, vol. 1, no. 2, pp. 171–198, 1996.

[9] D. Biber, S.Conrad, and R. Reppen, *Corpus Linguistics : Investigating language strcture and use.* Cambridge Unviersity Press,Cambridge, 1998.

[10] C. Mair and M. Hundt, *Corpus Linguistics and Linguistics theory.* Rodopi, Amsterdam, Atlanta, 2000.

[11] M. Stubbs, *Texts and Corpus analysis.* Oxford: Blackwell publishers, 1996.

[12] H. Kucera and W. N. Francis, *Computational Analysis of present-day American English.* Brown University Press, 1967.

[13] J. Aarts and W. Meijs, *Corpus Linguistics: Recent development in the use of Computer corpora in Enlgish Language Research.* Rodopi, Amsterdam, Atlanta, 1984.

[14] S. Johansson and A. B. Stenstrom, *English computer corpora: Selected papers and research guide*. Mouton de Gruyter, Berlin, 1991.

[15] G. Knowles, B. J. Williams, and L. Taylor, *A corpus of formal British English speech: The Lancaster/IBM spoken English Corpus*. Longman, London, 1997.

[16] M. Ljung, *Corpus-based studies in English*. Rodopi, Amsterdam, Atlanta, 1997.

[17] A. C. F. Meyer, *English Corpus Linguistics*. Cambridge Unviersity Press,Cambridge, 2002.

[18] R. Garside, G. Leech, and G. Sampson, *The computational analysis of English: A corpus based approach*. Longman, London, 1987.

[19] G. B. Kumar, K. N. Murthy, and B.B.Chaudhuri, "Statistical analyses of telugu text corpora," *To Appear in International journal of Dravidian Langauges (IJDL)*, vol. 36, no. 2, 2007.

[20] T. McEnery and A. Wilson, *Corpus Linguistics*. Edinbourgh University Press: Edinbourgh, 1996.

[21] G. Kennedy, *An introduction to Corpus Linguistics*. Addison-Wesley, 1998.

[22] D. Biber, "Representativeness in corpus design," *Literary and Linguistic computing*, vol. 8, no. 4, pp. 243–257, 1993.

[23] P. Hopple, *The structure of nominalization in Burmese,Ph.D thesis*. may 2003.

[24] "Student's english-english/myanmar dictionary." Ministry of Commerce and Myanmar Inforithm Ltd, Union of Myanmar, CD version, Version 1, 1999.

[25] "English-myanmar dictionary." Ministry of Education, Union of Myanmar,CD version.

[26] S. U Soe, *The Khit Thit English- English-Myanmar Dictionary with Pronunciation*. Yangon, Myanmar, apr 2000.

[27] http://www.syger.com/jsc/docs/stopwords/english.htm.

[28] K. N. Murthy, *Natural Language Processing - an Information Access Perspective*. Prentice Hall Engineering, Science and Math., 2006.

[29] V. Keselj, "Text ::ngrams." http://search.cpan.org/ vlado/ Text-Ngrams-1.8/.

[30] R. Anglell, G. Freund, and P. Willett, "Automatic spelling correction using a trigram similarity measure," *Information Processing & Management*, vol. 19, no. 4, pp. 305–316, 1983.

[31] P. et al., "Targeted s-gram matching: a novel n-gram matching technique for cross- and monolingual word form variants," *Information Research*, vol. 7, pp. 235–237, january 2001.

[32] H. H. Htay and K. N. Murthy, "Myanmar word segmentation," in *Fourth International Conference on Computer Applications*, (Yangon, Myanmar), pp. 353–357, February 2006.