THE ELECTRONIC DATA AND RETRIEVAL OF THE *SECRET HISTORY OF THE MONGOLS*

Di Jiang

Humanities and Communications College, Shanghai Normal University, Shanghai 200234, China
Lab. of Phonetics and Computational Linguistics, Institute of Ethnology & Anthropology,
CASS, Beijing, 100081, China
Email: jiangdi@cass.org.cn

ABSTRACT

This paper discusses the principle of electronic data and retrieval methods for the Secret History of the Mongols, which is a great classical historical work written in the 13th century with Chinese characters transliterated from Mongol. This handwritten work contains rather rich text information, which should be the contents of forming an electronic database. There are in the original book multi-types of information, including layouts, volumes, chapters, characters, interlinear translation, segments, and Chinese translation, each format of which has been approached in detail and divided separately with markers. On the basis of analysis, our project builds up a complete electronic retrieval system for this great book, which resolves the return to the original shape of the archaic handwriting form with three lines representing one content. The sorting methods of the system are also designed according to the original text formats, namely concordance technology, which can print out retrieved objects with their contexts, retrieve with statistical data, and freely browse search.

Keywords: Computer application, The *Secret History of the Mongols (SHM)*, Electronic text data, Data format, Retrieval system

1 OUTLINES

The *Secret History of the Mongols* (*SHM*) is a great classical historical document, which records the origins of Mongoloid nationality and the conquering history of Genghis Khan, the conqueror of Central Asia and southern Russia and the founder and pioneer of the Yuan Dynasty, with his army. The original document of *SHM* was in ancient Uighur letters or characters written in the middle of the 13th century. However, the original Uighur document was lost in history. Now people only can find a version of *SHM* in Chinese-transliterated characters, which was preserved to serve the Ming Dynasty's political, military, and diplomatic purposes. The Chinese *SHM* takes the name 忙豁仑·纽察·脱卜察安 (Monqolun Nihuča Tobčiyan), namely 'The Secret History of the Mongols'.

The *SHM* is the most important book about the society, politics, war, and social conventions of Mongolian history at that time. It is also a literary work, which portrays many typical characters of grassland ethnics with poetic texts. The *SHM* wins universal praise because it is a folk epic of the Mongolian nationality. From the Qing Dynasty, the *SHM* became a research focus in academic fields. The study of its contents includes the origins and development of its versions, historical events, the places and their people, textual research of

languages and lexemes, transliteration, and translation. According to statistics, the research achievements of SHM in the world add up to hundreds and thousands, and the translations are published in English, French, German, Japanese, Chinese, Hungarian, Russian, Polish, Czech, Turkish, Spanish, and other languages. The decision of UNESCO (United Nations Educational, Scientific, and Cultural Organization) points out that the SHM holds a lofty status in the world cultural history, and the *Secret History of the Mongols* could be considered not only as the remarkable masterpiece of the Mongolian literature but also as an outstanding literary monument of world significance (at the celebration of the 750th anniversary of the *Secret History of the Mongols*). So far, the *SHM* has become a learning domain in academic researches of the global world: the *SHM*-ology.

In the information society nowadays, the *SHM*, as a cultural heritage and an eternal classic, needs further research and understanding. The many puzzles, such as who wrote the book, are necessary to discover to clarify records. In this thesis we talk about making a full electronic version of *SHM*, which can help *SHM* experts perform deep research with the original book of Chinese characters. The electronic version is built on the basis of SibuCongkan (四部丛刊), a classical Chinese collection.

2 THE TEXT FEATURES OF *SHM* AND THE PRINCIPLES OF ELECTRONIZATION

The original information of handwritten *SHM* is rich in its contents and format. Its contents involve history, geography, religion, military, ethnics, and social life, and its format relates to versions, languages, grammatical phenomena and vocabularies, Chinese translation, transliteration, orthography, and so on. The most important principle for creating an electronic version is to keep all the information of the original book, including layouts, volumes, chapters, pages, characters, segmentation, interlinear translation, and Chinese translation.

Information about layouts: The so-called layouts mean that the original shape of the archaic handwriting form is one content represented by three lines of characters, which is a special complicated text with vertical direction of handwriting lines (Figure 1). The middle line is Chinese- transliterated characters from Mongol. The right (the first) line is side-for-side Chinese translation and grammatical annotation. The left (the third) line is initials representing initial consonants in pronunciations of the aligned characters. In addition, the small characters after Chinese- transliterated characters within the middle lines are endings representing pronunciation of final sounds. For example, “成吉思^{中}合罕” is a string of transliterated characters, the interlinear “太祖”(the first founder of a dynasty) is a side-for-side translation for “成吉思”(Genghis), and “皇帝”(emperor) interlinearizes “合罕”. And the interlinear character “^{中}” annotates the pronunciation of the initial consonant of character “合”. The following chart changes the original vertical lines to the lines sideways.

	中						-initials
成吉思	合罕	田迭	泰亦赤兀	的	倒兀里周	泰亦赤兀台	-transliteration
太祖	皇帝	那裏	種	行	擄着	種	-alignment

Let us have a look at the original handwritten format with a photograph of one page of the book from volume 5.

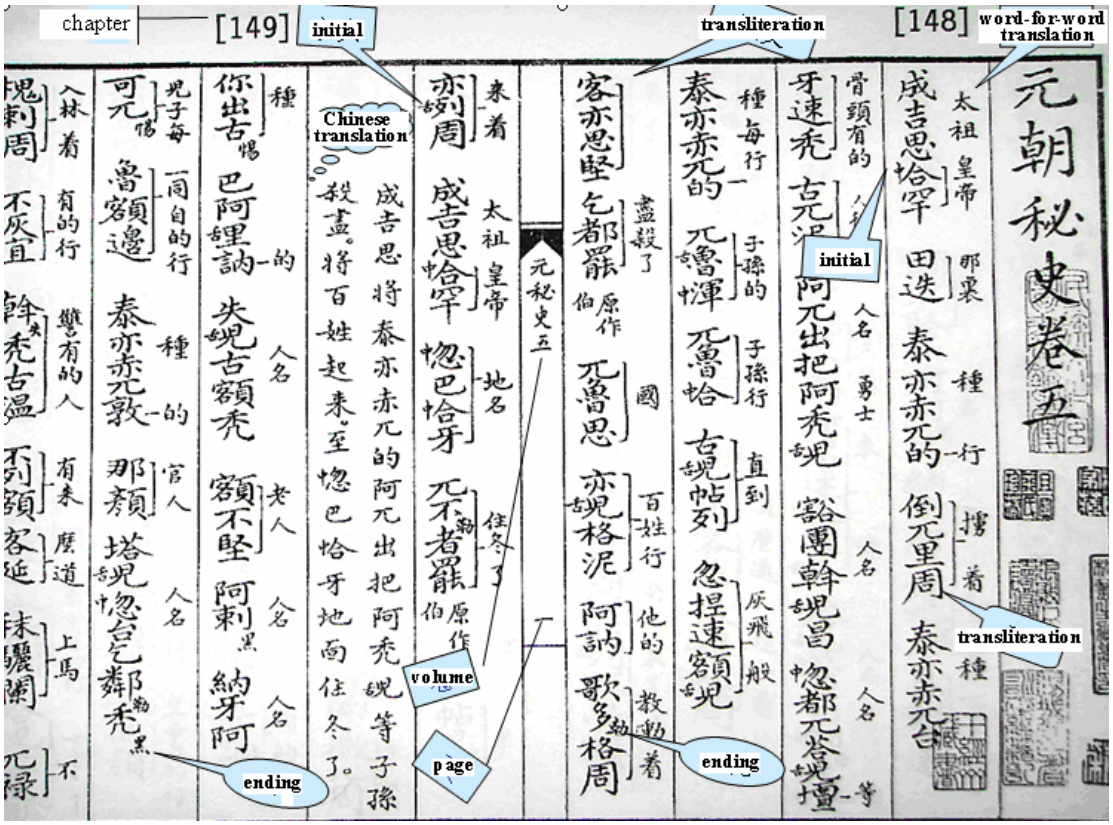

Figure 1. The original format of *SHM* (with English annotations)

Information about volumes: The original work is a traditional thread-bound Chinese book with titles and markers of volumes and pages in the middle of the layouts, which say "元秘史＊" or "元秘史卷＊" ("元秘史" means *SHM*; "＊" indicates number of volumes). Volume 11 and 12 show "元秘史续一" and "元秘史续二" ("续一, 二" =expanded volumes 1 and 2). Under the volume numbers are page numbers, which keep separate sequence within each volume. The number of pages of each volume is about the same. Volume 12 has a few more, 58 pages or so, and volume 10 has the fewest, 48 pages.

Information about chapters: The original numbers of chapters are at the top of the pages. Actually the division of chapters can be divided by text content. Namely, after Chinese- transliterated paragraphs appear the Chinese translation paragraphs appear; a chapter consists of both of them together. Altogether, there are 282 chapters in the whole book.

Information about characters: In order to transliterate Mongol with Chinese characters accurately, the transliterators made use of assisting symbols to represent the pronunciation of Mongol. The initials may help read the consonants of characters, and the endings may help read the finals of characters. Therefore, a complete character may consist of a transliterated Character, an initial, and an ending. As a result, there are four types of Characters: type C, which is a single character, such as "安", "客"; type xC, which consists of one initial and one normal character, such as "舌刺", "中豁"; type Cy, which is a normal character with an ending, such as "阿勒", "迭克"; and type xCy, which takes normal characters with both initials and endings, such as "舌魯黑", "中忽勒". According to the statistics, the number of each type is as follows.

Table 1. The statistics for four types of transliterated characters

	figures	rate	tokens	frequency		figures	rate	tokens	frequency
C	510	53.0146	68810	76.0693	Cy	317	32.9522	6114	6.7590
xC	83	8.6279	14477	16.0043	xCy	52	5.4054	1056	1.1674

Another aspect of characters are their graphic forms, which refer to traditional Chinese, simplified Chinese, and variant forms of Chinese characters. Take the following characters as examples. “趖” and “躲”, “備” and “俻”, “槍” and “鎗”, “鄰” and “隣”, “幾” and “几”, “讎” and “讐” or “仇”, “桑” and “桒”, “迆” and “逃”.

Information about interlinear translation: The interlinear translation of *SHM* can align with strings of transliterated characters, which is important so that people can compile a Mongol-Chinese dictionary of the 13th century from the text. In addition, the lines of interlinear translation include much grammatical information, such as “每” represents plural category, “自的行” represents one kind of objective case, “有來” shows the perfect aspect, etc.

Information about translation: *SHM*, as a complete historical and literary work, is not only a great Mongol book, but also an important Chinese document. The Chinese translation contains much cultural and historical information as well as information about the grammar and vocabulary of middle ancient times.

So far we have discussed only a little about the information in *SHM*. Here, we limited ourselves to the format and formal information, so that we can make an electronic version and a retrieval system for *SHM*.

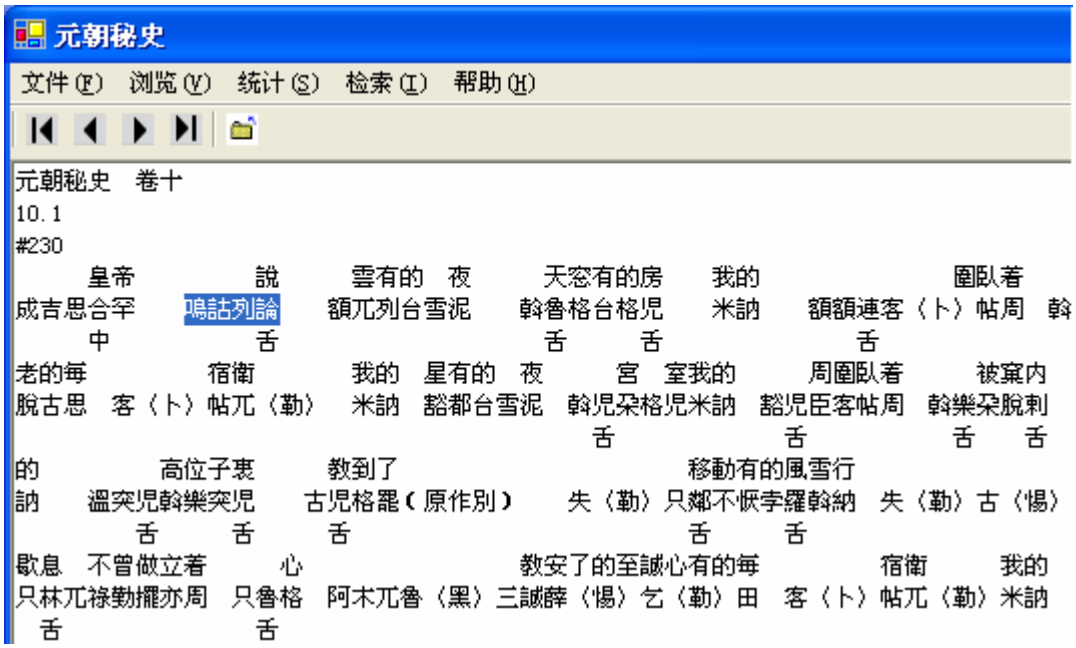

Figure 2. A piece of the electronic retrieval system of *SHM*

The electronic version of *SHM* keeps all the basic features and retains all of the original information. We designed an aligning format between transliteration and side-for-side translation, continued to have the types of characters with initials or/and endings, preserved all the simplified and traditional Chinese characters as well as

variant characters. As for information about volumes, pages, segments, chapters, and punctuations, all work as in the original book. For the electronic version, we made some revisions. Although we put all the volumes together into one system, users can directly select any volume wanted and operate within just that one volume. Within one volume, users do not need to turn pages, just roll the window bar, which will let them find the distinctive contents divided into pages. The most prominent revision is the reading direction of the characters, which changes to horizontal lines from vertical lines. Figure 2 is an example of the electronic data.

3 THE ELECTRONIC RETRIEVAL SYSTEM OF *SHM*

Based on the idea of modern corpus linguistics, the *SHM* is a complicated text, with many inlaid layers and sections. It is necessary to differentiate "sections" and to add markers between them for electronic processing. The main sections are sections of transliterated characters, of translation, and of interlinear translation. The other sections are sections of initials, of endings, of volumes, of pages, and of original notes. The sections of transliteration are the main body of the electronic project. The sections of translation refer to character strings after transliteration, which are different from transliteration for containing no interlinear characters. The sections of side-for-side translation are those strings, which interlinearize transliterated characters. When we digitalized the work, we added some markers within the text. The marker "/…/" is used for translation strings, "[]" is used for side-for-side translation strings, "{}" is used for initial strings, "<>" is used for ending strings, and "()" is used for strings of original notes, as shown in Figure 3.

帖迭[那]　亦児格[百姓]　泥[行]　阿合納児[哥哥每]　迭兀捏児[兄弟每]　塔不兀剌[五箇]　倒

兀里周[擄着]　阿都温[馬羣]　亦啞額捏[茶飯處]　哈闌{舌}[人口]　秃<惕>合{中}剌[使喚行]

Figure 3. Markers added to text

With the classification of sections, it is convenient to design different retrieval methods for different sections. While processing sections of transliteration, the alignment of transliterated characters and interlinear characters is difficult, as is the appropriate processing of initials and endings. We propose a comparison algorithm for these strings. After we get the length of a transliterated character string, we compare it with the length of an extracted interlinear string. If the transliterated string is longer than the interlinear string, we fix the position of the interlinear string at the location of the last character position of the transliterated string. If the interlinear string is longer than a transliterated string and clashes with previous strings, both strings are shifted rearward synchronously. With the realization of the section processing and positioning techniques, there is no doubt that we can determine the format of integrated content with three lines using the digitalized system there. For sections of translation, we can use the same method as for sections of transliterated character strings.

To retrieve the text of *SHM*, we have designed three retrieval techniques. They are: concordance, browsing, and statistical retrieval. Each method processes separately distinctive sections.

The method of concordance processes strings of transliterated characters. When the retrieval results are shown, their contexts need to be shown together. Figure 4 gives an example of retrieval string "必孫." Users may look at the previous character strings and the following character strings with interlinear characters and initials and endings.

Figure 4. An example of concordance

The method of concordance may be applied to strings of translation as well. Figure 5 shows the example of word "野獸" with ten characters in front of the word and ten characters after the word. Notice that the other accompanying information is numbers of volumes, pages, chapters, and paragraphs.

Figure 5. An example of translation strings

The method of statistic retrieval is designed for interlinear translation. Users can search for peoples' names, names of places, or other words and find which volumes or which chapters they appear in and how many times they occur. Figure 6 gives an example for the word "祭祀"(offer sacrifices to ancestors).

The method of browsing retrieval is suitable for any kind of sections or strings and finds objects as shown in a highlighted form, as in the example "鳴詁列論" in Figure 2.

Figure 6. An example of interlinear translation

4 Summary

The *SHM* was written in the remote past of the 13th century. It contains plenty of information with innumerable mysteries remaining under cover. We believe that the electronic version of the *SHM* will help people to explore the work more deeply and from any direction. Take the character statistics for instance; the number of all transliterated characters in the book is 540; the number of interlinear characters is 1,567, and the number of translated characters is 1,669. Removing repeated characters, the total number of characters in the book is 2,099, and the tokens of the characters in the whole book are 224,542. It is important for people to comprehend precisely the work with the figures and the information obtained from the electronic version. It is believed that the achievement of the electronic version of the *SHM* will carry the work a great step forward.

5 REFERENCES

Cleaves, F. W. (1982) *The Secret History of the Mongols*. London: Cambridge Massachusetts.

Eldengtei, etc. (1980) *A collated edition of the Secret History of the Mongols*. Huhehot: Inner Mongol People Press

Jiang, Di, Yan, Hailin, etc. (2006) Approach to the Retrieval System of the Secret History of the Mongols. *Journal of Chinese information processing 20*(3), 36-42.

Jiang, Di, Zhou, Xuewen. (2006) Full Inspection on Chinese Characters Transliterated from Secrete History of Mongols, In: Tingting He, Maosun Sun, Qunxiu Chen, (Eds.), *The 20th Pacific Asia Conference on Language, Information and Computation: Proceedings of the Conference,* pp 49-55. Beijing: Tsinghua University Press

Ozawa,S. (1993) *The course of Mongolian Grammar of the Secret History of the Mongols*. Tokyo:

Rachewiltz, I. (1972) *Index To The Secret History of the Mongols*. Bloomington

Shiro H. (1946) *A Study on Chinese-transliterated Characters from Mongolian in SHM*.