



# **PAN LOCALIZATION PROJECT**

## **POS TAGGED CORPUS FOR MONGOLIAN**

Phase 3.1

October 10, 2009

**CENTER FOR RESEARCH ON LANGUAGE PROCESSING  
NATIONAL UNIVERSITY OF MONGOLIA, ULAANBAATAR, MONGOLIA**

# Table of Contents

Table of Contents .....	2
List of Figures and Tables .....	3
Abstract.....	4
1. Introduction .....	5
2. Analysis on POS Tagged Corpus .....	6
Word frequency of corpus.....	6
Frequency of lexicon.....	8
3. Evaluation .....	9
4. Conclusion .....	10
Reference .....	11

## List of Figures and Tables

Figure 1. Proportion of numbers to parts of speech .....	6
Figure 2. Numbers of noun parts of speech .....	6
Figure 3. Frequency of verbs.....	7
Figure 4. General frequency of lexicon .....	8
Table 1. Frequency of nouns inflected case suffixes .....	6
Table 2. Frequencies of some verb forms .....	7
Table 3. Frequencies of adwords .....	7
Table 4. Frequencies of other parts of speech.....	8

## **Abstract**

This report provides an overview of POS tagged corpus for Mongolian. The corpus consisting of 5 million words is tagged by a bi-gram POS tagger and corrected by hand.

# 1. Introduction

In this project, we have to build a 5 million word tagged corpus from scratch and develop related tools such as text cleaning tools, spell-checker, POS tagger and POS tagset for Mongolian. Implementing such project is new experience for us because this corpus would be the first POS tagged corpus to build locally. Summarizing the previous phases of this project, following tasks have been done:

- Designed corpus for Mongolian
- Collected 5 million words from the Internet and printed sources
- Developed text cleaning tools and dictionary-based spell-checker
- Cleaned the collected texts
- Developed POS tagset
- Developed manual and bi-gram POS taggers
- Tagged 1 million words

(See the research reports [1, 2, 3, 4, 5] for more details)

As the last phase of the project, we have finished tagging the rest of the corpus, 4 million words, based on 1 million words that had been tagged in the previous phase. Moreover, the bi-gram POS tagger is trained on the corpus, and its accuracy is 92% for tagging around other 600k words as a test.

## 2. Analysis on POS Tagged Corpus

### Word frequency of corpus

In the last phase of the project, 5 million words, the whole corpus, are tagged by the bi-gram POS tagger (see [5] for more details) and checked and corrected by hand. In this section, the analysis on the tagged corpus is presented.

In the corpus, there are around 6 million tokens including punctuations such as periods, commas, question marks and so on. The most occurred parts of speech are nouns, counted more than 2 million. Figure 1 shows the proportion of numbers to parts of speech.

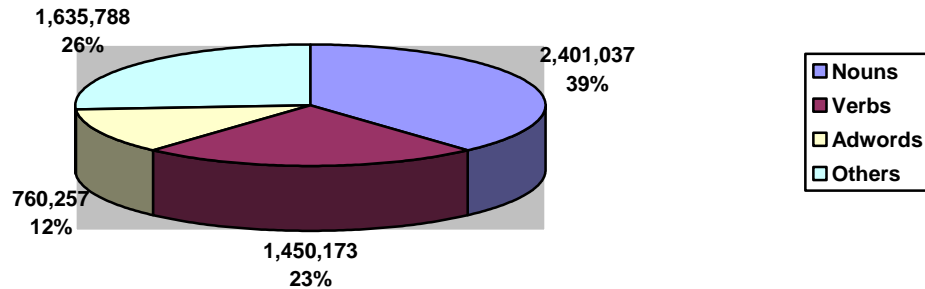


Figure 1. Proportion of numbers to parts of speech

In Figure 1, the numbers of the parts of speech in the tagged corpus are shown by dividing the parts of speech into four general groups that are nouns, verbs, adwords and others. Nouns are 39 percent, verbs are 23 percent, adwords are 12 percent, and other parts of speech such as punctuations, numbers, postpositions, particles and so on 26 percents of the 5 million words corpus, respectively.

The noun group consists of common nouns, pronouns, abbreviations, proper nouns and foreign words. Common nouns are 86 percent of the total nouns (see Figure 2).

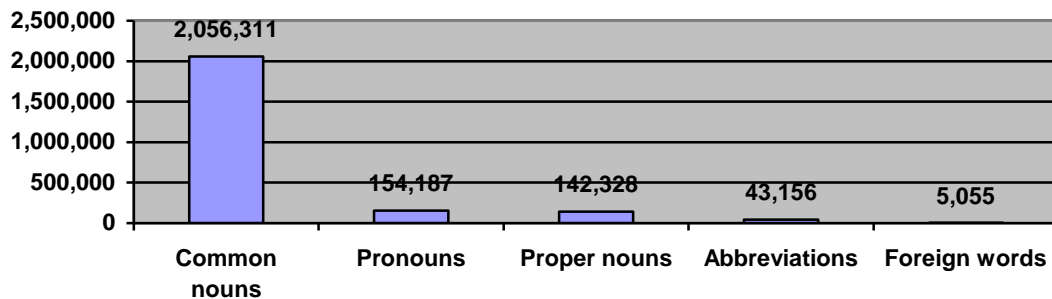


Figure 2. Numbers of noun parts of speech

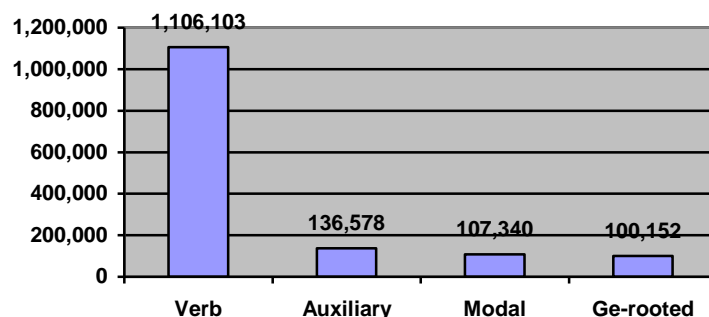
Pronouns and proper nouns are each 6 percent of the nouns, while abbreviations and foreign words are around 2 percent. In our POS tagset, common nouns are tagged with various tags according to their inflectional forms, while pronouns, abbreviations, proper nouns and foreign words are tagged with one certain tag. For example, pronouns are tagged only PN, while a common noun inflected nominative case is tagged by tag N, or a common noun inflected genitive case is tagged by tag NG. In addition, it can be concluded that the most effectively used case is genitive case, and followed by accusative case as shown in Table 1.

Table 1. Frequency of nouns inflected case suffixes

No.	Noun cases	Frequency
1.	Genitive	495,682
2.	Accusative	171,591
3.	Locative (dative)	146,475
4.	Instrumental	75,159
5.	Ablative	41,896
6.	Commutative	29,692
7.	Direction	25

In Table 1, frequencies of nouns inflected case suffixes are shown. The least used case in Mongolian is direction case in case of attaching to root or base words. On the other hand, the direction case is usually separated from content words.

The verb group shown in Figure 1 consists of verbs, auxiliary verbs, modals and ge-rooted verbs. The frequencies of the verbs are shown in Figure 3.



**Figure 3. Frequency of verbs**

As shown in Figure 3, verbs are 76 percent, auxiliary verbs are 10 percent, modals are 7 percent and ge-rooted (mentioned in the report of the previous phase [5]) verbs are also 7% of the total verbs of the corpus, respectively. Furthermore, the most effectively used inflectional suffix of verbs is a serial verb suffix. The second one is the infinitive form of verbs (see Table 2).

**Table 2. Frequencies of some verb forms**

No.	Verb form	Frequency
1.	Serial verb	417,488
2.	Infinitive verb	209,339
3.	Past tense	205,106
4.	Present tense	73,016
5.	Future tense	43,781

Mostly occurred verb forms are shown in Table 2. In Mongolian, compound verbs are often used. Thus, the serial verb forms are occurred more than tenses.

The adword group shown in Figure 1 consists of adjectives, adverbs, sentence-adverb, determiners, comparatives and superlatives. The frequencies of these parts of speech are shown below (see Table 3).

**Table 3. Frequencies of adwords**

No.	Adwords	Frequency
1.	Adjective	518,904

2.	Adverb	184,729
3.	Sentence-adverb	28,211
4.	Determiner	15,638
5.	Comparative	8,260
6.	Superlative	4,507

The most effectively used adword is the adjective, while the least used is superlative according to Table 3.

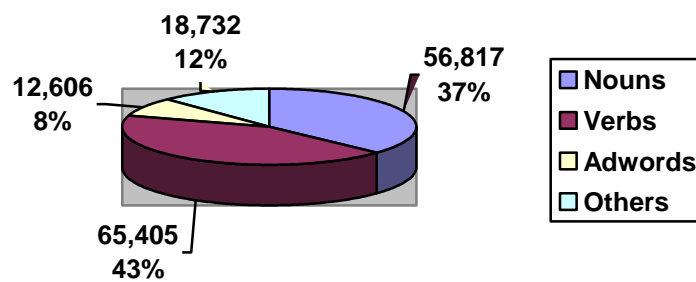
The other group mentioned in Figure 1 consists of other parts of speech (see Table 4) except above three groups.

**Table 4. Frequencies of other parts of speech**

No.	Other POSs	Frequency
1.	Punctuation	910,258
2.	Number	220,628
3.	Postposition	220,441
4.	Possessive particle	119,061
5.	Conjunction	59,523
6.	Question particle	23,877
7.	Interjection	21,056
8.	Negative	18,049
9.	Pre-position	8,721
10.	Content-separated inflectional suffix	34,056

### **Frequency of lexicon**

After tagging the 5 million words, there are totally 153,560 headwords in the lexicon used for the POS tagger. Although the nouns are the most occurred words in the corpus as shown in Figure 1, the most occurred parts of speech are the verbs in the lexicon (see Figure 4).



**Figure 4. General frequency of lexicon**

In Figure 4, the frequency of the lexicon is shown by dividing into four main groups that are nouns, verbs, adwords and others. Due to our tagged corpus, verbs are the parts of speech that have more morphological forms than others in Mongolian. There are 65,405 verb forms, 56,817 noun forms, 12,606 adword forms and 18,732 other parts of speech forms in the lexicon, respectively.

### 3. Evaluation

Using the training data trained on the corpus mentioned in section 2, a test corpus consisting of around 600k words is tagged by the POS tagger. The test corpus is collected from daily newspapers, and it includes 460 articles covering various general topics. As a result, the accuracy of the tagging is 92 percent. Incorrectly tagged words are caused from mistyped words and new words or word forms to the lexicon.

## 4. Conclusion

Within this project, we have built a POS tagged corpus for Mongolian. The corpus consists of 5 million words, and is collected from the Internet and printed sources. The text domains cover laws, literature and newspaper. More details about text collection and domains are mentioned in the research report [1].

One of the main tasks is to clean the raw corpus. Therefore, cleaning tools are developed for correcting file and character encodings, hyphenated and misspelled words, etc by the project team. More details about the error analysis on the raw corpus and cleaning tools are mentioned in the research report [3, 5].

The corpus tagging is divided into two tagging phases: one is to tag the first 1 million words and another is to tag the other 4 million words. In the first tagging, the POS tagset for Mongolian is designed several times, and the tagging tools that are manual and automatic POS taggers are also developed. More details about tagging the first 1 million words tagging are mentioned in the research report [6].

In this phase, 4 million words are tagged as the second part of the corpus tagging. The tagging procedure is the same to the first tagging part. After tagging and correcting the entire corpus, the POS tagger is trained on this corpus, and tags other 500k words corpus with 92 percent of the accuracy.

## Reference

- [1] **Research report on corpus collection design**  
*Center for Research on Language Processing, 2007, National University of Mongolia*
- [2] **Research report on manual tagging for Mongolian**  
*Center for Research on Language Processing, 2007, National University of Mongolia*
- [3] **Research report on corpus collection and cleaning tools**  
*Center for Research on Language Processing, 2008, National University of Mongolia*
- [4] **Research report on spell checker for Mongolian**  
*Center for Research on Language Processing, 2008, National University of Mongolia*
- [5] **Research report on error analysis on Mongolian corpus**  
*Center for Research on Language Processing, 2008, National University of Mongolia*
- [6] **Research report on automatic POS tagging for Mongolian**  
*Center for Research on Language Processing, 2009, National University of Mongolia*