Manual POS Tagging for Middle Chinese Corpus Processing

Yancheng Zhang[1]
College of Chinese Language and Literature
Wuhan University
Wuhan, China
e-mail: zyczyc@gmail.com

Wan Sun[2]
College of Chinese Language and Literature
Wuhan University
Wuhan, China
e-mail: arielsun1990@gmail.com

Abstract—**This paper introduces a program practice on Middle Chinese corpus for CLP and historical grammar research. It discusses the principles and approaches which focs on how to choose texts for corpus, word segmentation and POS tagging. It also describes the contents and developing processes on the Specification of POS Tag of Middle Chinese for CIP.**

Keywords—component; Part-of-speech(POS) manual tagging; Middle Chinese; Chinese word segmentation; Corpus processing

I. Introduction

POS tagging for Middle Chinese Language corpus processing is an intrinsic requirement in the study of the history of Chinese grammar, which has a precise and depth development tendency, furthermore, in the new view of language resource, it also a necessary choice for the study of the history of Chinese grammar in the information age. To support construction of various corpuses, the national 973 Plan vigorously advocates establishing Chinese Linguistic Data Consortium. Wang Yunlu(2010) presented 8 missions to the vocabulary study of the middle ancient Chinese Language, the 6th is to process and construct the corpus. Dong Zhiqiao(2011) indicated that the establishment of ancient Chinese electronic corpus is far behind the advances of research, currently, the number of separated ages Chinese corpuses available is not large, especially that used for middle ancient Chinese language research. Establishing the required corpuses is a urgent task for academia to deeper the research of ancient Chinese.

Modern Chinese word segmentation has accessed into a preliminary practical phase, whereas, there are shortcomings in the field of applications of ancient Chinese tests and the research achievements are also extremely rare in the field of middle ancient Chinese. A limited number of papers are concentrated in primitive-Chinese language and proto-Chinese language automatic word segmentation (Qiu Bin(2008), Liang Huang(2002) etc.). Taiwan's Academia Sinica structured proto-Mandarin Chinese tagged corpus including POS-tagging and word segmentation that have referenced significance. Dong Zhiqiao(2011) currently indicated finishing ancient Chinese Language research-based construction scheme which aims to establish a multi-level research corpus including information such as meaning tagging and grammatical status. It said the first phase is expected to intake 5000,000 words original corpus reaching 80,000,000 words' capacity which including following 6 sections: collating texts repository, word senses corpus, syntax trees corpus, parallel corpus of ancient and modern Chinese, middle ancient verse phonological corpus and integrated system. He proposed that at the beginning tagging word artificially, tagging word automatically after the tagging software trialed successfully, then there tagging manually. This project is closed to our idea. Nonetheless, this project put forward that the main target is each sense tagging mainly refers to *Han-yu ta tz'u-tien*, which differs from our purport. Their target is obviously served for interpreting meanings and dictionary compilation. There will be some differences between word segmentation according to this objective and CIP word segmentation. Dong proposed that construct syntax trees required tagging middle ancient corpus sentence-for-sentence and then word-for-word, after that in accordance with each word's grammar constitute hierarchical level, then, building syntax trees. This is a promising project.

From 2009, we start to work on manual speech tagging of middle ancient Chinese language which based on multiple paradigm tests. This work is a trial of establishing research-based middle ancient Chinese corpus. We have completed following works: initially formulated *the Standard of POS Tag of Middle Ancient Chinese for CIP* (revised for this conference currently), finished more than 155,000 words manual speech tagging of middle ancient Chinese language which based on multiple paradigm tests, finished middle ancient Chinese feature words list which contain literatures searching pages(4650 words). This work is the preparation of the automatic words segmentation on a larger scale in the next step. The expected results will not only promote Wang Yunlu's research on special book vocabulary, domain vocabulary, monographic vocabulary, synchronic vocabulary, special book dictionary compilation vocabulary, but also lay the foundations for establishing higher level corpus of ancient Chinese grammar, semantic, pragmatic. This paper reports our vision, methods and plan on middle Chinese language material selection, manual word segmentation and speech tagging.

II. Language material Selection

In the so-called *Digital-riching* age, semplice middle ancient Chinese texts is not hard to find: Guoxue Baodian, Han Ji full-text search system, Da Zhengzang retrieval system of CBETA, The Si Ku Quan Shu, Basic Ancient Books

Database of China, Han Dian System of Taiwan, Han Da system of Hongkong, are all available. But most texts in these are raw corpuses which can't meet the requirement of middle ancient Chinese research and information management because of lacking synchronic information and collation.

We have established a small scale corpus mainly based on *Middle Ancient Chinese Reader* written by Fang Yixin and Wang Yunlu, furthermore referred to other papers, The compiling of Reader inherited the vision of *Ancient Chinese Reader* (Tianjing Renmin Press, 1981) and *Proto-Mandarin Chinese Reader* (Shanghai Education Publishing,1985). Through extensive research and reading the Han Wei Jin and the Northern and Southern Dynasties texts, editors carefully chosen corpus reflecting the characteristics of oral English in this period. The Reader divided into following parts: Buddhist sutras, fictions, poetries, notes and others. All the works are broadly chronological, with including pandect of each category. The including pandects introduce name of the books and articles, list examples and annotations of difficult words, oral phrases, ensure it can comprehend by analogy and sufficient evidence. No matter checking from sampling properties, or from the quality of writing, the Reader is an excellent work of middle ancient Chinese corpus collection. Based on proofed Reader, we have chosen 100,100 words from each books, such as The Biographies of Eminent Monks, The Works of *Master Bao Pu*, *Southern Qi book*, *Poetry*, *Collection of Tao Yuanming*, *The Za Baozang Buddhist Sutras*, *Collection of Buddhist Sutras Fo Benxing*, *Epitaph of Northern Wei Dynasty* (All the books are used the general edition. The books are collated. The total amount of words of this corpus reached 155,000.

Although the corpus is not on a large scale, it almost contains all kinds of ancient Chinese language material. We have chosen some material such as *The Reader* and some epitaphs which have strong written color and various theme types. Facts proved that such chosen method has well served grammar research and information processing of corpus establishing. For example, such sentence "Three Zhang Two Lu, Double Pan One Zuo" in Poetry which is not common in oral language material. If such material not be chosen, there will be none example in the "abbreviation (j)" category of the *Standard of POS Tag of Middle Ancient Chinese for CIP* consider national codes that will result in some significant linguistic phenomena be omitted.

III. Segmentation and Tagging

A. *Because we do the language material handling work manually, the tagging and segmentation work was synchronized.*

Syncopate of words consulted *Contemporary Chinese Language Word Segmentation Specification for Information Processing* （GB/T13715-92）and *Segmentation Principle for Chinese Language Processing* of Taiwan. We don't draw up specialized segmentation principle, only using definition, descriptions, example words, example sentences and labeled instance in the formulating POS-tagging principle to embody principles and methods of segmentation. We will summarize practical tagging's experience to complete segmentation program.

Due to diachronic language material's particularity, such as lack of language support, incompletion and state alteration of word class system, difficult and confusing words, words' tagging of middle ancient Chinese language material is a professional and technical task based on words function. Based on classical distributed description theory, we have formulated the *Standard of POS Tag of Middle Ancient Chinese for CIP* by comparing and synthesizing the following principles: *Contemporary Chinese Language Word Segmentation Specification for Information Processing* （GB/T13715-92）, *Standard of POS tag of Contemporary Chinese for CIP*（GB/T20532-2006）, *The Southern and Northern Dynasties Historical Grammar*(Liu Shizheng, 1992), *Dictionary of Shi Shuo Xin Yu*(Zhang Wanqi,1998), *Ancient Chinese Function Words Dictionary*(1999), *A Grammar of Spoken Chinese*(Zhao Yuanren, 1996), *Specification for Corpus Processing at Peking University: Word Segmentation, POS Tagging and Phonetic Notation*(2003), Taiwan's *Information Processing Using Words Specification* (1997).

B. *Range, terms, definition, general provisions, parts of speech, other classification of segmentation units and markup codes are based on Standard of POS tag of Contemporary Chinese for CIP（GB/T20532-2006）and fine-tuned.*

The range of the Standard of POS Tag of Middle Ancient Chinese for CIP: It provided that the middle ancient Chinese's classification of segmentation units and tagging codes applied to middle ancient Chinese information processing and grammar research. It also provide a reference for the ancient Chinese and modern Chinese words tagging.

Terms and definition: Ancient Chinese indicates the documentary language from the Eastern Han Dynasty to Sui Dynasty. Segmentation units is a kind of unit having certain grammar function and using for information processing, including words, phrases, idioms, acronyms, enclitics, following compositions, punctuation tags, non-character symbols, unknown words, etc. Parts of speech mainly classify according to the grammatical functions, considering research requirement and the particularity of information processing, it can extended to character string of specific grammatical character strings. The tittle of Standard of POS-tagging is similar to that of Standard of POS tag of Contemporary Chinese for CIP（GB/T20532-2006）. It is a brief address including non-word segmentation units. Tagging is a code (sometimes used as verbs) used in the texts as segmentation units' classification.

Principles to be used in the process of naming, nominal, constant bit: rely on standard, base on function, consist name and actuality, have compatibility and intercommunication, differ details from summary. There are 13 kinds of first level parts of speech: nouns, numerals, measure words, adjectives, attributive words, verbs, adverbs, pronouns, prepositions, conjunctions, auxiliary words, interjection, onomatopoeias. There are 7 other kinds of first level parts of speech: idioms, acronyms, enclitics, following compositions, reduplication or sound repetition words, non-morpheme words, unknown words and others. First level parts of speech has similar number with *Standard of POS tag of Contemporary Chinese for CIP*（GB/T20532-2006）and there is a little difference in the content between them. Second level parts of speech have 44 parts of speeches, 28 parts more than *Standard of POS tag of Contemporary Chinese for CIP*. Third level parts of speech have 44 parts of speeches which is not installed in the *Standard of POS tag of Contemporary Chinese for CIP*. Finally the classification of segmentation units is tabulated as follows:

1 Nouns（N）

1.1 Common nouns（n）

1.1.1 Individual Nouns

1.1.2 Collective Nouns

1.1.3 Material noun

1.1.4 Abstract noun

1.2 Temporal noun（nt）

1.3 Direction noun（nd）

1.4 Place noun（nl）

1.5 Name（nh）

1.6 Geographical name（ns）

1.7 Nomen（nn）

1.8 Official position name（na）

1.9 name of book, title name, name of music instrument（nb）

1.10 Other proper noun（nz）

2 Numerals（M）

2.1 Cardinal numeral（m）

2.2 Ordinal numeral（mo）

2.3 Question numerals(mi)

2.4 Uncertain numerals (ma)

2.5 Fractional numeral (mf)

3 Measure words（Q）

3.1 Quantifier of substantive（qn）

3.1.1 Individual quantifier（qns）

3.1.2 Measure quantifier（qnm）

3.1.3 Collective quantifier（qnc）

3.1.4 Temporary quantifier（qnp）

3.2 Verbal quantifier（qv）

3.2.1 Specific verbal quantifier

3.2.2 Borrowed verbal quantifier

4 Adjectives（A）

4.1 Qualitative adjective（aq）

4.2 State adjectives（as）

5 Attributive words (F)

6 Verbs（V）

6.1 Common verbs（v）

6.2 Modal verb（vu）

6.3 Directional verb（vd）

6.4 Link verb（vl）

6.5 Intransitive verb（vi）

6.6 Transitive verb（v）

7 Adverbs（D）

7.1 Adverb of degree

7.2 Range adverbs

7.3 Time adverbs

7.4 Modal adverbs

7.5 Modality adverbs

7.6 Negative adverbs

7.7 Honorific adverbs

7.8 Referring adverbs

8 Pronouns（R）

8.1 Personal pronoun（rh）

8.1.1 First-person pronouns（rhf）

8.1.2 Second-person pronouns（rhs）

8.1.3 Third-person pronouns（rht）

8.1.4 Referral pronoun（rho）

8.1.5 Reflexive pronoun（rhs）

8.1.6 Polite personal pronoun（rhp）

8.2 Demonstrative pronoun（rd）

8.2.1 Demonstrative pronoun

8.2.2 Proximal pronoun

8.2.3 Other demonstrative pronouns

8.2.4 Phantom reference pronouns

8.2.5 Extend demonstrative pronoun

8.2.6 Nonspecific pronouns

8.3 Interrogative pronoun（ri）

9 Prepositions（P）

9.1 Time and place preposition

9.2 Reasoning preposition

9.3 Pattern preposition

9.4 Personnel preposition

10 Conjunctions（C）

11 Auxiliary words（U）

11.1 Structure auxiliary（us）

11.2 Dynamic auxiliary（ua）

11.3 Modal auxiliary（um）

11.4 Lining sound auxiliary（uy）

12 Interjection（E）

13 Onomatopoeias（O）

14 Idioms（I）

15 Acronyms（J）

16 Enclitics（H）

17 Following compositions (K)

18 Reduplication or sound repetition words（DY）

19 Non-morpheme words and unknown words（X）

20 Others(W)

20.1 Punctuation tags（w）

20.2 Non-character symbols（ws）。

20.3 Other unknown words（wu）

21 Multiple attribute tag for example

Compared with Standard of POS tag of Contemporary Chinese for CIP（GB/T20532-2006）, the single underlined units are polymerized, the double underlined units are different. The deleted first level parts of speech are as follows: ” Morpheme word（g）” and “Noun-morpheme word（gn）” under it, ” Verbal morpheme word（gv）”, ” adjectival morpheme word（ga）”. The deleted second level parts of speech are as follows: “Organization name（ni）”, “Noun-idiom（in）, Verbal idiom（iv）, Adjectival idiom（ia）, Connected idiom（ic）” under “Idiom”, “Noun abbreviation（jn）, Verbal abbreviation（jv）, Adjectival abbreviation（ja）. We changed some definitions and added large numbers of example words, example sentences. Most added units have undefined tagging code in order to help researchers grasp the second level parts of speech more accurately by listing the example words and example sentences. It also makes a preparation to improve its accuracy in the future. We have redefined some tagging codes according to the prior requirement of middle ancient Chinese research and the objective requirement of handling corpus. For example, Official position name（na）, Polite personal pronoun（rhp）are both outstanding linguistic phenomenon of ancient Chinese. We also added other tag which the tittle doesn’t have. For example, under the “numerals” formulated: “When Measure words and numerals used together, first tagging separately, than use mp to tagging entirely.”

C. There are some examples of the differences between the Standard of POS Tag of Middle Ancient Chinese for CIP and the Standard of POS tag of Contemporary Chinese for CIP（GB/T20532-2006）.

1) In the first level, name of parts of speech and tagging are the same as that in the Standard of POS Tag of Middle Ancient Chinese for CIP. The 14th to 20th terms in the Standard of POS Tag of Middle Ancient Chinese for CIP are correspond to “Other segment” in the Standard of POS tag of Contemporary Chinese for CIP（GB/T20532-2006）, but deleted “Morpheme word (g)”, added “Reduplication or sound repetition words”, recomposed “Non-morpheme word（x）” into “Non-morpheme words and unknown words（X）”. The “morpheme words” has little usefulness for middle ancient Chinese research. “Non-morpheme words” related to some transliteration words, although the number of unknown words is not large, it’s necessary to ancient Chinese texts research.

The “common noun” in the Second level is tagged by “n”, not by “ng”, according to economically tagging. For example, common nouns are tagged “ng”, but in the international principle “n” is a redundant principle, so this wii add double codes manual tagging work as noun’ tagging is too much. For the same reason, we tagged punctuation tag as “w” instead of “wp”, common verb and transitive verb are tagged with “v” instead of “vt”, transitive verb are tagged with “vi”. “Organization name（ni）” of the second level in national principle represents proper names of groups, organizations, institutions. Some words, such as “fu” in the sentence of “Wu jin qie bao fu(I go to my home today.) “which used not so frequently in middle ancient age as that in the modern times, moreover, some parts of the words function is closed to place words, so we put them into “Other proper noun（nz）”.

2) About parts of speech and definition of segmentation unit: Usually, The Standard of POS Tag of Middle Ancient Chinese for CIP later was changed more targeted. For example, “Time noun（nt）” is regard as “contain common time quantifiers”, in the Middle Ancient Standard is expressed as “Indicate time and contain following forms: name of dynasties, year name, Timing GanZhi, seasons, moths, moment, etc.” In the International Standard, “Place noun” is defined as “express place”, we changed it into “express place, always used together with locality noun and prepositions or *verbs which show the place. This actually gives distribution framework roughly. In the International Standard, “Name （nh）” defined as “express people’ names and proper names”. In the Middle Ancient Standard, “Name（nh）” was defined as “express people’s names and appellation” in view of the title words’ special value and the distribution of “name” are basically the same with it. This method omitted separated “appellation” from “name” and can tag such words: “Yao xian sheng”, “A mu”, etc.*

3) About example words and quotes: All kinds of modern Chinese words segmentation standard generally do not give example sentences. We think that the ancient Chinese word segmentation standard should appropriately use examples. The example sentences can give researchers examples and enlightenment. Quotes’ format don’t follow standard of historical grammar research papers: just quotation tags to tagging clear simple examples and don’t indicate the source of examples in order to show the examples are chosen from real text. (For instance, determined number of collective classifier “Xuan”(“Yi xuan shi jiu zhong”). To help researchers to estimate the meaning of examples, contemporaneity and style classify, sometimes need tag the source of material, we tag the source of documents only list book names and tittle of the articles (sometimes add author’s names), not standard to a certain volume as from one chapter or one pass (such as "Ci shan qu yan ran yuan jin “(Wei book)). At present, the operation method of tracing documents’ source is useful because the middle ancient Chinese corpus and digital resources have been quite rich, using key words is easy to find the examples’ primitive source and to check reliable text conveniently.

The chosen of the example words first consider its representative, diversity and epochal character (For example, “Jian du” in the Sanskrit represent “Zhang, jie” in Chinese.) The chosen of the example words consider its clarity, distinctiveness and necessity. For example, “Suo” was used as interrogative words firstly in middle ancient times, if we only list “Suo” under the parts of speech tags; it is too difficult to be easily understood. So we list example sentences: ” 为欲所至？”（《增一阿含经》）、“诸臣即问：所从得此儿？”（《太子须大拿经》）。As another example, “Yi” can be used as the first-person pronouns（“亮答以为：‘己等三人，同受顾命，岂可相残戮！’”（《宋书》））it also be used as the second person pronoun（“某甲为霸府佐，……妓人奏曲，赞之，己亦学人仰赞和。”（《笑林》）），so example sentences should be discriminated. In some occasions, we make notes behind example sentences, for example: demonstrative pronoun “Neng”(“卦不能佳！”（《三国志》注引《吴历》）指称样状).

D. About tagging operation

An inseparable segmentation unit has more than two kinds of tagging attributes and can be labeled continuously. The last item is marking main function of the current context. For instance, “A bo bo” is a hell’s name which comes from onomatopoetic word, so tagging it: A bo bo/o/nz(“三阿罗逻，此患寒声；四阿波波，亦患寒声；五名睺睺，亦是寒声。此之三种，从声以名。”（隋•慧远《大乘义章》））.

When to segmentation of parts of speech or the whole unit, which need to segment again, tagging it by using “[ ]”.For example, interjection[呼呼/dy 呵呵/dy]o（“汝等比丘，何故如是作大高声？犹如世人诸诤斗起呼呼呵呵，其声犹如钓鱼之师，各各相竞趁逐诸鱼，各相唱唤。”（《佛本行集经》））, place noun[中/nd 天竺]ns，premises nouns[清凉/a 台/n]np，personal pronoun [阿/h 侬]rh（Be careful not use“阿/h 侬/rh” in order to distinguish and statistic parts of speech，not use “[阿/n 侬/rh]rh”to simplify tags）.

The standard for the label can be relaxed when tagging important grammatical processes of parts of speech and units. Motivating grammar research is one of our goals to establish the corpus, historical material must contain dynamic and transitivity of grammatical category, this phenomenon should be reflected in corpus tagging. The dynamic auxiliary system of middle ancient Chinese (Contain attempting auxiliary, this classification is the practicability of information processing, not strict classification in theory.)Because of immaturity, there existed some ambiguity phenomena and intermediate state. Suggest tag dynamic auxiliary（ua）instead of common verbs（v）, such as: “汝/rhs 好/d 思量/vi 看/ua”（姚秦《十诵律》）,“锄/v 得/ua[五/m 遍/qv]mq 以上/nd”（《齐民要术》）. Again for instance, there are controversies of whether “Wu”,” Bu”,” Fei” at the end of a sentence is negative words or auxiliary words. Under misty circumstance, suggest tag it modal auxiliary（um）, for example, “身体/n 轻健/a 不/ua？”（《鼻奈耶》）.

Tagging idioms can process flexibly. Many later created idioms can be use separately. For example, “jiao shu tu cheng” in the Shi shuo xin yu was written “jiao shu er tu cheng” in the variants in the hand-copied books of Tang dynasty. In the San guo zhi, it was written “jiao shu tu er tu qi cheng”. So we tag it separately: 交/v 疏/n 而/c 吐/v 诚/n. In middle ancient Chinese, some idioms and structures changed into word strings of phrases, suggest tag them as idioms: [交/v 疏/n 而/c 吐/v 其/rht 诚/n]i，康伯/nh 未/d 得/v 我/rhf[牙/n 后/nd 慧/n]i（《世说新语》））.

Tag the ambiguity units according to the context. For example, "Gao chang" "Yan qi" "Qiu ci" can be used as place names and family name, choice one from the two according to the context.

IV. Summary and imagine

We plan to extension corpus and raise the precision in gradually. Prioritize the research projects which can hot and difficult that can motivate the research according to the middle Chinese research and information processing requirement. We plan to formulate Middle Ancient Chinese Information Processing Word List based on pre-existing word list (contain 15 kinds of treatises on middle ancient Chinese research). Moreover, we plan precipitate monosyllable and two-syllables with items, then tag their part of speech. This work will be based on The Great Chinese Dictionary and Chinese Etymology Dictionary to make preparation for automatic word segmentation of large scale middle ancient Chinese corpus.

Although we have done little work and the expected target lower than the project of Dong Zhiqiao, we still believe this work is necessary to middle ancient Chinese research. Only more and more people paying attention to or engage in middle Chinese information processing and application research can accelerate the process of accumulation and share experience, can gradually complete the different types, different levels of research or engineering tasks, finally comprehensive use our country language resources in ancient times.

REFERENCES

[1] Liang Huang etl. Part-of-Speech Tagging for Old Chinese. Papers (ID 115) accepted for TSD 2002.

[2] BaiYuling. ShiSanJing Dictionary Part-of-speech Tagging Problem. Dictionaries research, 2000(6).

[3] Dong Zhiqiao. To lay a solid foundation for middle Chinese research——discussion of “Middle Chinese research corpus”. Journal of Yanshan University, 2011, 12(1):1~6.

[4] Fang Yixin, Wang Yunlu. Middle Chinese Reader （corrected edition）.Shanghai: Shanghai education press, 2006.

[5] The education ministry of language application research institute of computational linguistics room "corpus processing" group. Information Processing with the Modern Chinese Parts-of-speech and Part-of-speech Tags Set Norms（Exposure Draft）,2002-04-08.

[6] Liu Shiru. The measure of the Southern and Northern Dynasties.Beijing: Zhonghua Book Company, 1965.

[7] Liu Shizhen. History Grammar of the Southern and Northern Dynasties.Nanjing: Nanjing University Press, 1992.

[8] Qiu Bing,Huang Pujuan. Based on Chinese Information Processing of Ancient Chinese Word Segmentation Research. Microcomputer Information.2008(24):100~102

[9] Shi Min.etc. CRF Based Research on a Unified Approach to Word Segmentation and POS Tagging for Pre-Qin Chinese. Journal of Chinese Information Processing, 2012(2):39~45.

[10] Wang Yunlu. Vocabulary History of Middle Chinese .Beijing: The Commerical Press, 2010.

[11] Wei Peiquan, Huang Juren, etc. Construct a Diachronic Research of Language with Oriented History Corpus. Chinese Journal of Computational Linguistics,1997(1):131~145

[12] Xiao Hong. Syntax research of Luo yang jia lan.Beijing: China Social Sciences Press, 2008.

[13] Xiao Guozhen. Information processing of Chinese Semantic Resource Construction the Current Situation Analysis and Prospect. Yangtze River Academic, 2007(2).

[14] Yu Shiwen, etc.The BasicProcessing of Contemporary Chinese Corpus at Peking University, Journal of Chinese Language and Computing, 2003,13(2):121~158

[15] Yu Shiwen, etc. Beijing University Modern Chinese Corpus Basic Specification. Journal of Chinese Information Processing, 2002(5)(6).

[16] Zhang Yingjie. Dictionary Based Learning Guidance and Half of Ancient Chinese Full text Meaning Labeling. The 11 th National Computational Linguistics Conference, 2011.