# Myanmar Word Segmentation using Syllable level Longest Matching

**Hla Hla Htay, Kavi Narayana Murthy**
Department of Computer and Information Sciences
University of Hyderabad, India
hla_hla_htay@yahoo.co.uk, knmuh@yahoo.com

## Abstract

***In* Myanmar *language, sentences are clearly delimited by a unique sentence boundary marker but are written without necessarily pausing between words with spaces. It is therefore non-trivial to segment sentences into words. Word tokenizing plays a vital role in most Natural Language Processing applications. We observe that word boundaries generally align with syllable boundaries. Working directly with characters does not help. It is therefore useful to syllabify texts first. Syllabification is also a non-trivial task in Myanmar. We have collected 4550 syllables from available sources . We have evaluated our syllable inventory on 2,728 sentences spread over 258 pages and observed a coverage of 99.96%. In the second part, we build word lists from available sources such as dictionaries, through the application of morphological rules, and by generating syllable n-grams as possible words and manually checking. We have thus built list of 800,000 words including inflected forms. We have tested our algorithm on a* 5000 *sentence test data set containing a total of (*35049 *words) and manually checked for evaluating the performance. The program recognized* 34943 *words of which* 34633 *words were correct, thus giving us a Recall of 98.81%, a Precision of 99.11% and a F-Measure is 98.95%.***

**Key Words:-** Myanmar, Syllable, Words, Segmentation, Syllabification, Dictionary

## 1 Introduction

Myanmar (*Burmese*) is a member of the Burmese-Lolo group of the Sino-Tibetan language spoken by about 21 Million people in Myanmar (Burma). It is a tonal language, that is to say, the meaning of a syllable or word changes with the tone. It has been classified by linguists as a mono-syllabic or isolating language with agglutinative features. According to history, Myanmar script has originated from *Brahmi* script which flourished in India from about 500 B.C. to over 300 A.D (MLC, 2002). The script is syllabic in nature, and written from left to right.

Myanmar script is composed of 33 consonants, 11 basic vowels, 11 consonant combination symbols and extension vowels, vowel symbols, devowelizing consonants, diacritic marks, specified symbols and punctuation marks(MLC, 2002),(Thu and Urano, 2006). Myanmar script represents sequences of syllables where each syllable is constructed from consonants, consonant combination symbols (i.e. Medials), vowel symbols related to relevant consonants and diacritic marks indicating tone level.

Myanmar has mainly **9** parts of speech: noun, pronoun, verb, adjective, adverb, particle , conjunction, post-positional marker and interjection (MLC, 2005), (Judson, 1842).

In Myanmar script, sentences are clearly delimited by a sentence boundary marker 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 post-position. In fact in the past spaces were rarely used. Segmenting sentences into words is therefore a challenging task.

Word boundaries generally align with syllable boundaries and syllabification is therefore a useful strategy. In this paper we describe our attempts on syllabification and segmenting Myanmar sentences into words. After a brief discussion of the corpus collection and pre-processing phases, we describe our approaches to syllabification and tokenization into words.

Computational and quantitative studies in Myanmar are relatively new. Lexical resources available are scanty. Development of electronic dictionaries and other lexical resources will facilitate Natural Language Processing tasks such as Spell Checking, Machine Translation, Automatic Text summarization, Information Extraction, Automatic Text Categorization, Information Retrieval and so on (Murthy, 2006).

Over the last few years, we have developed *monolingual text corpora* totalling to about 2,141,496 sentences and *English-Myanmar parallel corpora* amounting to about 80,000 sentences and sentence fragments, aligned at sentence and word levels. We have also collected word lists from these corpora and also from available dictionaries. Currently our *word list* includes about 800,000 words including inflected forms.

## 2 Myanmar Words

Myanmar words are sequences of syllables. The syllable structure of Burmese is C(G)V((V)C), which is to say the onset consists of a consonant optionally followed by a glide, and the rhyme consists of a monophthong alone, a monophthong with a consonant, or a diphthong with a consonant [1]. Some representative words are:

- CV [mei] girl
- CVC [me ′] crave
- CGV [mjei] earth
- CGVC [mje ′] eye
- CVVC [maun] (term of address for young men)
- CGVVC [mjaun] ditch

Words in the Myanmar language can be divided into simple words, compound words and complex words (Tint, 2004),(MLC, 2005),(Judson, 1842). Some examples of compound words and loan words are given below.

- Compound Words
  - head [u:] ဦး + pack [htou ′] ထုပ် = hat [ou ′ htou ′] ဦးထုပ်
    language [sa] စာ+ look,see [kji.] ကြည့် + [tai ′] building တိုက် = library [sa kji. dai ′] စာကြည့်တိုက်
  - sell [yaun:] ရောင်း + buy [we] ဝယ် = trading ရောင်းဝယ် $[yaun : we]$
- Loan Words
  - ကွန်ပျူတာ [kun pju ta] computer
  - ဆပ်ကော်မတီ [hsa ′ ko ma ti] sub-committee
  - ချယ်ရီ [che ri] cherry

## 3 Corpus Collection and Preprocessing

Development of lexical resources is a very tedious and time consuming task and purely manual approaches are too slow. We have downloaded Myanmar texts from various web sites including news sites including official newspapers, on-line magazines, trial e-books (over *300* full books) as well as free and trial texts from on-line book stores including a variety of genres, types and styles - modern and ancient, prose and poetry, and example sentences from dictionaries. As of now, our corpus includes *2,141,496* sentences.

The downloaded corpora need to be cleaned up to remove hypertext markup and we need to extract text if in pdf format. We have developed the necessary scripts in Perl for this. 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 and these pre-processed corpora are seen to be reasonably clean.

[1] http://en.wikipedia.org/wiki/Burmese_language

## 4 Collecting Word Lists

Electronic 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. Word lists and dictionaries in electronic form are of great value in computational linguistics and NLP. Here we describe our efforts in developing a large word list for Myanmar.

### 4.1 Independent Words

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. We have also collected from sources such as Myanmar Orthography(MLC, 2003), CD versions of English-English-Myanmar (Student's Dictionary)(stu, 2000) and English-Myanmar Dictionary (EMd, ) and Myanmar-French Dictionary (damma sami, 2004). Currently our word list includes 800,000 words.

### 4.2 Stop Word Removal

Stop words include prepositions/post-positions, conjunctions, particles, inflections etc. which appear as suffixes added to other words. They form closed classes and hence can be listed. Preliminary studies therefore suggested that Myanmar words can be recognized by eliminating these stop words. Hopple (Hopple, 2003) also notices that particles ending phrases can be removed to recognize words in a sentence. We have collected stop words by analyzing official newspapers, Myanmar grammar text books and CD versions of English-English-Myanmar (Student's Dictionary)(stu, 2000), English-Myanmar Dictionary (EMd, ) and The Khit Thit English-Myanmar dictionary (Saya, 2000). We have also looked at stop word lists in English (www.syger.com, ) and mapped them to equivalent stop words in Myanmar. See Table 1. As of now, our stop words list contains 1216 entries. Stop words can be prefixes of other stop words leading to ambiguities. However, usually the longest matching stop word is the right choice.

Identifying and removing stop words does not always necessarily lead to correct segmentation of sentences into words. Both under and over segmentation are possible. When stop-words are too short, over segmentation can occur. Under segmentation can occur when no stop-words occur between words. Examples of segmentation can be seen in Table 2. We have observed that over segmentation is more frequent than under segmentation.

| **Nominative personal pronouns** | |
|---|---|
| I | ကျွန်တော်[kjun do], ကျွန်မ[kja ma.], ငါ[nga], ကျုပ်[kjou′], ကျနော်[kja no], ကျွန်ုပ်[kjanou′], ကျမ[kja ma.] |
| **Possessive pronouns and adjectives** | |
| my | ကျွန်ုပ်၏[kjou′ i.],ကျွန်တော်၏[kjun do i.], ကျွန်မ၏[kja ma. i.],ကျနော်၏[kja nou′ i.], ကျမ၏[kja ma. i.],ငါ့ရဲ့[nga i.],ကျုပ်ရဲ့[kjou′ i.], ကျွန်ုပ်ရဲ့[kjou′ je.],ကျွန်တော့်ရဲ့[kjun do je.], ကျွန်မရဲ့[kja ma. je.],ကျနော့်ရဲ့[kja nou′ je.], ကျမရဲ့[kja ma. je.],ငါရဲ့[nga je.], ကျုပ်ရဲ့[kjou′ je.],ကျွန်တော့်[kjun do.], ကျနော့်[kja no.] |
| **Indefinite pronouns and adjectives** | |
| some | အချို့[a chou.],အချို့သော[a chou. tho:], တချို့[ta chou.],တချို့သော[a chou. tho:], တချို့ချို့[ta chou.ta chou.], တချို့တလေ[ta chou.ta lei] |

Table 1: Stop-words of English Vs Myanmar

| ဝိုင်းဝန်းချီးကျူးခံရသဖြင့်သူအနေကသည် | | |
|---|---|---|
| ဝိုင်းဝန်းချီးကျူးခံရ | | အနေက |
| [waing: win: chi: kyu: khan ya] | | [a nay khak] |
| received compliments | | abashed |
| $V_{pp}$ | | $V_{past}$ |
| ကျောင်းအုပ်ဆရာကြီးသည်အကြမ်းဖက်မှုကိုစက်ဆုပ်သည် | | |
| ကျောင်းအုပ်ဆရာကြီး | အကြမ်းဖက်မှု | စက်ဆုပ် |
| [kyaung: aop hsa ya kyi:] | [a kyan: phak mhu] | [sak sop] |
| The headmaster | violence | abhors |
| $N_{subj}$ | $N_{obj}$ | $V_{present}$ |

Table 2: Removing stop-words for segmentation

### 4.3 Syllable N-grams

Myanmar language uses a syllabic writing system 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 (Hopple, 2003).

Myanmar Natural Language Processing Group has listed 1894 syllables that can appear in Myanmar texts (Htut, 2001). We have observed that there are more syllables in use, especially in foreign words including Pali and Sanskrit words which are widely used in Myanmar. We have collected other pos-

sible syllables from the Myanmar-English dictionary(MLC, 2002). Texts collected from the Internet show lack of standard typing sequences. There are several possible typing sequences and corresponding internal representations for a given syllable. We include all of these possible variants in our list. Now we have over **4550** syllables.

| Bigram<br>bisyllables | Trigram<br>3-syllables | 4-gram<br>4-syllables |
|---|---|---|
| ဖန်အိမ်<br>lantern<br>[hpan ein] | ဗုန်းခနဲ<br>with a big sound<br>[boun: ga. ne:] | နှစ်နှစ်ကာကာ<br>whole-heartedly<br>[hni ′ hni ′ ka ga] |
| ဖန်သား<br>glassware<br>[hpan tha:] | ဆွဲ့ခနဲ<br>effortlessly<br>[swei. ga. ne:] | ထူးထူးကဲကဲ<br>outstanding<br>[htu: htu: ke: ke:] |
| ကန်စောင်း<br>bank of lake<br>[kan saun:] | ထောင်းခနဲ<br>fuming with rage<br>[htaun: ga. ne:] | များများစားစား<br>many,much<br>[mja: mja: sa: za:] |

Table 3: Examples of Collected N-grams

| No. of syllables | No of words | Example |
|---|---|---|
| 1 | 4550 | ကောင်း |
| | | Good (Adj) |
| | | [kaun:] |
| 2 | 59964 | လိပ်ပြာ |
| | | Butterfly, Soul (N) |
| | | [lei ′ pja] |
| 3 | 170762 | ပြတင်းပေါက် |
| | | Window (N) |
| | | [b a. din: bau ′ ] |
| 4 | 274775 | ပြည်တွင်းထုတ်ကုန် |
| | | Domestic Product (N) |
| | | [pji dwin: htou ′ koun] |
| 5 | 199682 | လျှပ်စစ်ထမင်းအိုး |
| | | [hlja ′ si ′ ht a. min: ou:] |
| | | Rice Cooker(N) |
| 6 | 99762 | သူနာပြုဆရာမ |
| | | Nurse(female) (N) |
| | | [thu na bju. hs a. ja ma.] |
| 7 | 41499 | ရင်းနှီးသွားကြပေတော့သည် |
| | | become friend (V) |
| | | [jin: hni: thwa: kya. pei to. thi] |
| 8 | 14149 | ပြည်ထောင်စုမြန်မာနိုင်ငံတော် |
| | | Union of Myanmar (N) |
| | | [pji daun zu. mj a. ma nain gan to ] |
| 9 | 4986 | သဘာဝတအရင်းအမြစ်များ |
| | | Natural Resources (N) |
| | | [than jan za ta. a. jin: a. mji ′ ] |
| 10 | 1876 | ခြေမကိုင်မိလက်မကိုင်မိဖြစ်သည် |
| | | be agitated or shaken(V) |
| | | [ chei ma. kain mi. le ′ ma. kain mi. hpji ′ thi] |

Table 4: Syllable Structure of Words

We have developed scripts in Perl to syllabify words using our list of syllables as a base and then generate n-gram statistics using Text::Ngrams which is developed by Vlado Keselj (Keselj, 2006). This program is quite fast and it took only a few minutes on a desktop PC in order to process 3.5M bytes of Myanmar texts. We have used "-type=word" option treating syllables as words. We had to modify this program a bit since Myanmar uses zero (as "(0) wa " letter) and the other special characters ( ",", "<", ">", ".", "&", "[","]" etc.) which were being ignored in the original Text::Ngrams software. We collect all possible words which is composed of n-grams of syllables up to 5-grams. Table 1 shows some words which are collected through n-gram analysis. Almost all monograms are meaningful words. Many bi-grams are also valid words and as we move towards longer n-grams, we generally get less and less number of valid words. See Table 3. Further, frequency of occurrence of these n-grams is a useful clue. See Table 4.

By analyzing the morphological structure of words we will be able to analyze inflected and derived word forms. A set of morphemes and morphological forms have been collected from (MLC, 2005) and (Judson, 1842) . See Table 5. For example, the four-syllable word in Table 3 is an adverb "ထူးထူးကဲကဲ" [htu: htu: ke: ke:] outstanding derived from the verb "ထူးကဲ". See Table 3.

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. Common names such as names of persons, cities, committees etc. can be also mined. Once sufficient data is available, statistical analysis can be carried and techniques such as mutual information and maximum entropy can be used to hypothesize possible words.

### 4.4 Words from Dictionaries

Collecting words using the above three mentioned methods has still not covered all the valid words in our corpus. We have got only 150,000 words. Words collected from n-grams needs exhaustive human effort to pick the valid words. We have therefore collected words from two on-line dictionaries - the English-English-Myanmar (Student's Dictionary) (stu, 2000), English-Myanmar Dictionary (EMd, ) and from two e-books - French-Myanmar(damma sami, 2004), and Myanmar Orthography (MLC, 2003). Fortunately, these texts can be transformed into winninnwa font. We have

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

Table 5: Example patterns of Myanmar Morphological Analysis

written Perl scripts to convert to the standard font. Myanmar Spelling Bible lists only lemma (root words). We have suffixed some frequently used morphological forms to these root 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 (MLC, 2002). 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)(see Table 3) are not listed in dictionary although we usually use those words in every day life.

With all this, we have been able to collect a total of about 800,000 words. As we have collected words from various sources and techniques, we believe we have fairly good data for further work.

| On screen | ကြီး | ကို |
|---|---|---|
| | ကြီး: | ကို |
| In ascii | MuD: | udk |
| | BuD: | ukd |

Table 6: Syllables with different typing sequences

## 5 Syllabification and Word Segmentation

Since dictionaries and other lexical resources are not yet widely available in electronic form for Myanmar language, we have collected 4550 possible syllables including those used in Pali and foreign words such as ဓမ္မတက္ကသီလ ), considering different typing sequences and corresponding internal representations, and from the 800,000 strong Myanmar word-list we have built. With the help of these stored syllables and word lists, we have carried out syllabification and word segmentation as described below. Many researchers have used longest string matching (Angell et al., 1983),(Ari et al., 2001) and we follow the same approach.

The first step in building a word hypothesizer is syllabification of the input text by looking up syllable lists. In the second step, we exploit lists of words (viewed as n-grams at syllable level) for word segmentation from left to right.

### 5.1 Syllabification

As an initial attempt we use longest string matching alone for Myanmar text syllabification. Examples are shown in Table 7.

**Pseudo code** Here we go from left-to-right in a greedy manner:

**sub** *syllabification*{
**Load** the set of syllables from syllable-file
**Load** the sentences to be processed from sentence-file
**Store** all syllables of length $j$ in $N_j$ where $j = 10..1$
**for-each** *sentence* **do**
length ← length of the sentence

```
    pos ← 0
    while (length > 0) do
        for j = 10..1 do
            for-each syllable in N_j do
                if string-match sentence(pos, pos + j) with syllable
                    Syllable found. Mark syllable
                    pos ← pos + j
                    length ← length − j
                End if
            End for
        End for
    End while
Print syllabified string
End for
}
```

We have evaluated our syllables list on a collection of 11 short novels entitled "*Orchestra*" သံစုံတီးဝိုင်း[than zoun ti: wain:], written by **"Nikoye"** (Ye, 1997) which includes 2,728 sentences spread over 259 pages including a total of 70,384 syllables. These texts were syllabified using the longest matching algorithm over our syllable list and we observed that only 0.04% of the actual syllables were not detected. The Table 6 shows that different typing sequences of syllables were also detected. Here are some examples of failure: ဖိုယ်း[rkdCf;]and ဖိုလ်း[rkdvf;] which are seldom used in text. The typing sequence is also wrong. Failures are generally traced to

- differing combinations of writing sequences
- loan words borrowed from foreign languages
- rarely used syllables not listed in our list

### 5.2 Word Segmentation

*W*e have carried out tokenization with longest syllable word matching using our 800,000 strong stored word list. This word list has been built from available sources such as dictionaries, through the application of morphological rules, and by generating syllables n-grams and manually checking. An example sentence and its segmentation is given in Table 8.

```
Load the set of words from word-file
    for-each word do
        i ← syllabification(word);
        Store all words of syllable length i in N_i where i = 10..1
    End for

Load the sentences to be processed from sentence-file
for-each sentence do
    length ←syllabification(sentence);
    #length of the sentence in terms of syllables
    pos ← 0
    while (length > 0) do
        for j = 10..1 do
            for-each word in N_j do
                if string-match sentence(pos, pos + j) with word
                    word found. Mark word
                    pos ← pos + j
                    length ← length − j
                End if
            End for
        End for
    End while
Print tokenized string
End for
```

## 6 Evaluation and Observations

We have segmented *5000* sentences including a total of (*35049* words) with our programs and manually checked for evaluating the performance. These sentences are from part of the English-Myanmar parallel corpus being developed by us (Htay et al., 2006). The program recognized *34943* words of which *34633* words were correct, thus giving us a Recall of 98.81% and a Precision of 99.11%. The F-Measure is 98.95%. The algorithm suffers in accuracy in two ways:

**Out-of-vocabulary Words:** Segmentation error can occur when the words are not listed in dictionary. No lexicon contains every possible word of a language. There always exist out-of-vocabulary words such as new derived words, new compounds words, morphological variations of existing words and technical words (Park, 2002). In order to check the effect of out-of-vocabulary words, we took a new set of 1000 sentences (7343 words). We have checked manually and noticed *329* new words, that is about 4% of the words are not found in our list, giving us a coverage of about 96%.

**Limitations of left-to-right processing:**
Segmentation errors can also occur due to the limitations of the left-to-right processing. See the example 1 in Table 9. The algorithm suffers most in recognizing the sentences which have the word *He* သူ [thu] followed by a *negative verb* starting with the particle မ[ma.]. The program wrongly segments *she* as *he*. Our text collection obtains from various sources and the word "she" is used as သူမ [thu ma.] in modern novels and Internet text. Therefore, our

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ကော်ဖီသောက်ရင်းအနံတိုနှုင်အလပသလပပြောနေခဲ့သည် | | | | | | | | | | | | | | | | |
| aumfzDaomuf&if;tefwDESihftvyovyajymaecJhonf | | | | | | | | | | | | | | | | |
| ကော် | ဖီ | သောက် | ရင်း | အနံ | တို | နှုင် | အ | လ | ပ | သ | လ | ပ | ပြော | နေ | ခဲ့ | သည် |
| aumf | zD | aomuf | &if; | tef | wD | ESihf | t | v | y | o | v | y | ajym | ae | cJh | onf |
| [ko] | [hpi] | [thau′] | [jin:] | [an] | [ti] | [hnin.] | [a] | [la] | [pa.] | [tha.] | [la] | [pa.] | [pjo] | [nei] | [khe.] | [thi] |
| Having his coffee, he chit-chat with the lady. | | | | | | | | | | | | | | | | |

Table 7: Example syllabification

| ကျောင်းအုပ်ဆရာကြီးသည်အကြမ်းဖက်မှုကိုစက်ဆုပ်သည် | | | | |
|---|---|---|---|---|
| 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 8: A sentence being segmented into words

word list contains she သူမ. This problem can be solved by standardization. Myanmar Language Commission (MLC, 1993) has advised that the words “she” and “he” should be written only as သူ and the word သူမ representing a feminine pronoun should not be used. For example 2 in Table 9, the text အားပေးသည် can be segmented into two ways. 1) အားပေးသည် [a: pei: thi] which means “encourage” and 2) အား [particle for indicating dative case] and ပေးသည် give [pei: thi]. Because of greedy search from left to right, our algorithm will always segment as အားပေးသည် no matter what the context is.

In order to solve these problems, we are plan to use machine learning techniques which 1) can also detect real words dynamically (Park, 2002) while we are segmenting the words and 2) correct the greedy cut from left to right using frequencies of the words from the training samples.

Although our work presented here is for Myanmar, we believe that the basic ideas can be applied to any script which is primarily syllabic in nature.

## 7 Conclusions

Since words are not uniformly delimited by spaces in Myanmar script, segmenting sentences into words is an important task for Myanmar NLP. In this paper we have described the need and possible techniques for segmentation in Myanmar script. In particular, we have used a combination of stored lists, suffix removal, morphological analysis and syllable level n-grams to hypothesize valid words with about 99% accuracy. Necessary scripts have been written in Perl. Over the last few years, we have collected monolingual text corpora totalling to about 2,141,496 sentences and English-Myanmar parallel corpora amounting to about 80,000 sentences and sentence fragments, aligned at sentence and word levels. We have also built a list of 1216 stop words, 4550 syllables and 800,000 words from a variety of sources including our own corpora. We have used fairly simple and intuitive methods not requiring deep linguistic insights or sophisticated statistical inference. With this initial work, we now plan to apply a variety of machine learning techniques. We hope this work will help to accelerate work in Myanmar language and larger lexical resources will be developed soon.

## References


Richard C. Angell, George W. Freurd, and Peter Willett. 1983. Automatic spelling correction using a trigram similarity measure. *Information Processing & Management*, 19(4):255–261.

Pirkola Ari, Heikki Keskustalo, Erkka Leppnen, Antti-Pekka Knsl, and Kalervo Jrvelin. 2001. Targeted s-gram matching: a novel n-gram matching technique for cross- and monolingual word form variants. *Information Research*, 7(2):235–237, january.

U damma sami. 2004. *Myanmar-French Dictionary*.

English-myanmar dictionary. Ministry of Education, Union of Myanmar,CD version.

Paulette Hopple. 2003. *The structure of nominalization in Burmese,Ph.D thesis*. May.

Hla Hla Htay, G. Bharadwaja Kumar, and Kavi Narayana Murthy. 2006. Building english-myanmar parallel corpora. In *Fourth International Conference on Computer Applications*, pages 231–238, Yangon, Myanmar, Feb.

| Example 1: ဓားပြမှုတွင်သူမပါဝင်ခဲ့ဘူး | | | | | |
|---|---|---|---|---|---|
| ဓားပြမှု | | တွင် | သူမ | ပါဝင်ခဲ့ဘူး | |
| [damja. hmu.] | | [twin] | [thu ma.] | [pa wun khe. bu;] | |
| robbery | | in | she | did not involve | |
| N | | Particle | $N_{subj}$ | $V_{pastneg}$ | |
| Example 2: မိမိမလိုချင်သောတာဝန်ကိုသူတစ်ပါးအားပေးသည်။ | | | | | |
| မိမိ | မလိုချင်သော | တာဝန် | ကို | သူတစ်ပါး | အားပေးသည် |
| [mi. mi.] | [ma. lou chin tho:] | [ta wun] | [gou] | [thu daba:] | [a: pei: thi] |
| I,myself | don't want | duty,responsibility | | others | encourage |
| $N_{subj}$ | $V_{neg}$ | $N_{obj1}$ | Particle | $N_{obj2}$ | V |

Table 9: Analysis of Over-Segmentation


Zaw Htut. 2001. All possible myanmar syllables, September.

Adoniram Judson. 1842. *Grammatical Notices of the Buremse Langauge*. Maulmain: American Baptist Mission Press.

Vlado Keselj. 2006. Text ::ngrams. http://search.cpan.org/ vlado/ Text-Ngrams-1.8/, November.

MLC. 1993. *Myanmar Words Commonly Misspelled and Misused*. Department of the Myanmar Language Commission,Ministry of Education, Union of Myanmar.

MLC. 2002. *Myanmar-English Dictionary*. Department of the Myanmar Language Commission, Ministry of Education, Union of Myanmar.

MLC. 2003. *Myanmar Orthography*. Department of the Myanmar Language Commission,Ministry of Education, Union of Myanmar, June.

MLC. 2005. *Myanmar Grammer*. Department of the Myanmar Language Commission, Ministry of Education,Union of Myanmar, June.

Kavi Narayana Murthy. 2006. *Natural Language Processing - an Information Access Perspective*. Ess Ess Publications, New Delhi, India.

Youngja Park. 2002. Identification of probable real words : an entropy-based approach. In *ACL-02 Workshop on Unsupervised Lexical Acquisition*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

U Soe Saya. 2000. *The Khit Thit English-English-Myanmar Dictionary with Pronunciation*. Yangon, Myanmar, Apr.

2000. Student's english-english/myanmar dictionary. Ministry of Commerce and Myanmar Inforithm Ltd, Union of Myanmar, CD version, Version 1, April.

Ye Kyaw Thu and Yoshiyori Urano. 2006. Text entry for myanmar language sms: Proposal of 3 possible input methods, simulation and analysis. In *Fourth International Conference on Computer Applications*, Yangon, Myanmar, Feb.

U Tun Tint. 2004. Features of myanmar language. May.

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

Ni Ko Ye. 1997. *Orchestra*. The two cats, June.