Theoretical Framework of Mongolian Word Segmentation Specification for Information Processing

Tong Laga

Library ,Department of Minority Language and Literaturel

Quanzhou Normal University, Minzu University of China

Quanzhou ,Beijing ,China

bolor@163.com

Xiaobing Zhao

College of Information Engineering,

Minzu University of China

Beijing China

nmzxb_cn@163.com

Abstract: The establishment of Contemporary Mongolian word segmentation specification for information processing has a great significance in the standardization of information processing, the compatibleness of different systems, the sharing of corpus, grammatical analysis, and POS tagging. The present paper studies the framework of Mongolian word segmentation including guidelines, formulating principles, styles, scopes of segmentation units, establishment foundation, structure of the specification and so on, and lays the theoretical foundation for this specification.

Key words: Contemporary Mongolian word segmentation specification for information processing, theoretical framework, guidelines

INTRODUCTION

Language information processing is a project which covers a large scale. Since the research work in every field is based on the segmentation of word units, its correctness is of great significance. Word segmentation specification for information processing is the rules for word units segmentation (including words and non-word units), and provides the unified criteria of word units defining for language information processing. It also promotes the compatibleness and sharing between different systems, standardizes the establishment of corpus, speeds up the rate of information processing, and helps solve the problems of ambiguity and OOV words. Language information processing needs a mature word segmentation specification as its foundation, so does Mongolian information process, which also calls for the establishment of corresponding word segmentation specification.

The national standard *Contemporary Chinese Language Word Segmentation Specification for Information Processing* （ *Word Segmentation Specification* for short）was issued in 1993, while there are no word segmentations which are in conformity with the linguistic features of Chinese minority languages issued till now. The paper tries to study the Mongolian word segmentation and the present corresponding research results, and also studies the various kinds of Chinese languages word segmentations and the segmentations of the units which are larger or smaller than words in Mongolian. With the support of corpus, the segmentation will be verified and perfected in time and then will help to lay the foundation for the first segmentation for Chinese minority language.

The present thesis studies the theoretical framework of the establishment of Mongolian word segmentation specification for information processing (Mongolian word segmentation specification for short), which relates to the basic problems, such as the guidelines, formulating principles, styles, scopes of segmentation units, establishment foundation, structure of the specification and so on. It also provides the basis for the formulation of a generally-accepted, practical and systematic word segmentation specification which will be in line with the specific linguistic features.

.GUIDELINES AND FORMULATING PRINCIPLES OF THE EATABLISHMENT OF WORD SEGMENTATION SPECIFICATION

All the Chinese language word segmentation specifications set up their guidelines and formulating principles at the initial stage of their establishment. Mongolian word segmentation specification should take the Chinese word segmentation specifications as references, and set up its own guidelines and formulating principles according to its own linguistic features.

A. Guidelines

- Information processing centered and linguistic researching subordinated: the segmentation should take

978-0-7695-4554-7/11 $26.00 © 2011 IEEE
DOI 10.1109/IALP.2011.45

information processing as its center.

- This specification should adopt the findings of linguistic research, and make itself in accordance with the linguist theories.
- People-oriented and serving the people: ultimately, the segmentation and segmentation specification are put forward to serve the people; therefore, it should take people's language sense into consideration.
- Combination of quantitative principles and qualitative principles: At the time when computational linguistics and corpus linguistics are developing rapidly, the research method combined both stipulation and calculation, both quantitative principle and qualitative principle, should be the mainstream research method and the leading trend; therefore, the segmentation specification should combine both the quantitative principles and qualitative principles.

B. Formulating principles

- Scientificalness and preciseness: the systematical and theoretical preciseness should be guaranteed; segmentation should be well-grounded and in line with the traditional practice as a whole.
- Completeness: it should be applied to deal with all the linguistic phenomena.
- Consistency: all the rules should be in consistency with each other,

. THE STYLE OF SEGMENTATION DEFINING OF WORD UNITS

The national standard *Contemporary Chinese Language Word Segmentation Specification for Information Processing* （ *Word Segmentation Specification* for short）first figured out the style of word segmentation specification. Facing with the problem that the word segmentation should be described from the perspectives of grammar or from the perspectives of word formation, the standard setters adopted the former at last, while, on the contrary, *Specification for Basic Processing of Contemporary Chinese Corpus at Peking University* adopted the latter. Mongolian word segmentation specification is in favor that "the rules should be demonstrated from the perspectives of grammar". In order to be in accordance with *Word Segmentation Specification*, it is suggested to take the perspectives of grammar. The segmentation units cover the 14 basic words defined in *Information-processing-oriented Mongolian Tag Set*, and word segmentation units which is smaller than words, such as characters, connective letters and affixation, and word segmentation units which is equal to or larger than words, such as compound words, idioms, proper nouns, set phrases, idiomatic expressions, acronyms or abbreviations and nominal terms. That is to say, the segmentation units cover 14 basic words, and 10 non-word units, 24 groups in total.

ESTABLISHMENT FOUNDATION AND BASIC STRUTURE OF THE SPECIFICATION

A. Establishment foundation of the specification:

the specification should center on information processing, and set up the segmentation rules for Contemporary Mongol according to its features and laws. On the one hand, the foundations of Mongolian word segmentation specification are as follows: *Mongolian Grammatical Information Dictionary*, *Information-processing-oriented POS Tag Set*, word formation, orthoepy and orthography, segmentation and reduction rules, scopes and segmentation of compound words, segmentation rules of etyma of compound words, various kinds of Mongolian dictionaries; on the other hand, it should also be based on the findings of the Chinese and other agglutinative language researches, such as word segmentation specification, POS tag set, and lexical analysis and POS tagging system.

B. Basic structure of the specification

The basic structure of the specification is made up with subject contents, scope of application, quoted standards, general rules, detailed rules and remarks.

The general rules stipulate the rules for space and punctuations （《 ERDENI-YIN T0BCI 》）, loosely-integrated two-character words（MARAYIN=JIDGUJV）, closely-integrated or meaning- Transferred two-character words (HAR_A=ARIHI), words with more than two characters, set phrases, idioms, proverbs, mottos and sayings （MAGV=HOMON=U=AMA=ACA=M0GAI= MELEHEI=CVBVRAN_A）, abbreviation（GVRBAN =YEHE= ARADCILAL）, transliterated foreign words (deYItA =HOMORGE）, non-Mongolian codes（2.67）and closed words（HAYA）.

The detailed rules follow the way of dividing the words into 14 groups, that is, nouns, adjectives, verbs,

numerals, quantifiers, pronouns, tense morphemes, adverbs, modals, postposition modifiers, modal particles, conjunction, and interjections, and give a detailed definition of the scopes and segmentation of the-above-mentioned 14 groups of words as well as people's names, place names, names of organizations, other proper nouns, nominal terms, and temporal nouns. Detailed segmentation rules of connective letters, generic words, reduplicative words, comparative adjectives and ambiguous word strings are also set up in this part.

Ⅴ.CONCLUSION:

The standardization of language is the foundation of and the major problem to be solved in the linguistic study. The lag of issuance of the standard and specification for Mongolian will have a direct effect on the development of Mongolian information processing. The paper tries to build up the framework of Mongolian word segmentation and study its guidelines, formulating principles, styles, scopes of segmentation units, establishment foundation and structure. By doing this, it is hoped that it will make contribution to the build-up of Mongolian word segmentation and lay the foundation for the national standard.

NOTE: words in parentheses are examples.

REFERENCES

[1] Liu Yuan, *Contemporary Chinese Language Word Segmentation Specification for Information Processing and Automatic Word Segmentation Method* [M]. Beijing: Tsinghua University Press, 2000,11

[2] Xu Shun .the research and implementation of computability on the Chinese segmentation specification[D]. Master thesis of Suzhou University，2006,21

[3] Yu Shiwen Etc.The basic processing of contemoracy Chinese corpus at peking university[J].journal of Chinese information processing[J].2002（5）49-64

[4] Yin Hailiang **.A Corpus-based Preliminary Study of Automatic Recognition of Modern Chinese Affixes and Derivative Words**[J].Applied Linguistics, 2010（1）127

[5] Li Yu-mei,CHEN Xiao, JIANG Zi-xia, **Three Complements to Make Better Guideline of Chinese Word Segmentation** [J].Journal of Chinese Information Processing,2007（5）3-7

[6] Xu Jialu, Fu Yong.Modern Chinese Information Processing of Chinese Vocabulary

[M].Guangzhou:Guangdong:Education Publishing House,2006

[7] Administration of Technical Supervision.Contemporary Chinese language word segmentation specification for information processing [S].Beijing:National Standard Press,1993

[8] Qenggeltai.Monggolian Granmar[M]. Hohhot :Inner Mongolia People's Publishing House，1992

[9] Nasanurt , Shuqin .The Standardization for Mongolian Information Processing[J]. Journal of the Central University for Nationalities(Philosophy and Social Sciences Edition)，2007(6)115-122