Exploitation and Application of the Mongolian Linguistic Knowledge Resource*

Na Shunwuritu
010021, Mongolian Language Institute, Hohhot, China
qingjirvm@126.com

Abstract

It is important to exploit and make full use of knowledge-based resources, which can be used for Mongolian information processing. The exploitation of these knowledge-based resources is focused on the following aspects: a compilation of a Mongolian grammatical information dictionary, semantic research for Mongolian information processing, in-depth processing of Mongolian corpus, etc. A knowledge-based system includes the knowledge itself as well as the formalization of the knowledge. The knowledge-based system can be a fundamental basis for correctly parsing and generating Mongolian words and sentences, since it provides a full description of the grammatical and semantic features of the Mongolian language. In this paper, the author discusses his opinions from the angle of establishing Mongolian codes and developing all types of Mongolian applied systems.

1. Introduction

The research on Mongolian information processing, which started in the 1980s, consists of three main parts: to exploit the data resource, to develop the knowledge resource, and to train competent researchers.

The exploitation of the data resource is to establish various corpora. For example, middle age Mongolian corpus, contemporary Mongolian corpus, the word-frequency statistics of one million Mongolian words, etc. Although these corpora are on a small scale and shallowly processed, they provide credible linguistic data for the research of Mongolian information processing. More importantly, they help make it possible to change Mongolian language research from a traditionally “qualitative” analysis to a “quantitative” one (though this kind of change is still only beginning). The exploitation of the knowledge-based resource is composed of some basic criteria: a compilation of Mongolian grammatical information dictionary, semantic research for Mongolian information processing, and in-depth processing of Mongolian corpus, etc.

The methodology of computational linguistics is still a completely new concept in Mongolian language study, therefore, above all, we have to fully “exploit” the usable “knowledge” which is derived from traditional grammar research. At the same time, with new theories and methods we should go on intensely studying the Mongolian language and constantly transform these fruits of our research into a knowledge base, in order to continually provide system developers and language researchers with the knowledge of the Mongolian language; that is the so-called “exploitation”.

We aim to train high-level researchers who are devoted to Mongolian information processing. In 2002, Inner Mongolia University began recruiting doctoral students to specialize in Mongolian information processing. It is worth mentioning here that the Institute of Computational Linguistics of Peking University greatly supported us in the training of the professionals for Mongolian information processing.

The exploitation and application of the linguistic knowledge resource will be detailed later in this paper.

The research on the language information processing is going deeper and the requirement for the

* Supported by National 863 Program 2003AA115510 and the National Natural Science Foundation of China under Grant No.36963005
language knowledge is also increasing. “The language knowledge base is indispensable to a NLP system and the scale and quality of the base can be a decisive element of the system,” said Professor Yu Shi Wen. The 20-year history of the Mongolian information processing reconfirms this idea.

For the first time, the term “the knowledge base of the Mongolian language” is brought forward. The research on Mongolian information processing has continued for over 20 years. Some graphitic and linguistic knowledge has been used during the character and word processing phases, but thus far the knowledge is still disorderly. It might be used for a concrete system or included in a concrete project. It has not yet been systematically organized in a knowledge-based system which can meet all kinds of needs and which is comprehensive and accurately described. There is an urgent need to establish a large-scale systematical knowledge base for Mongolian information processing. How did the knowledge of Mongolian language serve the Mongolian information processing systems in different phases? What kinds of problems are left for us to resolve? These can be discussed in three steps.

2. Graphitic knowledge and character processing

It is well known that the Mongolian writing system is quite different from the Western writing systems as well as the Chinese writing system. First, Mongolian characters are written vertically and a new line is entered from left to right. That is obviously not compatible with general operating systems and has always been a headache for the Mongolian system developers. So Mongolian systems have been implemented on operating systems, like WINDOWS, with a technique called “hanging outside”. Second, Mongolian characters are written in succession, meaning that depending on where the letter is placed in a word, it may have different forms. There are at least three different forms for each letter in Mongolian, some having up a dozen different forms. These forms are namely isolate, initial, middle, and final form. All of these are decided by their positions in a word. This is particularly one of the hardest problems in Mongolian character processing, and has been a bottleneck in developing input methods and in the writing of the international and national standards of a Mongolian code.

2.1 Inputting

When we first began to in-put Mongolian characters in the 1980s, we were confronted with the question about whether character pronunciation in-put or character shape in-put was more advisable. As mentioned above, each Mongolian letter has at least three forms (except “NG”) in different positions (initial, middle and final) of a word. Indeed, most of the characters have more than three forms, which means at least 82 keys would be needed if each form was arranged by a key. However, we need only 35 keys if using character pronunciation in-put. In fact, character shape in-put methods were also developed at that time, and it is evident that character pronunciation in-put is more suitable.

With that we come to a new problem of how to settle the different forms of a letter with only one key. Although Mongolian characters are complicated with dozens of different variants, there still exist some rules to be followed: the current letter’s position in a word or a syllable; the position of the syllable which contains the current letter in a word; whether it is a second root or not; the masculinity and femininity of the vowels included in a word etc. These rules can be a knowledge resource to settle the different variants of a letter with only one key.

For example, key “A” is arranged for Mongolian letter “мон”. We realize the output of the independent form “мон” by inputting the string of  
and the initial form “мон” by  
, and the middle form “мон” by  
, and the final form “мон” by  
 respectively. Another example is, key “G” can be used to output 9 different forms of “мон”. Each certain form of “G” is decided by whether there is “A, 0, V” or “E, O, I, e” in the appointed character string, that is, the initial form of masculine “мон” is produced by inputting the string of  
(A/0/V) and the initial

\(^1\) denotes a space.
\(^2\) *denotes a random Mongolian character
form of feminine “♀” by Ⱪ(E/O/U)*, etc.

Therefore, the graphitic knowledge, especially the orthography knowledge, is a basis for data input. To a great extent, the “Mongolian full word input method”, which was developed in the 1990s, did not only make use of the Mongolian graphitic knowledge, but also of its grammatical knowledge.

2.2 Code proposal

Experts from China and Mongolia began writing the International Standard Proposal of Mongolian Code in 1994. In 2000, the proposal passed the accreditation of the ISO/IEC JTC1/SC2 and was officially adopted in the ISO/IEC 10646.1-2000. Actually, we encountered many problems when writing the proposal. According to the prescripts of ISO/IEC JTC1/SC2/WG2, “Basic Character” is the only one that can be encoded among all the variants of a letter while the “Presentation Forms” will not be encoded. The “Presentation Forms” have to be visualized by the encoded form and we need a series of rules to fulfill the visualization. After six years of hard work, we wrote a set of detailed rules of transforming the Basic Characters into the Presentation Forms. A variant can be decided in most cases according to its context (or its position in a word). However, we defined and set up some Control Characters to meet the situation when the fitness of a variant cannot be decided by its context only. For instance, the initial, middle and the final forms of Mongolian letter “♀” can be easily distinguished according to the rules. The problem is that the final form of Mongolian “♀” has two different sub forms, namely attached form “♀” and detached form “♀”, therefore both rules and Control Characters have to be used together to decide which sub form is more suitable in a certain situation. All in all, Mongolian graphitic knowledge, especially Mongolian orthography knowledge, plays an important role in Mongolian encoding.

We have compiled this knowledge as an “information list” and attached it to the International Standards of Mongolian Code. At the same time, the knowledge is stored in the form of a “field and field value” in the “Mongolian Information Grammatical Dictionary”.

3 Mongolian grammatical knowledge and word processing

Mongolian grammar researchers have done a lot of creative work in discovering and describing Mongolian language laws. They have found that nouns are broken into grammatical categories of numbers, case and reflexive genitives, and that verbs are broken into categories of voice, aspect, modality, person, adjective verbs, converbs, etc. However, not all nouns have number categories. Which nouns have number categories and what are their concrete plural forms? And why are they in those forms? To date, there is no single book available which can explain the above questions reasonably and systematically.

Let’s take a look at some examples here. Example 1: “0D0 (star)” and “AGAR (air)” all belong to the same noun class, naming things in nature. The former has a plural form of “0D0D (stars)” while the latter does not have a plural form at all. Example 2: “SVRVGCI (student)” and “BAGSI (teacher)” are all numerable appellative nouns in Mongolian and they all have number categories. But their plural forms are different from each other. Specifically, the plural form of “SVRVGCI (student)” is “SVRVGCID” while the plural form of “BAGSI (teacher)” is “BAGSI NAR”. Next, let’s take verbs as an example. Verbs have categories of voice and aspect as a whole, though things might be different when we focus on a certain verb. We can easily find the following forms of verbs in Mongolian like “C0HIGDAJAI (be hit)” or “BAYISHIGED (after a while)” or “C0HILDVHV (fight each other)” or “INIYELCEHU (laugh, especially to say that lots of people laugh together)”, but it is not proper to say “VNTAGDAJAI (be slept)” or “OTElushiged (turn old for a while)” or “SAGVLDVHV (sit each other)” or “HOGJILCEHU (develop together)” in Mongolian language. So we must have a knowledge-based system that includes the knowledge itself and its formalization, describing the grammatical attributes of each word accurately and totally, or we cannot parse and generate Mongolian words and sentences correctly.

Mongolian grammar research has a long history, yet it is still far from meeting the needs of natural language processing. The basic research lags behind in two aspects: it does not go far enough
to meet the needs of natural language processing; and the research angles and the description methods are not favorable for natural language processing. The things we have done, such as, corpus processing, machine translation, etc., and the things we plan to do in the future indicate that we are in urgent need of a multipurpose Mongolian knowledge base on a large scale. So, we are now compiling a “Mongolian Grammatical Information Dictionary” (electronic version) with information on pronunciation, structure, parts of speech, orthography, morphological forms, syntactic function, semantic class and collocation for each word.

The “Mongolian Grammatical Information Dictionary” is compiled for the automatic parsing and generating of the Mongolian language. Comprehensive and detailed grammatical knowledge is available for all kinds of applied systems including morphological forms, syntactic function, collocation features, orthographical rules, etc. At present, the framework of the dictionary is almost complete. A “general bank” of 25,800 words has been established, and the “verb bank” will soon be finished. Some of the attribute fields of the “general bank” have been set in the “Chinese-Mongolian MT system-Darhan system” (Supported by National 863 Program 863-306-ZT04-05-3 ; 2003 AA115510) and they have become an important knowledge-based resource for the Mongolian language model, generating rules and bilingual corpus processing.

On the other hand, developing a deep study of Mongolian phrase types, phrase features, and the relationships between phrases and compound words and between phrases and sentences is inevitable because it relates a lot to the development of the Mongolian MT systems and the process of Mongolian corpora. The types of Mongolian phrases (like np, vp, ap, mp, rp, dp,......etc.), the structure of the each type (np→n n, np→a n, np→m n, np→t n, np→v n……etc.) and the syntactic features of the different types of phrases such as, attribute head, adverbial head, statement complement, subject predication etc. are indispensable pieces of information to the Mongolian language processing systems, because the phrase rules are the most basic knowledge resources for parsing and generating Mongolian sentences. Unfortunately, there are no sources available at this point to elaborate on the above. The criteria for segmenting and tagging Mongolian phrases are limited. So we have attached more importance to phrase research from recent years and preliminary studies are already being made on types and structures of Mongolian phrases. We are also doing some phrase tagging experiments on the Mongolian corpus. We realized after years of work that Mongolian syntactic knowledge is significant in the research of phrase rules and tagging criteria. Establishing a Mongolian knowledge base will be an essential way to thoroughly resolve these problems. The purpose of the task is to build up a Mongolian Tree Bank.

4 Semantics knowledge and sentence processing

Some of the Mongolian grammar phenomena have virtually semantic reasons behind them. For example, “VLAGAN (red)” and “JEGERDE (this is a special word in Mongolian and it is only used to describe the color of a horse, something like jujube red-; chestnut )” are both adjectives, though, “VLAGAN” can have a comparative form of “VB=VLAGAN (very red, reddest)”, but “JEGERDE” can never have a comparative form of “JEB=JEGERDE”; “YEHE (big, large...)” and “T0M0 (large, huge...)” are both qualitative adjectives representing “degree”. We can say “YEHE IDEHU (eat a lot)” or “T0M0 NIDU (big eyes)”, but we can never say “YEHE NIDU (large eyes)” or “T0M0 IDEHU (eat big)”. We have to depend on semantics since the phenomena cannot solely be explained from the grammar angle. Mongolian semantic research today cannot offer us enough information to explain the linguistic case. Formalizing and describing the case is next to impossible at present, but these problems have to be solved to implement the automatic parsing and generating of the Mongolian language. Therefore, we can see that semantic research plays an important role in Mongolian information processing. We are now developing research on Mongolian semantics which is information processing oriented, and we aim to establish a Mongolian semantic knowledge base with rich information, including Mongolian semantic fields, the semantic classes and the tag sets, the collocation attributes, the valence quantity and quality (of verbs, nouns and adjectives), sememe analysis and its annotating system etc.

Correspondingly, semantic units exist with linguistic units, namely morphemes have morpheme
meanings, words have word meanings, phrases have phrase meanings and sentences have sentence meanings. The semantic information dictionary describes only word meanings, but the meanings of the bigger units (like phrases or sentences) or the smaller ones (like morphemes) have to be described too. The laws of how the surface structure and the deep structure of languages correspond with each other can not be clearly denoted unless the semantic structures and the collocation features are exactly described.

I am taking two pairs of Mongolian sentences as examples to clarify the point.

(1) “H0NI HARIGVLBA (to graze sheep)” and “H0NI HVRAGALABA (A sheep lambs)”. The two Mongolian sentences seem to have the same surface structures: s -> np (nominative form) vp (indicative mood, the past tense). But in fact the structure of the former sentence is vp -> np (accusative form) vp (indicative mood, the past tense). The structure of the latter sentence is s -> np (nominative form) vp (indicative mood, the past tense). To disambiguate this kind of ambiguous structures, the semantic attributes of the two verbs should be retrieved from the semantic knowledge base: HARIGVL (to graze): [valence quantity: 2; semantic class: human action; agent: humankind; object: domestic animals]; HVRAGALA (to lamb): [valence quantity: 1; semantic class: animal action; agent: sheep]. So, this information in the knowledge base can offer a judgment that the two sentences have two different syntactic and semantic structures.

(2) “BAGSI INIYEBE (The teacher laughed.)” and “BAGSI MAN-I INIYELGEBE (The teacher made us laugh.).” The surface structures of these two sentences are of little difference: s -> np (nominative form) vp (indicative mood, the past tense). But the vp in the first sentence is composed of only one vp while the vp in the second sentence is composed of a np and a vp which is a “predicate object” structure. The word “BAGSI (teacher)” plays the role of subject in both sentences. However, the agent of the first “INIYE (laugh)” is “BAGSI (teacher)” when the agent of the second “INIYE (laugh)” is MAN (we). This kind of deep semantic structure is realized by the meaning of the verb suffix “LGE” (causative voice). So we can see that Mongolian causative voice can not only transform the intransitive verbs into transitive verbs but also increase the valence quantity of the same verb. (For example: MARTAHV (forget) is a two-valence verb and MARTAGVLHV (make forget) is a three-valence verb.) At the same time, Mongolian voice category can also change the semantic role of a component. Therefore, we should not only describe in detail the semantic attributes of the word stems, but also give a good description of the semantic attributes of the suffixes and offer all this information to the system developers. The syntactic and semantic information of the different Mongolian linguistic units are essentially important for establishing a semantic framework of Mongolian sentences. It is profoundly significant to make research on Mongolian valences and semantic cases for natural language parsing.

Generally speaking, the establishment of the Mongolian knowledge base, or the exploitation and application of the Mongolian linguistic resource is a significantly important field, for it is indispensable to Mongolian language research, especially in Mongolian information processing. If we attached importance to only system development and industrialization and neglected the resource exploitation and basic research, we would certainly lose something in the long term.

Therefore, I hope that the relevant departments, research institutes and enterprises make good plans and allocate budgets for the cause of Minority language information processing, so that the work in this field can go on steadily and healthily in this new era and environment.
References:

俞士汶 朱学峰: 关于汉语信息处理的认识及其研究方略，语言文字应用，2002 年第二期，p51-58
Nasun-urtu: Mongolian knowledge base and Mongolian information processing，The 17th International conference KOREA and MONGOLIA, Seoul 2004, p51-58
S. 苏雅拉图: 蒙古文整词编码研究，中文信息学报，2001 年第二期，p57-64
确精扎布: 蒙古文编码，内蒙古大学出版社，2000 年，呼和浩特
那顺乌日图 确精扎布: 蒙古文国际标准编码的构成原则，内蒙古大学学报 1997 第六期，p1-7
那顺乌日图 确精扎布: 蒙古文国际标准编码诸规则，内蒙古大学学报 1998 年第四期，p77-84
那顺乌日图: “蒙古语语法信息词典”的开发与应用，Workshop on Mongolic Languages, Taiwan, 2002.9,p53-60
那顺乌日图：关于面向信息处理的蒙古语语义研究，Proceeding of the 5th Seoul International Altaistic Conference – Information processing in Altaic Languages –，Seoul National University, 2002.10,p43-51