# A Study of Linguistic Analysis for Classical Chinese Texts

Tomohiko Morioka, Christian Wittern, Koichi Yasuoka
Center for Informatics in East Asian Studies
Institute for Research in Humanities
Kyoto University
Kyoto 606-8265 JAPAN

Naoki Yamazaki
Faculty of Foreign Language Studies
Kansai University
Suita 564-8680 JAPAN

***Abstract*—A method to analyze classical Chinese texts is proposed. In the method a morphological analyzer MeCab is used. A four-level word-class system for classical Chinese on MeCab is also proposed, and an XEmacs-based editor to make a corpus on the word-class system is presented.**

## I. Introduction

The most difficult point in the linguistic analysis of classical Chinese texts is that they don't have any spaces or punctuation marks between words or between sentences [1]. They consist of continuous strings of Chinese characters from the start to the end of texts. Contrary to the analysis of modern Chinese texts, which have several punctuation marks and can be fragmented into phrases with these punctuation marks, the analysis of classical Chinese texts has to begin with finding out the ends of sentences.

In this paper we propose a method to analyze classical Chinese texts using a morphological analyzer MeCab [2]. Our method, shown in this paper, is concentrated on the proses of classical Chinese texts, but not on the rhymes.

## II. Brief Description of MeCab

MeCab is a language-independent morphological analyzer engine. It was originally developed for Japanese and was subsequently extended to become language-independent. MeCab reads a string of letters or characters from the input stream, segments them into a morpheme sequence, and then outputs the sequence, while adding linguistic information such as word-classes.

We can build any kind of morphological analyzer for any natural language using MeCab, preparing a proper morphological dictionary. We can obtain a better analyzer with a proper morphologically tagged corpus for MeCab. The morphological dictionary and the tagged corpus require a word-class system, which can be stratified at most at four levels.

## III. Development of Our Analyzer for Classical Chinese Texts

### A. Outline of Development

Morphological dictionaries and tagged corpora for MeCab are structured as machine-readable data, and they are difficult for human to read and write. In other words, they are difficult to input from scratch by hand. On the other hand, the format of morphological corpora of MeCab is same as the default output format of MeCab. It means that a prototype analyzer based on MeCab helps developing a more powerful and efficient corpus for MeCab.

Therefore we designed our development of a classical Chinese analyzer into two steps: the first step is building a prototype and the second step is refactoring of this prototype.

At the first step, we developed a prototype dictionary from IPA Japanese Dictionary [3] and defined a prototype word-class system for classical Chinese. We also developed a MeCab-corpus editor to input a prototype corpus easily and to reduce mistakes.

At the second step, we examined the prototype corpus, and redefined our four-level word-class system to be more suitable and systematic for classical Chinese. Then we developed our new dictionary on our new word-class system, and eliminated the old prototype dictionary. We are also developing our new corpus for classical Chinese using our MeCab-corpus editor with our new dictionary.

In the following sections we mainly describe our new dictionary and corpus, developed in the second step.

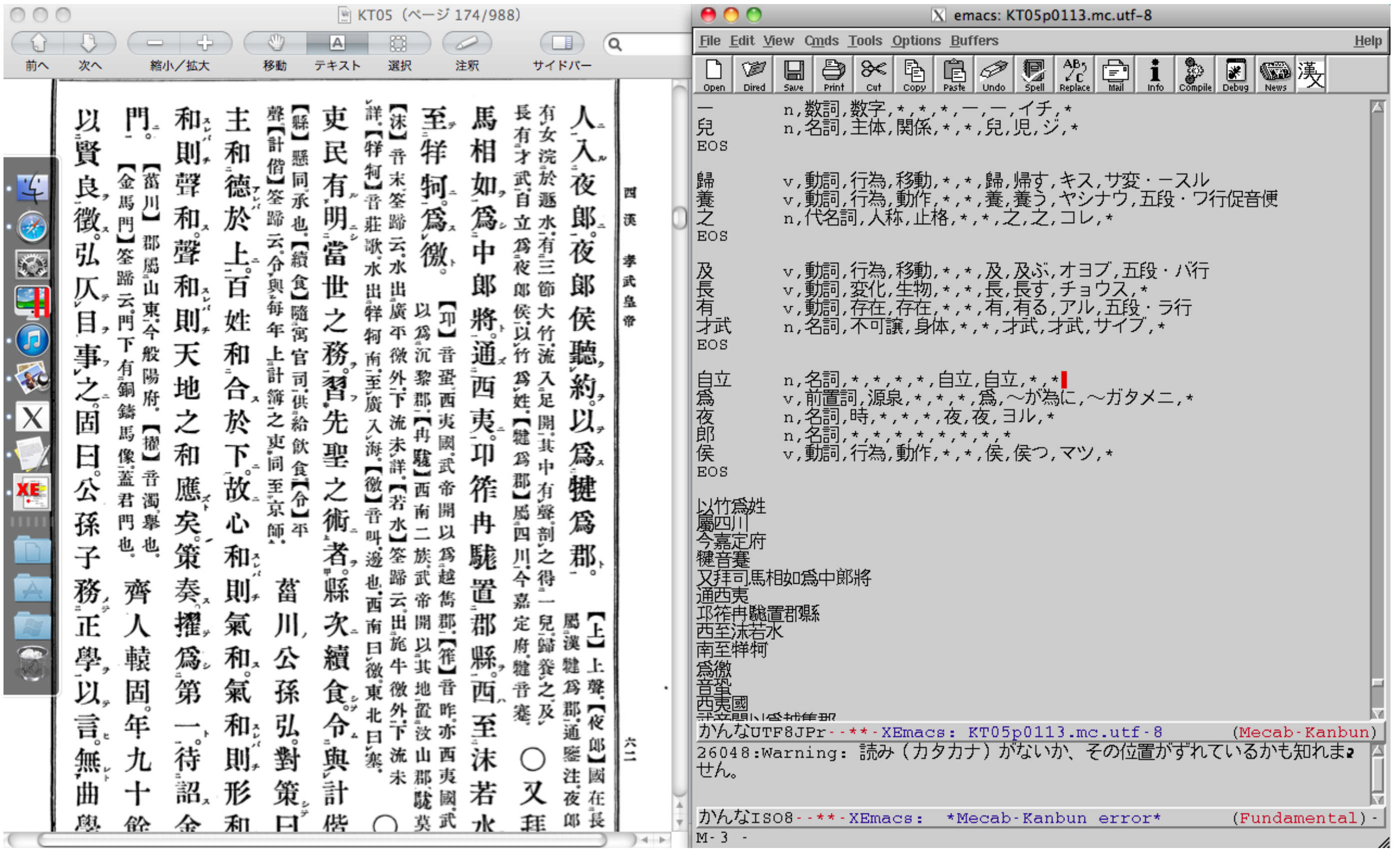
### B. MeCab Word-Class System for Classical Chinese

Here we present our four-level word-class system suitable for classical Chinese on MeCab.

The top level, which we call "word-superclass," is designed to represent the predicate-object structure of classical Chinese: "n" represents objectives, "v" represents predicates, and "p" represents others.

The second level is the ordinary word-class of classical Chinese: 名詞 (noun), 代名詞 (pronoun), 数詞 (numeral), 動詞 (verb), 前置詞 (preposition), 副詞 (adverb), 助動詞 (auxiliary verb), 助詞 (particle), and 感嘆詞 (interjection). In our system, 名詞, 代名詞 and 数詞 compose "n" word-superclass; 動詞, 前置詞, 副詞 and 助動詞 compose "v" word-superclass; 助詞 and 感嘆詞 compose "p" word-superclass. We have excluded 形容詞 (adjective) from our system and have joined 形容詞 into 動詞, since, in classical Chinese, there exists no essential distinction between 動詞 and 形容詞 [4].

The third and fourth levels are word-subclasses to describe detailed behavior of the words in classical Chinese



西漢　孝武皇帝　六二

人入夜郎。夜郎侯聽約。以爲犍爲郡。【上】上聲。【夜郎】國在長屬漢犍爲郡。通鑑注。夜郎有女浣於遯水。有三節大竹流入足開。其中有聲。剖之得一兒。歸養之。及長有才武。自立爲夜郎侯。以竹爲姓。【犍爲郡】屬四川。今嘉定府。犍音蹇。○又拜司馬相如爲中郎將。通西夷。邛筰冉駹置郡縣。西至沫若水至牂牁。爲徼。【邛】音蛩。西夷國。武帝開以爲越巂郡。【筰】音昨。亦西夷國。武以爲沉黎郡。【冉駹】西南二族。武帝開以其地置汶山郡。駹莫【沫】音末。箋踰云。水出廣平徼外。下流未詳。【若水】箋踰云。出旄牛徼外。下流未詳。【牂牁】音莊歌。水出牂牁南。至廣入海。【徼】音叫。邊也。西南曰徼。東北曰塞。○吏民有明當世之務。習先聖之術者。縣次續食。令與計偕【縣】縣同承也。【續食】隨寓官司。供給飲食。【令】平聲。【計偕】箋踰云。令與每年上計簿之吏。同至京師。菑川公孫弘。對策曰主和德於上。百姓和合於下。故心和則氣和。氣和則形和和則聲和。聲和則天地之和應矣。策奏。擢爲第一。待詔金門。【菑川】郡屬山東。今般陽府。【擢】音濁。舉也。【金馬門】箋踰云。門下有銅鑄馬像。蓋君門也。齊人轅固年九十餘以賢良徵。弘仄目事之。固曰。公孫子務正學以言。無曲學

```
一          n,数詞,数字,*,*,*,一,一,イチ,*
兒          n,名詞,主体,関係,*,*,兒,児,ジ,*
EOS

歸          v,動詞,行為,移動,*,*,歸,帰す,キス,サ変・ースル
養          v,動詞,行為,動作,*,*,養,養う,ヤシナウ,五段・ワ行促音便
之          n,代名詞,人称,止格,*,*,之,之,コレ,*
EOS

及          v,動詞,行為,移動,*,*,及,及ぶ,オヨブ,五段・バ行
長          v,動詞,変化,生物,*,*,長,長す,チョウス,*
有          v,動詞,存在,存在,*,*,有,有る,アル,五段・ラ行
才武        n,名詞,不可譲,身体,*,*,才武,才武,サイブ,*
EOS

自立        n,名詞,*,*,*,*,自立,自立,*,*
爲          v,前置詞,源泉,*,*,*,爲,～が為に,～ガタメニ,*
夜          n,名詞,時,*,*,*,夜,夜,ヨル,*
郎          n,名詞,*,*,*,*,*,*,*,*
侯          v,動詞,行為,動作,*,*,侯,侯つ,マツ,*
EOS

以竹爲姓
屬四川
今嘉定府
犍音蹇
又拜司馬相如爲中郎將
通西夷
邛筰冉駹置郡縣
西至沫若水
南至牂牁
爲徼
音叫
西夷國
武帝開以爲越巂郡
かんなUTF8JPr--**-XEmacs: KT05p0113.mc.utf-8        (Mecab-Kanbun)
26048:Warning: 読み（カタカナ）がないか、その位置がずれているかも知れま
せん。

かんなISO8--**-XEmacs:  *Mecab-Kanbun error*       (Fundamental)-
M-3  -
```

Fig. 1. Screenshot of an Authentic Textbook (漢文大系) and Our MeCab-Corpus Editor

texts. For example, we show our ideal result for a sentence “自立爲夜郎侯” below in MeCab format.

| | |
|---|---|
| 自 | v, 副詞, 範囲, 限定,*,*, 自, 自ら, ミズカラ,* |
| 立 | v, 動詞, 行為, 役割,*,*, 立, 立つ, タツ, 五段・タ行 |
| 爲 | v, 動詞, 行為, 役割,*,*, 爲, 為る, ナル, 五段・ラ行 |
| 夜郎 | n, 名詞, 主体, 国名,*,*, 夜郎, 夜郎, ヤロウ,* |
| 侯 | n, 名詞, 人, 役割,*,*, 侯, 侯, コウ,* |

For the word “自” we categorize its top level (word-superclass) into “v” (predicates), its second level (word-class) into 副詞 (adverb), its third level (word-subclass) into 範囲 (describing field or range), and its fourth level (word-subsubclass) into 限定 (restriction). The fifth to tenth columns are annotations for human consumption and are not used by MeCab.

### C. Making Corpus and Dictionary for MeCab

In order to make a corpus of classical Chinese for MeCab, we have developed a MeCab-corpus editor based on XEmacs CHISE [5], [6]. In our MeCab-corpus editor we first input typical sentences from classical Chinese texts. Second we push 漢文-button of the editor, then we obtain a morpheme sequence temporarily segmented by MeCab. Third we edit the sequence to categorize its words, looking up authoritative textbook references of the sentences (Fig. 1). And last we include the morpheme sequence in our corpus for classical Chinese.

Our corpus for classical Chinese on MeCab now includes about 18,000 sentences, written in our four-level word-class system. Our dictionary for classical Chinese on MeCab includes about 5,000 words, which we categorized into our four-level word-class system. We keep increasing our corpus, and we also keep selecting new words from our corpus to add them into our dictionary.

## IV. Conclusion

In this paper we proposed a method to analyze classical Chinese texts using a morphological analyzer MeCab. We also proposed our four-level word-class system and constructed a MeCab-corpus editor to make our corpus for classical Chinese. Now we keep increasing our dictionary and corpus for classical Chinese to make the analyzer more powerful and efficient.

## References

[1] K. Yasuoka, “Toward a Syntactic Analysis of Classical Chinese Texts,” *Osaka Symposium on Digital Humanities 2011*, pp. 34.

[2] T. Kudo, K. Yamamoto and Y. Matsumoto, “Applying Conditional Random Fields to Japanese Morphological Analysis,” *2004 Conference on Empirical Methods in Natural Language Processing*, pp. 230–237.

[3] T. Kudo. (2007) mecab-ipadic [Online]. Available: http://code.google.com/p/mecab/downloads/detail?name=mecab-ipadic-2.7.0-20070801.tar.gz

[4] N. Yamazaki, T. Morioka and K. Yasuoka, “Refactoring of Wordclasses for Morphological Analysis of Classical Chinese,” *The Computers and the Humanities Symposium じんもんこん 2012*, pp. 39–46.

[5] T. Morioka, “CHISE: Character Processing Based on Character Ontology,” *3rd International Conference on Large-Scale Knowledge Resources LKR 2008*, pp. 148–162.

[6] T. Morioka, “古典中国語形態素コーパス編集システムの開発,” *23rd Annual Workshop for Oriental Studies Computing* 東洋学へのコンピュータ利用, 2012, pp. 75–83.