A Full Inspection on Chinese Characters Used in the Secrete History of the Mongols*

Di Jiang Xuewen Zhou

Lab. of Phonetics and Computational Linguistics, Institute of Ethnology & Anthropology,
Chinese Academy of Social Sciences, Beijing, 100081
jiangdi@cass.org.cn

Abstract. The Secrete History of the Mongols (SHM) is a special Mongolian historical document transliterated with Chinese characters as phonetic symbols, which feature causes many multivariate analyses in later ages. The discussion of this paper mainly focuses on the data of used Chinese characters, the frequencies, and the rules of transliteration on SHM. All the statistic data and the analysis are based on the electronic text of SHM, which provides those important data in an all-round way, including the whole Chinese-transliterated characters, aligned Chinese glosses by the side of Chinese-transliterated characters, and Chinese translational paragraphs. Furthermore, four types of Chinese characters as phonetic symbols and their statistic information have been discussed, which are type C, xC, Cy, and xCy as in the main body of the text: 安, ^{中}豁, 阿$_{勒}$, and ^{舌}魯$_{黑}$. Shortly, this statistics is so far the most completed inspection on the text of the Secrete History of the Mongols.

Keywords. the Secrete History of the Mongols, Chinese-transliterated characters from Mongolian, types of Characters, Character statistics

1 An Overview of the Researches on Chinese-Transliterated Characters

Almost every researcher who studies The Secrete History of the Mongols (SHM), whatever his researching interests are, pays more or less attentions on Chinese-transliterated characters. Yet till now, the exposition specially written for Chinese-transliterated characters as phonetic symbols are only papers of Mr. Chen Yuan's, namely *An Inspection on Chinese Characters Used in SHM* [1], which was published seventy years ago, and of Shiro Hattori's *A Study on Chinese-transliterated Characters from Mongolian in SHM* [2], which was published in 1946.

In terms of the original text of SHM prior to the appearance of the version of Chinese-transliterated characters, the well-known viewpoint is that the original text of SHM is in Uighur letters or characters. The SHM of Chinese-transliterated character is at that times a cultural activity serving Ming Dynasty's political, military and diplomatic purposes. The transliteration with Chinese characters definitely causes problems such as accuracy, pronunciation harmony and context consistency. Furthermore, it is difficult to avoid multi-errors such as copies of handwriting with wrongly written words, or missed-out words, because of the procedures of manual copies in ancient times.

So far, there are neither an identical view on the classification of Chinese-transliterated characters, nor an accurate statistics and frequencies of Chinese characters and symbols in SHM. This paper tries to conduct comprehensive statistics and analysis of Chinese-transliterated characters on the bases of an electronic version of SHM and presents accurate character data of SHM [3].

The discussing object is based on the version from SibuCongkan (四部丛刊), a classical Chinese collection. Before discussion, we regulate some terminologies. The characters in main body of the text are called Chinese-transliterated characters from original Mongolian. The small characters in the left side of Chinese-transliterated characters represent different consonant pronunciation, called word initials. The small characters after Chinese-transliterated characters represent consonant differences in

* This research is supported by the CASS Funds to Key Laboratory (phonetics and computational linguistics, MZ-101).

word endings, called word endings. It is necessary to point out that the word initials and the word endings do not really stand for Mongolian sounds or syllables, in stead, they are supplementary symbols for transliterating Mongolian letters to Chinese characters. Therefore, they will be treated differently while conducting statistics.

2 The Classification & its Principles of Chinese-Transliterated Characters

Firstly, let's focus on the form types of the Chinese-transliterated characters in SHM. There are two studies on the types of Chinese-transliterated characters between Shuangfu and Shiro Hattori, which show much differences from their viewpoints. Shuangfu lists altogether 502 Chinese-transliterated characters [4], which are characters without word initials. Such as "豁", "羅", "驪". In Shiro Hattori's study, however, they may be "豁", "羅" or "中豁", "舌羅" etc.. Hattori Shiro lists 43 characters with word initials "舌", 22 characters with initials "中" in his table of 563 items. For instance, "洏", "舌洏"; "翁", "中翁". Apparently, Shuangfu's classification is simpler. Only characters in main body of the text were embedded into his word table, without distinguishing character with or without supplementary symbols. In addition, Shuangfu also excluded those characters which are used as auxiliary symbols and never appear as independent characters, namely "舌", "惕", "泐", "黑".

Table 1. Some difference examples between the two character lists

A	B	A	B	A	B	A	B	A	B	A	B
中崑	崑	中孩	孩	舌廉	廉	舌驪	驪	舌路	路	舌瀧	瀧
中晃	晃	中海	海	舌怜	怜	舌林	林	舌瞳	瞳	中宦烏	宦烏
中荒	荒	中鶤	鶤	舌澧	澧	舌驎	驎	舌論	論	舌羊歷	羊歷
中灰	灰	中康	康	舌離	離	舌騾	騾	舌輪	輪	舌馬羅	馬羅

Note: A stands for Shiro Hattori'list; B stands for Shuangfu's list

From the classification approach, we know that Shiro Hattori's standpoint of classification is clear and definite, because he regards Chinese characters as phonetic symbols, and characters with word initials as a whole. But it is obscure that he lists word initial symbols or word endings as characters into his table, such as "泐", "黑". In fact, these characters never appear as independent Chinese-transliterated characters in the main body of the text.

Secondly, there are some differences of character numbers and character forms between the two researchers' lists. Please check table 1.

The number in the list of Shuangfu is smaller, in which there are no "吃", "洏", "峒", "濟", "思", "兒", "盾", "翼", "閊" and so on. Some difference may be caused by different character font-types among simplified Chinese and traditional Chinese. For example, "翌" may be corresponding to Shiro Hattori's "翊". On the contrary, there are no characters "闐", "𤰞", "馬翼"in Shiro Hattori's list.

It should be pointed out that both of the two researchers mark characters with phonetic symbols. So character "洪" appears twice in Shiro Hattori's list, one is "qung", and another is "hung".

In terms of transliteration principle, the list of Shiro Hattori is far from perfect. Because he does not attach word endings to characters of the text main body, and characters with word initials are not the whole of them as well. For example, "撒惕" and "中忽勒". The researcher here to be mentioned of is Ozawa Shigeo, who gives a perfect process to all the characters in his work [5]. In his work Ozawa Shigeo compiles a dictionary, in which all the characters are in the format of "word initial + character + word endings". For example, "都舌剌惕中罕", "中忽卜察速", "客舌列勒", "尕脱黑石", etc..

In this paper all the characters in the text main body are put into four types, and the statistics will be collected upon this principle. The four character types are as follows.

1) Type C: Single characters. E g.“安”, “客”,

2) Type xC: Characters with initials. E g. “舌刺”, “中豁”,

3) Type Cy: Characters with endings. E.g. “阿勒”, “迭克”,

4) Type xCy: Characters with initials and endings. E.g. “舌魯黑”, “中忽勒”

3 The Number and Frequency of Chinese-Transliterated Characters

According to the statistics on the original text, there are only 540 items of Chinese-transliterated characters in the main body part of the whole book, and they appear for 90,457 times. Now read the details in Table 2.

Table 2. Statistics of Chinese-transliterated characters, word initials and endings

Character forms	Number	Token	Proportion	Token rate
Chinese-trans. char.	546*	113160*	100.00	100.00
Single words	540	90457	96.9479	79.9373
Word initials	2	15533	0.3591	13.7266
Word endings	15**	7170	2.6929	6.3362
Character Types	Number	Token	Proportion	Token rate
Chinese-trans. char.	962	90457	100.00	100.00
Type C	510	68810	53.0146	76.0693
Type xC	83	14477	8.6279	16.0043
Type Cy	317	6114	32.9522	6.7590
Type xCy	52	1056	5.4054	1.1674

Notes: * if initials and endings are not treated as Chinese-transliterated characters, the number is 577. ** The 6 endings “黑, 防, 汃, 溯, 室, 惕” never appear in the main body of the text.

With regard to Chinese-transliterated characters, the numbers of their types are 540 and they appear 90,457 times in the main body of the text. On average, every character appears 167.5129 times. However, the token for each character is quite different. Such as, there are 124 characters, which only appear once. In addition, 56 characters appear twice and 21 characters appear three times among them. Although the 201 characters take 37.22% amongst all types of characters, yet their token is only 0.33%. The character at the highest token takes 4,826 times, which takes rate of 5.33% in frequency. Another case, which should be considered of, is that there are 27 characters, each of which appears more than 1000 times, and each proportion is 5.00%. Furthermore, there are 51 characters, each of which appears more than 500 times, and the proportion for each one is 9.44%. In addition, the occurrence token for the two types are separately 56.46% and 74.67%. See table 3.

Table 3. The first 20 Chinese-transliterated characters in frequency

Chr.	Token	Rate	Chr.	Token	Rate	Chr.	Token	Rate	Chr.	Token	Rate
児	4826	5.3351	周	2395	2.6476	里	1642	1.8152	格	1317	1.4559
兀	4263	4.7127	列	2355	2.6034	斡	1525	1.6858	訥	1305	1.4426
額	3567	3.9433	亦	1995	2.2054	忽	1447	1.5996	迭	1291	1.4271
合	3223	3.5630	客	1985	2.1944	苔	1409	1.5576	帖	1277	1.4117
阿	2912	3.2192	剌	1813	2.0042	赤	1321	1.4603	禿	1230	1.3597

It is sure that the characters in high frequencies are mainly related with Mongolian phonetics, and with the contents of the text. For example, character “周” is usually used as a grammatical participle, meaning progressive aspect.

About the types of Chinese-transliterated character, as we mentioned above, they are categorized into four types. Because there are only two types of word initials, one of which is “舌”(8,896 times), another “中”(6,637 times), the total types of characters with word initials are very limited. The number of xC type is 83 items, accounting for 8.6278% of all types. However, the frequencies of these types are far more than types of characters with word endings. Type xC appear 14,477 times, accounting for 16.0043% from all four types. Otherwise, there are 317 items in type Cy, accounting for 32.9521% of the whole types. Yet the occurrency token is only 6,114 times, it only takes the rate of 6.7590%. The number of type xCy is only 52 items, and appears 1,056 times, accounting for 5.4054% of all types, taking the rate of 1.1674%. As to type C, the 510 items, accounting for 53.0145% of all types, appear 68,810 times, taking the rate of 76.0693%. Pay attention to the following data.

Table 4. The first 20 Chinese-transliterated characters for four types in frequency

	C	token	rate	xC	token	rate	Cy	token	rate	xCy	token	rate
1	兀	3429	3.79	舌児	3707	4.09	兀勒	498	0.55	中合惕	147	0.16
2	額	3072	3.39	中合	2816	3.11	孛勒	388	0.43	中合勒	116	0.13
3	周	2395	2.65	中忽	1113	1.23	兀惕	310	0.34	中忽勒	94	0.10
4	阿	2389	2.64	舌里	1059	1.17	額惕	205	0.23	中合黑	92	0.10
5	亦	1862	2.06	舌列	873	0.96	阿勒	200	0.22	舌魯黑	55	0.06
6	客	1844	2.04	中豁	857	0.95	額勒	189	0.21	舌剌黑	48	0.05
7	荅	1328	1.47	舌剌	765	0.85	里黑	178	0.19	舌鄰勒	47	0.05
8	訥	1301	1.44	中罕	487	0.54	斡克	150	0.17	舌列克	39	0.04
9	列	1287	1.42	舌侖	458	0.51	斡惕	136	0.15	中豁勒	38	0.04
10	赤	1261	1.39	舌魯	407	0.45	阿惕	118	0.13	中忽惕	37	0.04

It is believed that some phenomenon in this table will arouse interest of researches. For example, character “児” with word initial “舌” (type xC) shows the number of 3,707 times, but “児” with initial “中” only 3 times. Thus, it is definite that there are some errors within the handwriting process. Furthermore, with the comparison to type C and xC of “児”, the frequency is still open to suspicion. The former is in frequency of 1,116 times, and the latter 3,707. Are they two characters? In the light of the SHM experts’ study, the two types of “児” are due to “inconsistency character usage” [6]. Therefore, from the angle of statistics, any character “児” without an initial is an error. Eldengtei has pointed out that type xC of “児” rarely emerges within volume 1 or 2, which has been proved by the computing statistic data. In volume 1 “児” emerges 362 times, none of which takes word initials. In volume 2, the number of “児” is 478, among which there are only 8 items with initials. Yet from volume 3, things changed dramatically. “児” appears 408 times in volume 3, among which there are 391 items of type xC. The same is to volume 4 and 5 and so on so forth.

Altogether, there are 15 word endings in the main body of the text. However, in fact, some of them are homophones, which are caused by the writers for representing original word meaning with Chinese radical. For instances, “泐”, “汐”, “防” and “勒” are of the same pronunciation and of a little different senses. Some word endings may be fabricated arbitrarily or by miscopy. Table 5 shows the data and rates of the endings in the text.

Table 5. Data of word endings

endings	token	rate	endings	token	rate	endings	token	rate	endings	token	rate
勒	2804	39.10	卜	465	6.48	渤	13	0.18	乞	1	0.01
惕	1851	25.81	思	35	0.48	赤	4	0.05	失	1	0.01
黑	1122	15.64	木	32	0.44	防	1	0.01	室	1	0.01
克	815	11.36	你	24	0.33	汐	1	0.01			

Some attentions should be paid to on the limitations among word initials, endings, and Chinese-transliterated characters. For example, Only 20 items of Chinese characters can take both initials and endings with themselves. They are “罕, 合, 忽, 渾, 豁, 刺, 闌, 驪, 里, 澧, 連, 列, 劣, 洌, 鄰, 魯, 侖, 捋, 羅, 思”. On the other hand, the number of characters attached to with initials is 87, and that of characters with endings is 317.

Table 6. Four types of distributional statistics in different volumes

	C	token	rate	xC	token	rate	Cy	token	rate	xCy	token	rate
V.1	250	5949	6.5766	37	820	0.9065	111	367	0.4057	19	62	0.0685
V.2	236	6241	6.8994	34	907	1.0026	113	396	0.4377	13	45	0.0497
V.3	253	5309	5.8690	37	1354	1.4968	115	406	0.4488	21	87	0.0961
V.4	236	5670	6.2681	32	1175	1.2989	114	465	0.5140	25	76	0.0840
V.5	224	6088	6.7302	30	1252	1.3840	114	505	0.5582	24	109	0.1205
V.6	251	6220	6.8761	42	1417	1.5665	115	495	0.5472	27	144	0.1592
V.7	231	5460	6.0360	28	1192	1.3177	134	517	0.5715	22	94	0.1039
V.8	230	5217	5.7674	33	1216	1.3443	128	491	0.5428	24	97	0.1072
V.9	208	5183	5.7298	27	1098	1.2138	110	590	0.6522	22	93	0.1028
V.10	219	4765	5.2677	32	1096	1.2116	115	551	0.6091	23	68	0.0752
V.11	240	5931	6.5567	32	1337	1.4781	117	538	0.5947	23	86	0.0951
V.12	235	6777	7.4919	33	1613	1.7832	119	793	0.8766	25	95	0.1050
Total	510	68810	76.069	83	14477	16.004	317	6114	6.7590	52	1056	1.1674

To have a general understanding of character usage in SHM and character inconsistency issues caused by different transcribers in different volumes, some descriptions are to be shown in table 6, which lists in volumes and types with rate. Please pay more attention to the data with underlines.

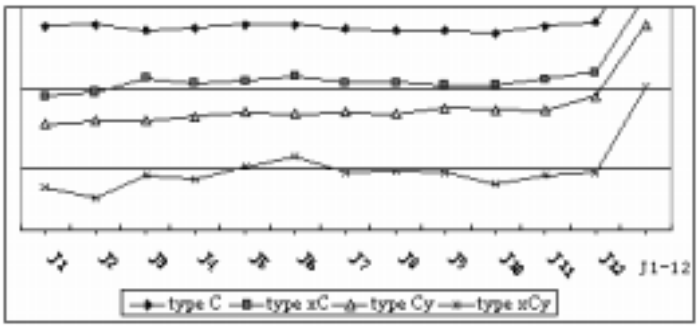


Fig. 1. Distribution curve of frequencies of all types (J stands for volume)

Generally speaking, based on the above statistics, character types are distributed evenly in every volume. It is only something that makes surprising results that the number of type xC is too low in first

two volumes. The reason is that character “児” always appears in the format of C in these two volumes, and it is the character with the highest frequency in the text. In volume 6, type xC and type xCy are slightly high. The reason need to be explored further. In volume 12, the occurrences of type C, type xC, and type Cy are higher than that in other volumes, which probably relate with different transcribers or with the length of the volumes. The following is a curve diagram for each volume, in which curvilinear of type xCy rises and falls so much, such as the figure in volume 6 is 2.3 times as much as that in volume 1, and 3.2 times as much as that in volume 2.

4 Analyses of Characters in Aligned Chinese Glosses and Translated Chinese Paragraphs

The amount of aligned Chinese glosses (ACG, 1,567 items) is much higher than that of characters in the text main body, close to the amount of translated Chinese paragraphs (TCP, 1,665). On the other hand, the token of aligned Chinese glosses (61,893) is a little higher than that of translated Chinese paragraphs (40,674). Please check up the following data in table 7.

Table 7. Statistics of TCP and ACG

Char. types	Amount	Token	Proportion	Rate	Rate
Types of TCP	1816	40674*	100.00	100.00	
Type C	1657	38839	91.2445	95.4885	
Type xC	33	1227	1.8172	3.0167	
Type Cy	108	533	5.9471	1.3104	
Type xCy	18	75	0.9912	0.1844	
Char. in TCP[#]	1669	47907**			21.3354
Initials	2	1302			
Endings	9**	608			
Character	1665	40674			
Full stop	1	5323			
Char. in ACG	1567	61893			27.5641
Annotative char.	10	1582			0.7045
Chin.trans.char.	546	113160			50.0396
Total Chinese char.	2099[##]	224542			100.00

Note: * full stop is not included in character type statistics. ** word endings “汐, 渕, 惕” do not appear in TCPs. Statistics are based on 1677 item, including characters, initials, endings, and full top. [#] full stops are included. [##] “阞, 汐, 渕, 惕”, which only emerge as endings, are included.

The four types of characters in translated Chinese paragraphs are caused mainly by names of persons and places. However, characters with initials or endings take small proportion, emerging only 1,835 times and accounting for 4.5115% of characters, in translated Chinese paragraphs.

Now, some additional statistics should be mentioned here. The amount of characters in TCP and ACG is 1950 items, in which there are 1,280 items in common, and 285 items with 524 token in ACG, and 383 items with 4,641 token in TCP. The total token of characters is 102,567 in TCP and ACG.

With respect to the Chinese characters and symbols all over the book, there are altogether 2,099 items of Chinese characters and symbols, including Chinese-transliterated characters in the text main body, supplementary symbols as initials and endings, annotative characters, characters in aligned Chinese glosses, characters in translated Chinese paragraphs, to the exclusion of reduplication. In addition, all the above characters emerge 224,542 times, which is the number of the whole book. At last, it is necessary to point out that there are only 145 items of Chinese-transliterated characters (2,648 token), which never emerge in TCP and ACG, among 540 items in the main body of the text.

5 Conclusion

This paper mainly discusses statistics of Chinese-transliterated characters and all the other characters and symbols in SHM. It is certain that the data and statistics are highly reliable, for the check is a full and exhaustive inspection. As for other issues in SHM, such as error characters, mistaken attachment with symbols, redundant characters and symbols, reversals and omission of characters in the origin, and for some explanations to the reason of high or low frequencies, they are all worth to be explored further in other papers.

Acknowledgments. Thanks to Prof. Junast, Prof. Maosong Sun, Prof. Wulan, and Dr. Shengli Zu for their constructive suggestion to the research.

References

1. Chen, Yuan: An Inspection on Chinese Characters Used in SHM. In: *An Academic Collections of Chen Yuan* (volume 2), Beijing: Zhonghua Book Company Press (1982)
2. Shiro Hattori: A *Study on Chinese-transliterated Characters from Mongolian in SHM*. (服部四郎，1946：元朝祕史の蒙古語を表はす漢字の研究，Tokyo: 龍文書局(1946)
3. Jiang, Di, etc.: Approach to the Retrieval System of the Secret History of the Mongols, *Journal of Chinese Information processing*, 2006: Vol. 20: 3
4. Shuang, Fu: *The translateration and studies on the Secret History of the Mongols,* Huhehot: Inner Mongol People Press (in Mongolian) (2002)
5. Ozawa Shigeo: The course of Mongolian Grammar of the Secret History of the Mongols, Tokyo: 风间书房 (1993)
6. Eldengtei, ect.: *A collated edition of the Secret History of the Mongols*. Huhehot: Inner Mongol People Press (1980)
7. Anonymous: The Secret History of the Mongols, the version from Sibu Congkan(四部丛刊) (13th Century)