Corpus Building for Mongolian Language

Purev Jaimai
Center for Research on Language Processing,
National University of Mongolia, Mongolia
purev@num.edu.mn

Odbayar Chimeddorj
Center for Research on Language Processing,
National University of Mongolia, Mongolia
odbayar@num.edu.mn

Abstract

This paper presents an ongoing research aimed to build the first corpus, 5 million words, for Mongolian language by focusing on annotating and tagging corpus texts according to TEI XML (McQueen, 2004) format. Also, a tool, MCBuilder, which provides support for flexibly and manually annotating and manipulating the corpus texts with XML structure, is presented.

1 Introduction

Mongolian researchers quite recently have begun to be involved in the research area of Natural Language Processing. All necessary linguistic resources, which are required for Mongolian language processing, have to be built from scratch, and then they should be shared in public research for the rapid development of Mongolian language processing.

This ongoing research aims to build a tagged and parsed 5 million words corpus for Mongolian by developing a spell-checker, tagger, sentence-parser and others (see Figure 1 and 2). Also, we needed to develop a tagset for the corpus because there was not any tagset for Mongolian and the traditional words categories are not appropriate to it. Thus, we designed a high level tagset, which consists of 20 tags, and are further classifying them. Currently, we have collected and populated 500 thousand words, 50 thousand of which have been manually tagged, into the corpus (see Figure 1).

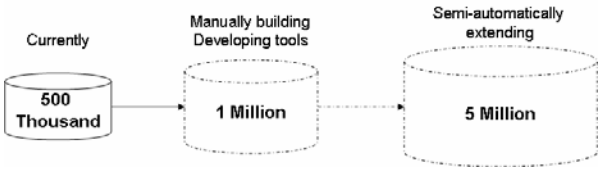

Figure 1. Current and future states of building a Mongolian corpus.

And, we manually build the corpus until collecting and annotating 1 million words and tagging 100 thousand words of them for semi-automatically building the corpus in the future.

2 Corpus Building Design

We are building the corpus as sub-corpora, which are a raw corpus, a cleaned corpus, a tagged corpus and a parsed corpus, separately for various kinds of studying and use on Mongolian language (Figure 2).

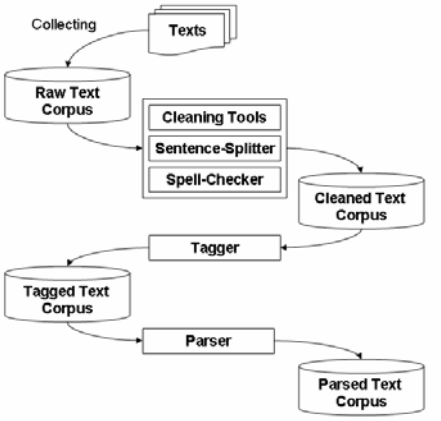

Figure 2. Schema of building a Mongolian corpus.

At first, we are collecting the editorial articles of Unen newspaper (Unen publish), which is one of

the best written newspapers in Mongolia, by using OCR application. We will also collect laws, school book, and literary text (see Figure 3).

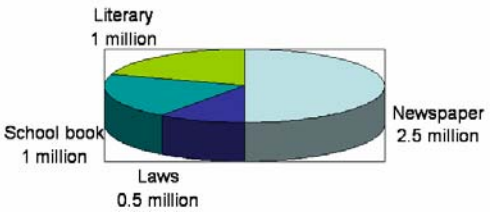

Figure 3. Text sizes included in the corpus.

The corpus annotation follows TEI XML standard. According to the work scope, the annotating part is divided into two parts that are structural annotation such as paragraphs, sentences, and so on, and POS tagging.

The structure of the text annotation is presented in Figure 4.

```
<tei>
  <teiHeader>
    <fileDesc/ >
  </teiHeader>
  <text>
    <body>
    <s>
    <word id="" pos="tag">WORD</word>
    </s>
    </body>
  </text>
</tei>
```

Figure 4. XML Structure of corpus text.

For annotating two parts, once a manual corpus builder, called MCBuilder, were planned to develop, we have developed the first version and used to annotating 500 thousand word texts and tagging 50 of them (see Figure 5).

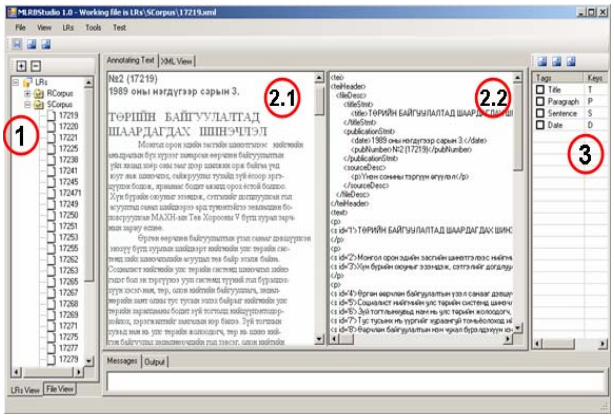

Figure 5. Screenshot of the corpus organizer and its main view.

MCBuilder has three main windows that are (1) manipulating and organizing the corpus, (2) annotating sample texts and (3) manipulating tagset as shown in Figure 5.

3 Conclusion

Mongolian language has hardly studied by computer, and its traditional rules such as inflectional, derivational, part of speech, sentence constituents, etc are extremely difficult to computerize. Our research works in the last few years showed it (Purev, 2006). Therefore, we are revising them by creating a corpus for computer processing.

The proposals of this ongoing research are the first Mongolian 5 million words corpus, and tools that are spell-checker, tagger and parser.

Currently, we have done followings:

- Defined the corpus design, XML structure of the corpus text, and the high level tagset
- Collected and annotated 500 thousand words text
- Tagged 50 thousand words
- Released the first version of a Mongolian corpus building tool called MCBuilder
- First versions of Syllable-parser and Morph-analyzer for Mongolian

We are planning to complete the corpus in the next two years.

4 Acknowledgement

Here described work was carried out by support of PAN Localization Project (PANL10n).

References

PANL10n: PANLocalization Project. National University of Computer and Emerging Sciences, Pakistan.

Purev J. 2006. *Corpus for Mongolian Language*, Research Project, Mongolia.

Purev J. and Odbayar Ch.. 2006. *Towards Constructing the Corpus of Mongolian Language*, Proceeding of ICEIC.

Sperberg-McQueen, C. M. and Burnard, L.. 2004. *Text Encoding Initiative. The XML version of the TEI Guidelines*, Website.

Unen press. 1984-1989. *Editorial Articles*. Mongolia