Selection of XML tag set for Myanmar National Corpus

Wunna Ko Ko
AWZAR Co.
Mayangone Township, Yangon,
Myanmar
wunnakoko@gmail.com

Thin Zar Phyo
Myanmar Unicode and NLP Research
Center
Myanmar Info-Tech, Hlaing Campus,
Yangon, Myanmar
myanmar.nlp5@gmail.com

Abstract

In this paper, the authors mainly describe about the selections of XML tag set for Myanmar National Corpus (MNC). MNC will be a sentence level annotated corpus. The validity of XML tag set has been tested by manually tagging the sample data.

Keywords: Corpus, XML, Myanmar, Myanmar Languages

1 Introduction

Myanmar (formerly known as Burma) is one of the South-East Asian countries. There are 135 ethnic groups living in Myanmar. These ethnic groups speak more than one language and use different scripts to present their respective languages. There are a total of 109 languages spoken by the people living in Myanmar [Ethnologue, 2005].

There are seven major languages, according to the speaking population in Myanmar. They are Kachin, Kayin/Karen, Chin, Mon, Burmese, Rakhine and Shan [Ko Ko & Mikami, 2005]. Among them, Burmese is the official language and spoken by about 69% of the population as their mother tongue [Ministry of Immigration and Population, 1995].

Corpus is a large and structured set of texts. They are used to do statistical analysis, checking occurrences or validating linguistic rules on a specific universe.[1]

In Myanmar, there are a plenty of text for most of the languages, especially Burmese and major languages, since stone inscription.

Myanmar Language Commission and a number of scholars had been collected a number of corpora for their specific uses [Htay et al., 2006]. But there is no national corpus collection, both in digital and non-digital format, until now.

Since there are a number of languages used in Myanmar, the national level corpus to be built will include all languages and scripts used in Myanmar. It has been named as Myanmar National Corpus or MNC, in short form.

During the discussion for the selection of format for the corpus, XML (eXtensible Markup Language), a subset of SGML (Standard Generalized Markup Language), format has been chosen since XML format can be a long usable and possible to keep the original format of the text [Burnard. 1996]. The range of software available for XML is increasing day by day. Certainly more and more NLP related tools and resources are produced in it. This in turn makes the necessity of selection of XML tag set to start building of MNC.

MNC will include not only written text but also spoken texts. The part of written text will include regional and national newspapers and periodicals, journals and interests, academic books, fictions, memoranda, essays, etc. The part of spoken text will include scripted formal and informal conversations, movies, etc.

During the selection of XML tag sets, the sample for all the data which will be included in building of MNC, has been learnt.

2 Myanmar National Corpus

Myanmar is a country of using 109 different languages and a number of different scripts [Ethnologue, 2005]. In order to do language processing for these languages and scripts, it becomes a necessity to build a corpus with

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

languages and scripts used in Myanmar; at least with major languages and scripts, which will include almost all areas of documents.

Among the different scripts used in Myanmar, the popular scripts include Burmese script (a Brahmi based script), Latin scripts. Building of MNC will be helpful for development of Natural Language Processing (NLP) tools (such as grammar rules, spelling checking, etc) and also for linguistic research on these languages and scripts. Moreover, since Burmese script is written without necessarily pausing between words with spaces, the corpus to be built is hoped to be useful for developing tools for automatic word segmentation.

2.1 XML based corpus

XML is universal format for structured documents and data, and can provide highly standardized representation frameworks for NLP (Jin-Dong KIM et al. 2001); especially, the ones with annotated corpus based approaches, by providing them with the knowledge representation frameworks for morphological, syntactic, semantics and/or pragmatics information structure. Important features are:

- XML is extensible and it does not consist of a fixed set of tags.
- XML documents must be well-formed according to a defined syntax.
- XML document can be formally validated against a schema of some kind.
- XML is more interested in the meaning of data than its presentation.

The XML documents must have exactly one top-level element or root element. All other elements must be nested within it. Elements must be properly nested [Young, 2001]. That is, if an element starts within another element, it must also end within that same element.

Each element must have both a start-tag and an end-tag. The element type name in a start-tag must exactly match the name in the corresponding end-tag and element name are case sensitive.

Moreover, the advantages of XML for NLP includes ontology extraction into XML based structured languages using XML Schema. The great benefit about XML is that the document itself describes the structure of data. [2]

Three characteristics of XML distinguish from other markup languages:[3]

- its emphasis on descriptive rather than procedural markup;
- its notion of documents as instances of a *document type* and
- its independence of any hardware or software system.

Since MNC is to be built in XML based format, the selection process for tag set of XML become an important process. The XML tagged corpus data should also keep the original format of the data.

In order to select XML tag set for MNC, the sample data for the corpus has to be collected. The format of the sample corpus data has been studied for the selection of the XML tag set in appropriate with the data format.

2.2 Structure of a data file at MNC

The structure of a data file at MNC will include two main parts: information of the corpus file and the corpus data.

The first part, the header part of a corpus file, describes the information of a corpus file. The information of the corpus file includes the header which will provide sensible use of the corpus information in machine readable form. In this part, the information such as language usage and the description of the corpus file will be included.

The second part, the document part, of a corpus file will include the source description of the corpus data and the corpus data, the written or spoken part of the text, itself. The information of the corpus data such as bibliographic information, authorship, and publisher information will be included in this section. Moreover, the corpus data itself will also be included in this section.

The hierarchically structure of a corpus file at MNC will be as shown in figure 1.

[2] http://www.tei-c.org/P5/Guidelines/index.html
[3] http://www.w3.org/TR/xml/

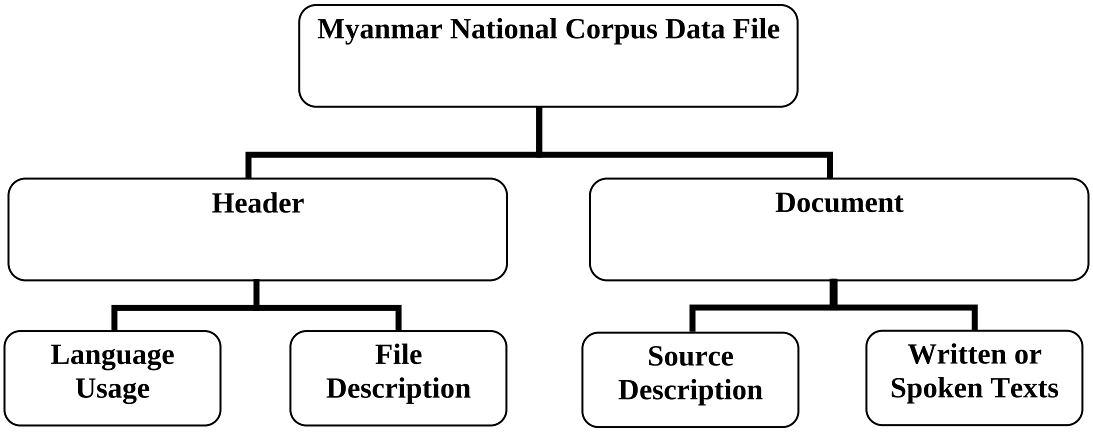

Figure 1. Hierarchically structure of a data file at MNC

3 Selection of necessary XML tag set

After studying original formats and features of texts, to be used in corpus, and the structure of the corpus file has been determined, the selection procedure for XML tag set has been started.

British National Corpus (BNC)[4], American National Corpus (ANC)[5] had been referenced for selection of XML tag set.

The selection of XML tag set is based on the nature of the structure of a data file. The main tag for the data file will be named as <mnc> which is the abbreviation of Myanmar National Corpus.

A data file contains two main parts, the header part and the document part.

```
-<mnc>
  +<teiHeader></teiHeader>
  +<myaDoc></myaDoc>
</mnc>
```

Figure 2. Root and element tags of MNC

3.1 Header Part

The XML tag for the header part of the corpus data file is named as <teiHeader>. Text Encoding Initiative (TEI) published guidelines for the text encoding and Interchange[6]. TEI encoding scheme consists of a number of rules with which the document has to adhere in order to be accepted as a TEI document.

This header part contains language usage of the data file <langUsage> and the file description <fileDesc> which includes machine readable information of the data file.

```
-<mnc>
  -<teiHeader>
      +<langUsage></langUsage>
      +<fileDesc></fileDesc>
    </teiHeader>
 +<myaDoc></myaDoc>
</mnc>
```

Figure 3. Element and Child tags of MNC

The language usage part contains such information as language name <langName>, script information <script>, International Organization for Standardization (ISO) code number <ISO>, encoding information <encodingDesc> and version of encoding <version>.

```
-<mnc>
  -<teiHeader>
      <langUsage>
            <langName> </langName>
            <script> </script>
            <ISO></ISO>
           <encodingDesc> </encodingDesc>
```

[4] Lou Burnard. 2000. Reference Guide for the British National Corpus (World Edition). Oxford University Computing Services, Oxford.

[5] Nancy Ide and Keith Suderman. 2003. The American National Corpus, first Release. Vassar College, Poughkeepsie, USA

[6] TEI Consortium. 2001, 2002 and 2004 Text Encoding Initiative. In The XML Version of the TEI Guidelines.

```
          <version> </version>
      </langUsage>
    +<fileDesc></fileDesc>
  </teiHeader>
+<myaDoc></myaDoc>
</mnc>
```

Figure 4. 2nd level Child tags in language Usage part of MNC

The file description part contains such information as title information of the corpus file <titleStmt>, edition information <editionStmt> and publication information about the corpus file <publicationStmt>. The detail information will be tagged using more specific lower level child tags under the previously described tags.

```
-<mnc>
  -<teiHeader>
     +<langUsage></langUsage>
     -<fileDesc>
          +<titleStmt></titleStmt>
          +<editionStmt></editionStmt>
          +<publicationStmt></publicationStmt>
       </fileDesc>
    </teiHeader>
 +<myaDoc></myaDoc>
</mnc>
```

Figure 5. 2nd level Child tags in file description part of MNC

3.2 Document Part

The XML tag for the document part of the corpus data file is named as <myaDoc> which is the short form of Myanmar Document. It contains two sub parts: the source description of the data <sourceDesc> and the original data itself which in turn can be divided into two types; written text <wtext> and the spoken text <stext>.

```
<mnc>
   +<teiHeader></teiHeader>
   -<myaDoc>
      +<sourceDesc></sourceDesc>
      +<wtext></wtext>
    </myaDoc>
</mnc>
```

Figure 6. Element and Child tags of MNC

The first part, the source description part of the data <sourceDesc>, will contain the bibliographic information, such as title, name of author, publisher, etc., of the original data.

```
<mnc>
 +<teiHeader></teiHeader>
 -<myaDoc>
    -<sourceDesc>
       -<bibl>
               <title></title>
               <author></author>
               <editor/></editor>
            -<imprint>
                  <publisher></publisher>
                  <pubPlace></pubPlace>
                  <date></date>
             </imprint>
         </bibl>
    </sourceDesc>
  +<wtext></wtext>
  </myaDoc>
</mnc>
```

Figure 7. 2nd level Child tags for source description part of MNC

The second part, the original data part <wtext> or <stext> will contain the whole original data. The original format information such as heading <head type="MAIN">, sub-heading <head type="SUB">, paragraph number <paragraph n="1">, sentence number <s n="1"> will be saved in this part.

```
<mnc>
   +<teiHeader></teiHeader>
   -<myaDoc>
      +<sourceDesc></sourceDesc>
       -<wtext>
          -<head>
                <s></s>
             +<paragraph></paragraph>
             +<head></head>
           </head>
        </wtext>
      </myaDoc>
</mnc>
```

Figure 8. 2nd level Child tags for original data part of MNC

Since MNC is going to be annotated in sentence level, each sentence will be annotated and numbered.

```
<mnc>
  +<teiHeader></teiHeader>
   -<myaDoc>
      +<sourceDesc></sourceDesc>
       -<wtext>
          -<head>
                     <s></s>
                  -<paragraph>
                         -<s></s>
                     </paragraph>
             </head>
            +<head></head>
         </wtext>
     </myaDoc>
</mnc>
```

Figure 9. Down to the sentence level Child tags of MNC

3.3 Sample MNC data file

The Myanmar National Corpus is a major resource for linguistic research, as well as computational linguistics research, lexicography, corpus linguistic research and a resource for the development of Myanmar Language teaching material because we expect the corpus to be continually expanded in the future.

A sample MNC data is use the Universal Declaration of Human Rights (UDHR) texts in Burmese and Karen, which is one of the major languages in Myanmar, has been used to sample tagging with the selected XML tag set.

The following figure is show for the sample MNC.

```
<? xml version="1.0"?>
<mnc>
  -<teiHeader>
          -<langUsage>
                    <langName> Myanmar </langName>
                    <script>Burmese</script>
                    <ISO> 10646</ISO>
                    <encodingDesc> utf-8</encodingDesc>
                    <version>Unicode 5.0</version>
           </langUsage>
         -<fileDesc>
                  -<titleStmt>
                         <title>Myanmar National Corpus</title>
                        -<respStmt>
                              <resp>Corpus built by</resp>
                              <name>Myanmar NLP Team</name>
                         </respStmt>
                    </titleStmt>
                  -<editionStmt>
                        <edition> First TEI-conformant version </edition>
                        <extent/>
                   </editionStmt>
                 -<publicationStmt>
                        <address>Myanmar Info-Tech, Yangon, Myanmar</address>
                        <availability status="restricted">
                              Availability limited to Myanmar NLP Team
                        </availability>
                       -<creation>
                              <date>07/06/2007</date>
                        </creation>
                        <distributor>Myanmar NLP Team </distributor>
                        <idno type="mnc">MNC101</idno>
                   </publicationStmt>
           </fileDesc>
 </teiHeader>

-<myaDoc xml:id="TEXTS">
    -<sourceDesc>
                  -<bibl>
                       <title>
                              အပြည်ပြည်ဆိုင်ရာလူ့အခွင့်အရေးကြေညာစာတမ်း
                              (meaning: Universal Declaration of Human Rights)
                       </title>
                       <author/>
                       <editor/>
                     -<imprint vol="64" n="46">
                              <publisher></publisher>
                              <pubPlace></pubPlace>
                              <date></date>
                       </imprint>
                  </bibl>
    </sourceDesc>

   -<wtext type="OTHERPUB">
              -<head type="MAIN">
                       <s n="1">
                            အပြည်ပြည်ဆိုင်ရာလူ့အခွင့်အရေးကြေညာစာတမ်း
                            (meaning: Universal Declaration of Human Rights)
                       </s>
                    +<paragraph n="1"></paragraph>
                    -<head type="SUB">
                            <s n="1"> စကားချီး (meaning: Preamble) </s>
                         -<paragraph n="1">
                              -<s n="1">
                           လူခပ်သိမ်း၏ မျိုးရိုးဂုဏ်သိက္ခာနှင့်တကွလူတိုင်းအညီအမျှခံစားခွင့်ရှိသည့်အခွင့်အရေးများကို
                           အသိအမှတ်ပြုခြင်းသည်လူခပ်သိမ်း၏ လွတ်လပ်မှု၊တရားမျှတမှု၊ငြိမ်းချမ်းမှုတို့၏ အခြေခံအုတ်မြစ်
                           ဖြစ်သောကြောင့်လည်းကောင်း၊ …….
                            (meaning: Whereas recognition of the inherent dignity and of the equal and
                           inalienable rights of all members of the human family is the foundation of
                           freedom, justice and peace in the world,…….)
                               </s>
                            </paragraph>
                        +<paragraph n="2"></paragraph>
                       </head>
                    -<head type="SUB">
                            <s n="2"> အပိုဒ် ၁ (meaning: paragraph 1) </s>
                         -<paragraph n="1">
                              <s n="1">
                           လူတိုင်းသည်တူညီလွတ်လပ်သောဂုဏ်သိက္ခာဖြင့်လည်းကောင်း၊တူညီလွတ်လပ်သောအခွင့်အရေး
                           များဖြင့်လည်းကောင်း၊မွေးဖွားလာသူများဖြစ်သည်။
                              (meaning: All human beings are born free and equal in dignity and rights.)
```

```
                              </s>
                              <s n=”2”>
                              ထိုသူတို့၌ပိုင်းခြားဝေဖန်တတ်သောဉာဏ်နှင့်ကျင့်ဝတ်သိတတ်သောစိတ်တို့ရှိကြ၍ထိုသူတို့သည်
                              အချင်းချင်းမေတ္တာထား၍ဆက်ဆံကျင့်သုံးသင့်၏။
                              (meaning: They are endowed with reason and conscience and should act towards
                              one another in a spirit of brotherhood.)
                             </s>
                         </paragraph>
                    -<head type="SUB">
                         <s n="3"> အပိုဒ် ၂ (meaning: paragraph 2) </s>
                       +<paragraph n="1"></paragraph>
                       +<paragraph n="2"></paragraph>
                     </head>
                    -<head type="SUB">
                         <s n="4"> အပိုဒ် ၃ (meaning: paragraph 2) </s>
                      -<paragraph n="1">
                            <s n="1">
                              လူတိုင်း၌အသက်ရှင်ရန်လွတ်လပ်မှုခွင့်နှင့်လုံခြုံစိတ်ချခွင့်ရှိသည်။
                              (meaning: Everyone has the right to life, liberty and security of person.)
                            </s>
                        </paragraph>
                     </head>
                   +<head type=”SUB”></head>
                   +<head type=”SUB”></head>
                 </head>
       </wtext>
  </myaDoc>
</mnc>
```

Figure 10. Sample MNC Corpus file (Burmese UDHR text in MNC XML format)

4 Conclusion and Future work

In this paper, the authors have clearly described about the selection of XML tag set for building of MNC. Since the word level segmentation for Burmese script is not yet available, the corpus data will be annotated only up to the sentence level in order to be in the same format for all Myanmar languages and scripts.

In order to check whether the selected the XML tag set will be enough and useful for tagging the corpus data, the sample corpus data has been collected by manually tagging the data which includes newspapers and periodicals, Universal Declaration of Human Rights (UDHR), novels and essays.

Since the manual tagging to the sample corpus data proves that the selected XML tag set is enough to cover a variety of data sources, the next step is to develop an algorithm for automatic tagging the data.

Acknowledgement

This study was performed with the support of the Government of the Union of Myanmar through Myanmar Natural Language Implementation Committee. Thanks and gratitude towards the members of Myanmar Language Commission for providing necessary information to write this paper.

References

Ethnologue. 2005 *Languages of the World*, 15th Edition, Dallas, Tex.: SIL International. Online version: http://www.ethnologue.com/. Edited by Raymond G. Gordon, Jr.

Hla Hla Htay, G. Bharadwaja Kumar and Kavi N. Murthy. 2006. *Constructing English-Myanmar Parallel Corpora*. The Fourth International Conference on Computer Application 2006 (ICCA 2006) Conference Program.

Jin-Dong KIM, Tomoko OHTA, Yuka TATEISI, Hideki MIMA and Jun'ichi TSUJII. 2001. *XML-based Linguistic Annotation of Corpus* . In the Proceedings of the first NLP and XML Workshop held at NLPRS 2001. pp. 47--53.

Lou Burnard. 1996. *Using SGML for Linguistic Analysis: the case of the BNC*. ACM Vol 1 Issue 2 (Spring 1999) MIT Press ISSN: 1099-6621. pp. 31-51.

Michaek J. Young. 2001. *Step by Step XML.*Prentice Hall of India Private Limited Press. ISBN-81-203-1804-B

Ministry of Immigration and Population. 1995. *Myanmar Population Changes and Fertility Survey 1991*. Immigration and Population Department

Wunna Ko Ko, Yoshiki Mikami. *2005 Languages of Myanmar in Cyberspace*, In Proceedings of TALN & RECITAL 2005 (NLP for Under-Resourced Languages Workshop), Dourdan, FRANCE, 2005 June, pp. 269-278.