The Electronic Data and Retrieval of the Secret History of the Mongols

Di Jiang

Institute of Ethnology & Anthropology
Chinese Academy of Social Sciences

Outline

- The background of SHM
- The original format of Chinese-transliterated document
- Make the electronic data for SHM
- Design a data retrieval system for SHM
- What may the electronic data tell us?

What is SHM?

- SHM is a great classical historical work in 13th century. The name of the book is called *Monqol-un Nihuča Tobčiyan* in ancient Mongol (The Secret History of the Mongols, SHM)

- The book tells a historical story about Mongols and Genghis Khan, the conqueror of Central Asia and southern Russia, the founder and pioneer of Yuan Dynasty.

SHM of the world

- The work SHM is so famous that it has become a branch of learning: SHK-ology all over the world.
  - There are more than ten language translations of SHM in the world
  - There are more than twenty Latin transliterations of SHM from Chinese characters
  - There are more than ten thousands of papers and books on SHM
  - There are more than one hundred experts of SHM in the world
  - ……

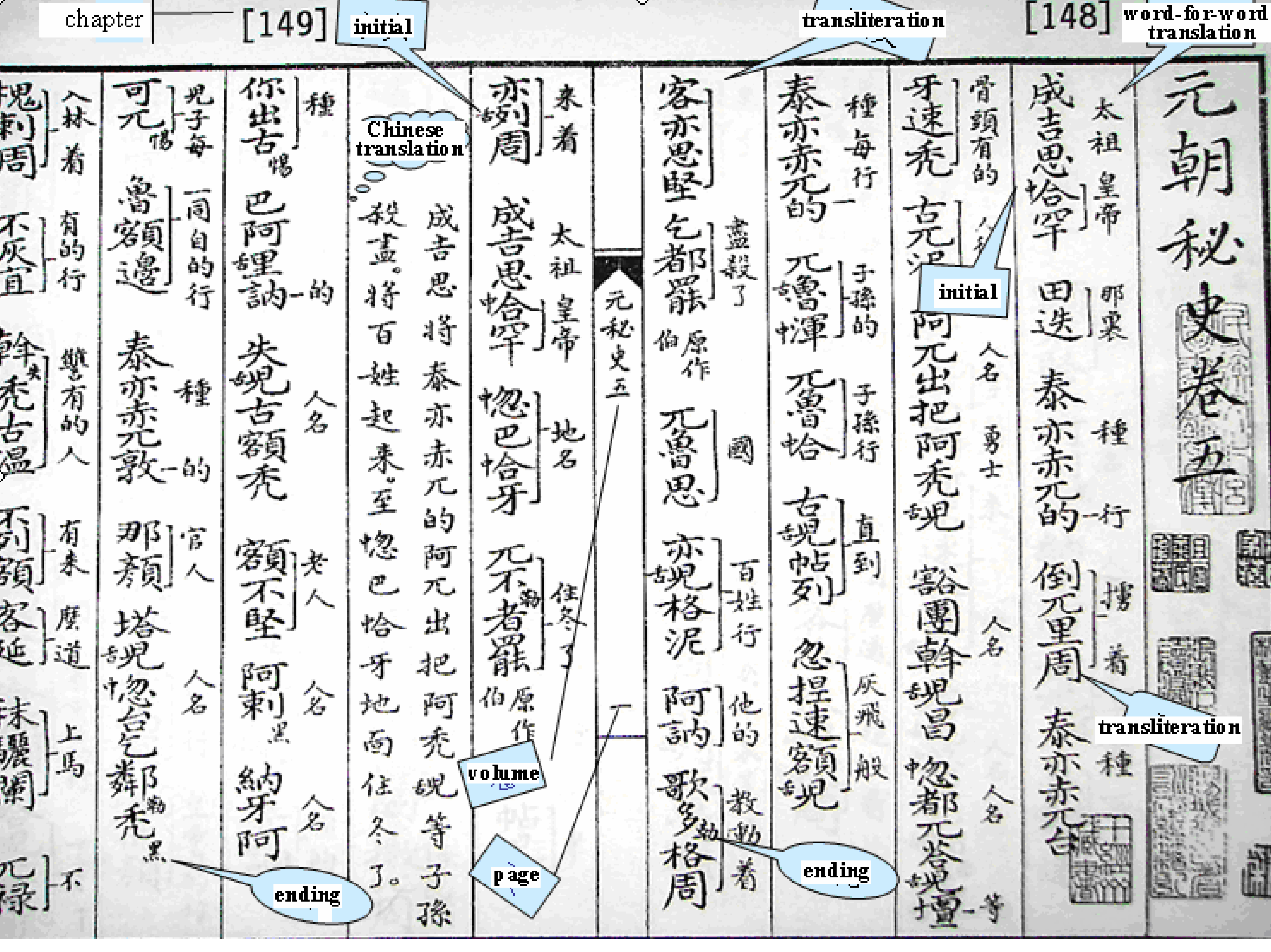
Secrets of the Secret History

- Why does SHM arouse so much interests of scholars? Its secrets or riddles?
- People say: you may find the tomb of Genghis Khan, however, you may have no way to find the original SHM.
  - The original SHM has been lost. Can it be found again?
  - Who is the author of the book?
  - Why was the book handed down to generations in Chinese characters?
  - How can the book be restored to original Mongol from Chinese-transliterated characters?
  - What is concealed among the Chinese characters of SHM?
  - ……

The format of Chinese-transliterated document

- the length of the book: 300 thousand Chinese characters (in guess before statistics).
- There are 12 volumes and 282 chapters by transliterators

- Format:
  - layouts: characters written vertically
  - the original shape of the archaic handwriting form with three lines representing one content
    - The first line is word-for-word Chinese characters (interlinearize)
    - The second line is Chinese-transliterated characters from Mongol
    - The third line is initials of the pronunciation of Mongol words
    - There are also endings indicating pronunciation of Mongol words within the second lines



元朝秘史卷五

成吉思合罕 田迭 泰亦赤兀的 倒兀里周 泰亦赤兀台

牙速秃 古兒 阿兀出把阿秃兒 豁團斡兒昌 忽都兀答兒

泰亦赤兀的 兀魯黑 兀魯恢 古兒帖列 忽提速客兒

客亦思堅 乞都罷 兀魯思 亦兒格泥 阿訥 歌篾勒格周

元秘史五

亦列周 成吉思合罕 忽巴合牙 不者罷

成吉思將泰亦赤兀的阿兀出把阿秃兒等子孫殺盡。將百姓起來。至忽巴合牙地面住冬了。

你出古惕 巴阿禮訥 失乞古額秃 額不堅 阿剌黑 納牙阿

可兀惕 魯客額邊 泰亦赤兀敦 那顏 塔兒忽台乞鄰勒秃黑

Can we make an electronic SHM?

- SHM is an ancient handwriting document
- SHM is a typical complicated document
- The carrier of information is unintegrated
- The characters and texts of SHM are non-standard and irregular

Principles of making Electronic data for SHM

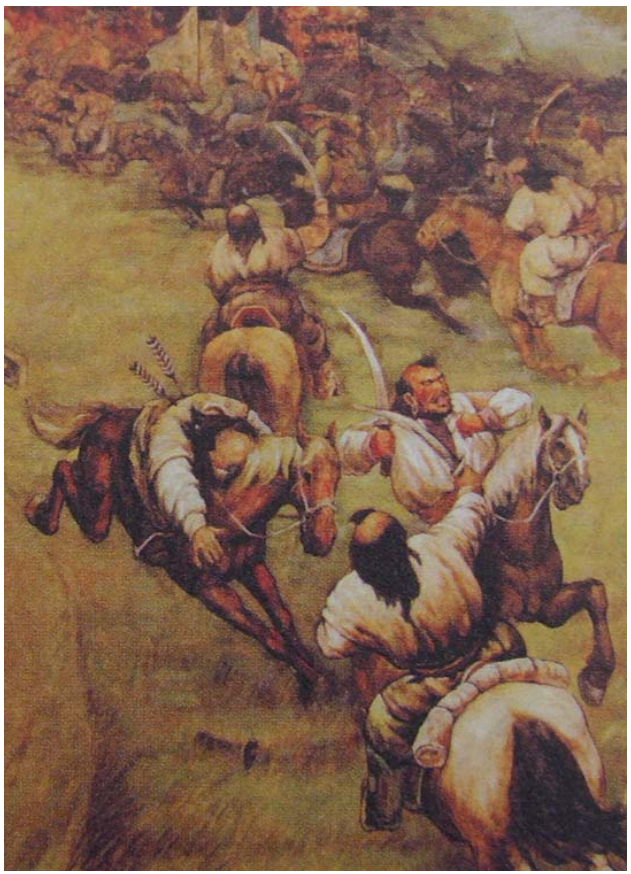
- Preserve the original format of SHM
  - Version, direction of handwriting, and annotation of pronunciation
- Keep all the information of characters
  - Simplified Characters, traditional Characters, and variant characters
- Keep all the information of volumes, chapters, and pages
- Keep all the information of segmentation and word-for-word format

The information of the handwriting document

- Layouts: three lines for one content
  - direction of handwriting: from vertical lines to lines sideways
  - Initials stand at one line
  - Word-for-word translation stands at one line

	中						－initial
成吉思	合罕	田迭	泰亦赤兀	的	倒兀里周	泰亦赤兀台	－transliterate
太祖	皇帝	那裏	種	行	擄着	種	－word-for-word trans

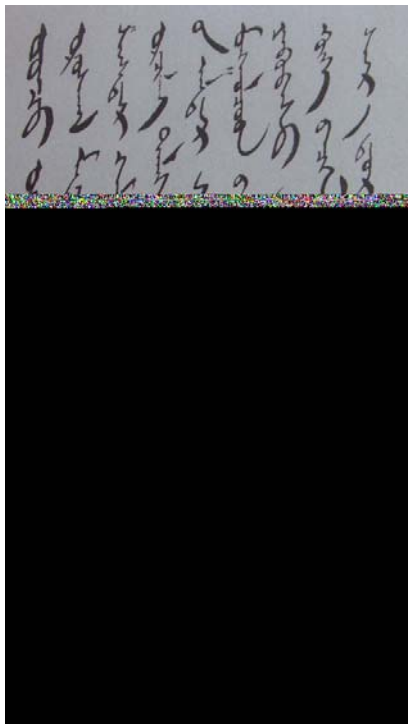
The information of the handwriting document

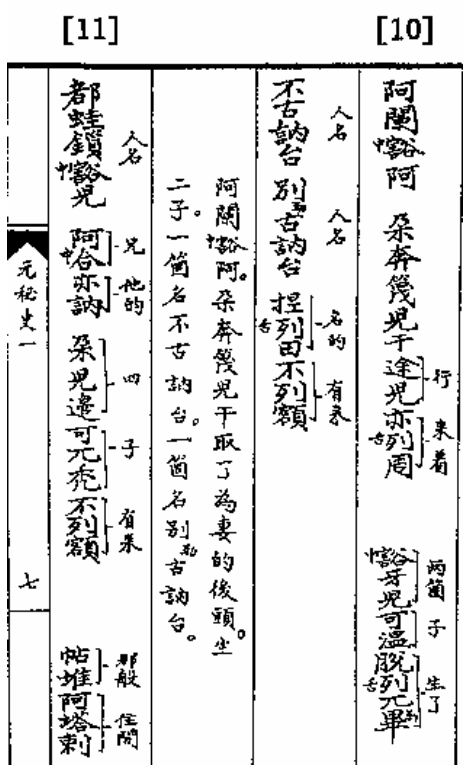


- Volumes:
  - add a volume number before each volume
- Chapters:
  - keep the numbers of chapters
- Chinese translation
  - Be separated from the main texts

The information of the handwriting document

- Characters:
  - keep all the forms of characters
  - “斷”/“断”
  - “驖”/“鐵”
  - “幾”/“几”
  - “仇”/“讎”/“讐”
  - ……





The format of Electronic version (deep structure)

page

chapter

translit eration

initial

ending

align

元朝秘史　　卷一　忙豁侖纽察脱察安

1.1

＃1

成吉思[名] 合罕訥[皇帝的] 忽札兀兒[根源]

迭額列{舌}[上]騰格理{舌}[天] 額扯[處] 札牙阿禿[命有的] 脱列{舌}<克>先[生了的] 孛兒帖[蒼色] 赤那[狼] 阿主兀[有] 格兒該[妻]亦訥[他的] 豁{中}埃[悋白色]馬闌{舌}<勒>[鹿] 阿只埃[有來] 騰汲思客[水名]禿<勒>周[渡着]亦列{舌}罷[來了] 斡難[河名] 沐漣{舌}訥帖里{舌}兀捏[河的源行] 不峏罕哈<勒>敦[山名]納[行]嫩兀<黑>剌周[營盤做着] 脱列{舌}<克>先[生子的] 巴塔赤罕[人名] 阿主兀[有來]

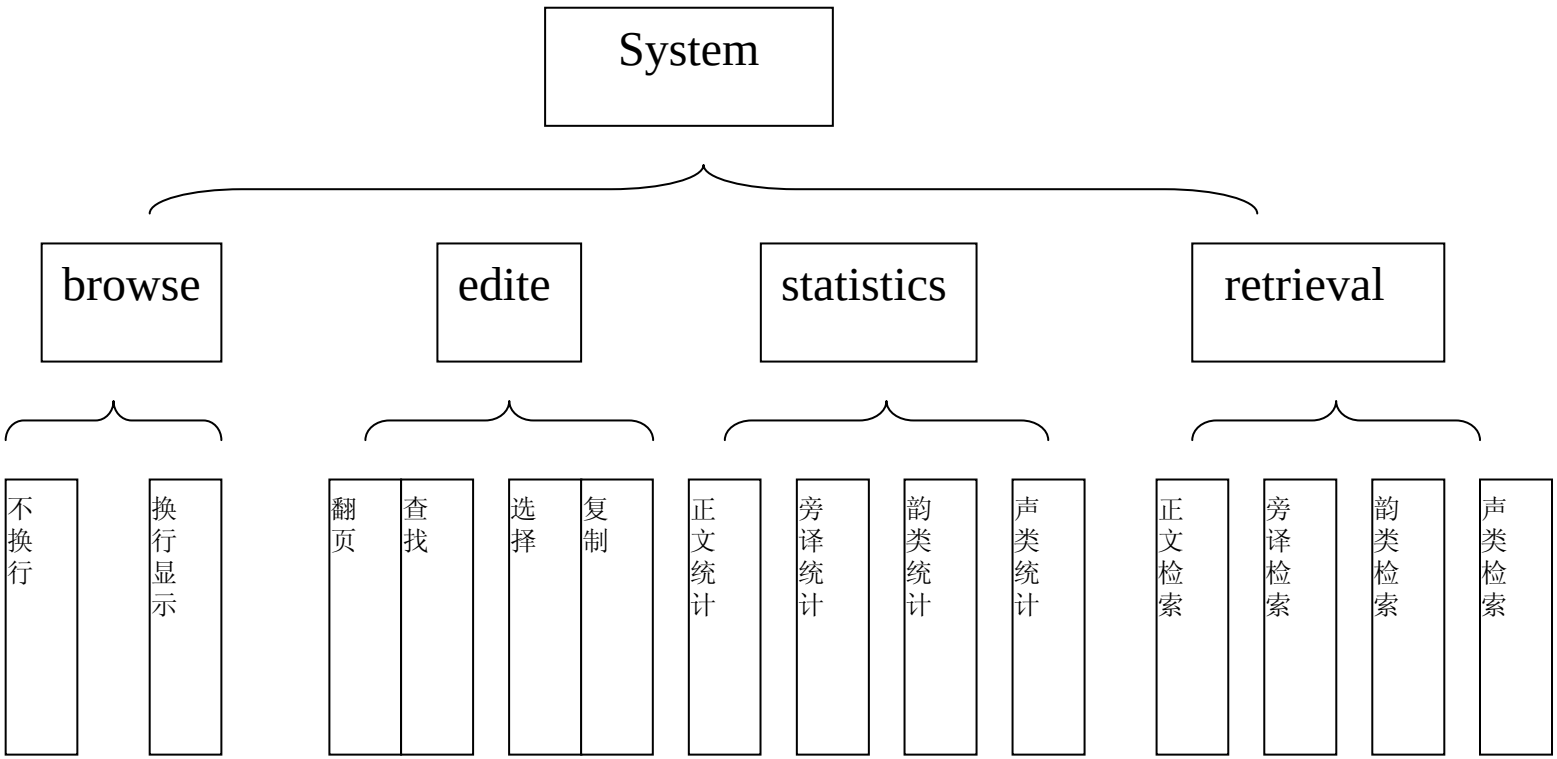
//當初元朝的人祖。是天生一箇蒼色的狼。與一箇悋白色的鹿相配了。同渡過騰吉思名字的水來。到於斡難名字的河源頭。不兒罕名字的山前住着。產了一箇人名字喚作巴塔赤罕。//

translation

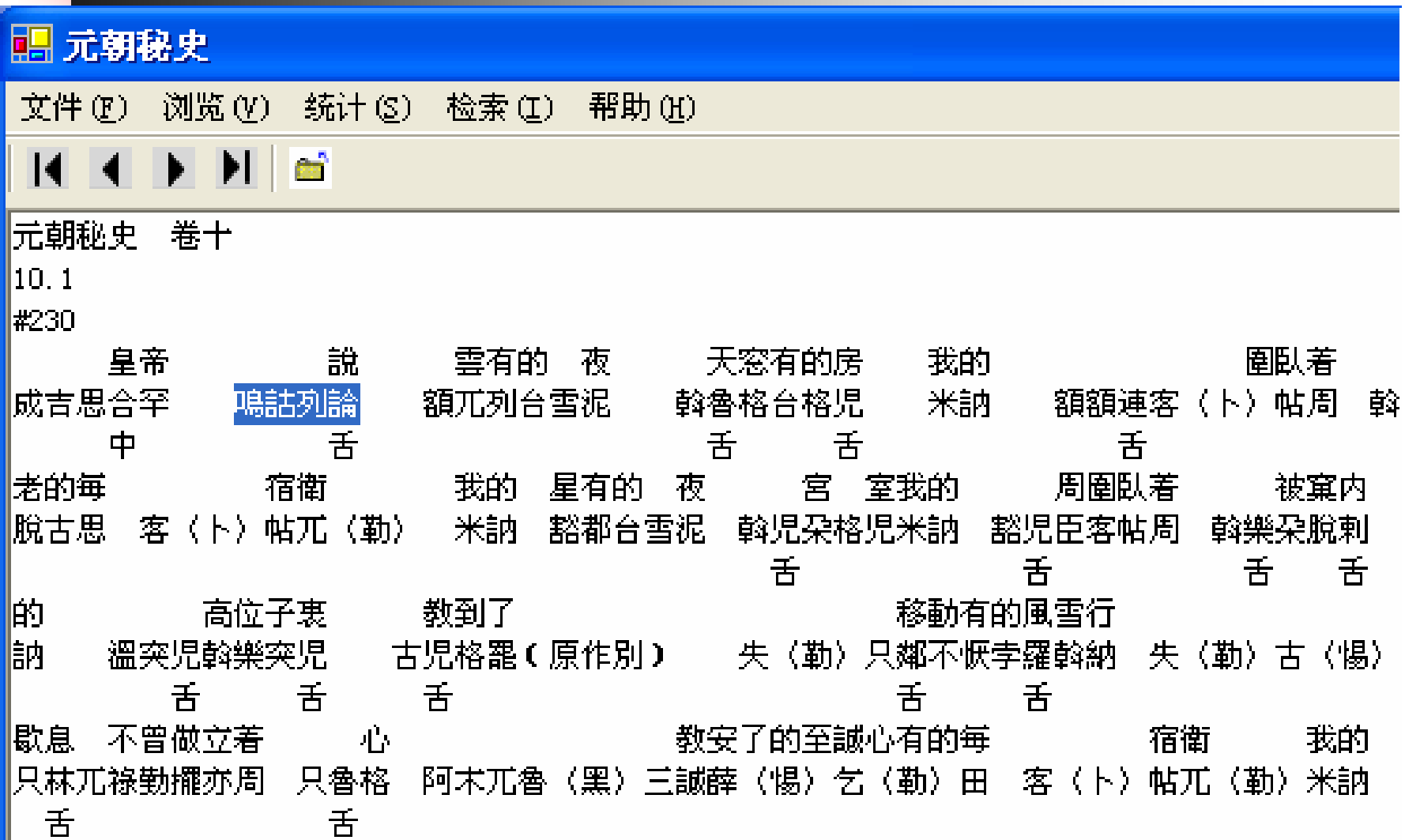
Design for the retrieval system

- Functions of Browsing and seeking
  - Any characters or words
- Functions of statistics
  - Transliteration, Translation, Initials, endings
- Functions of retrieval
  - Transliteration, Translation, Initials, endings
- Functions of data output
- ……

Design for the retrieval system



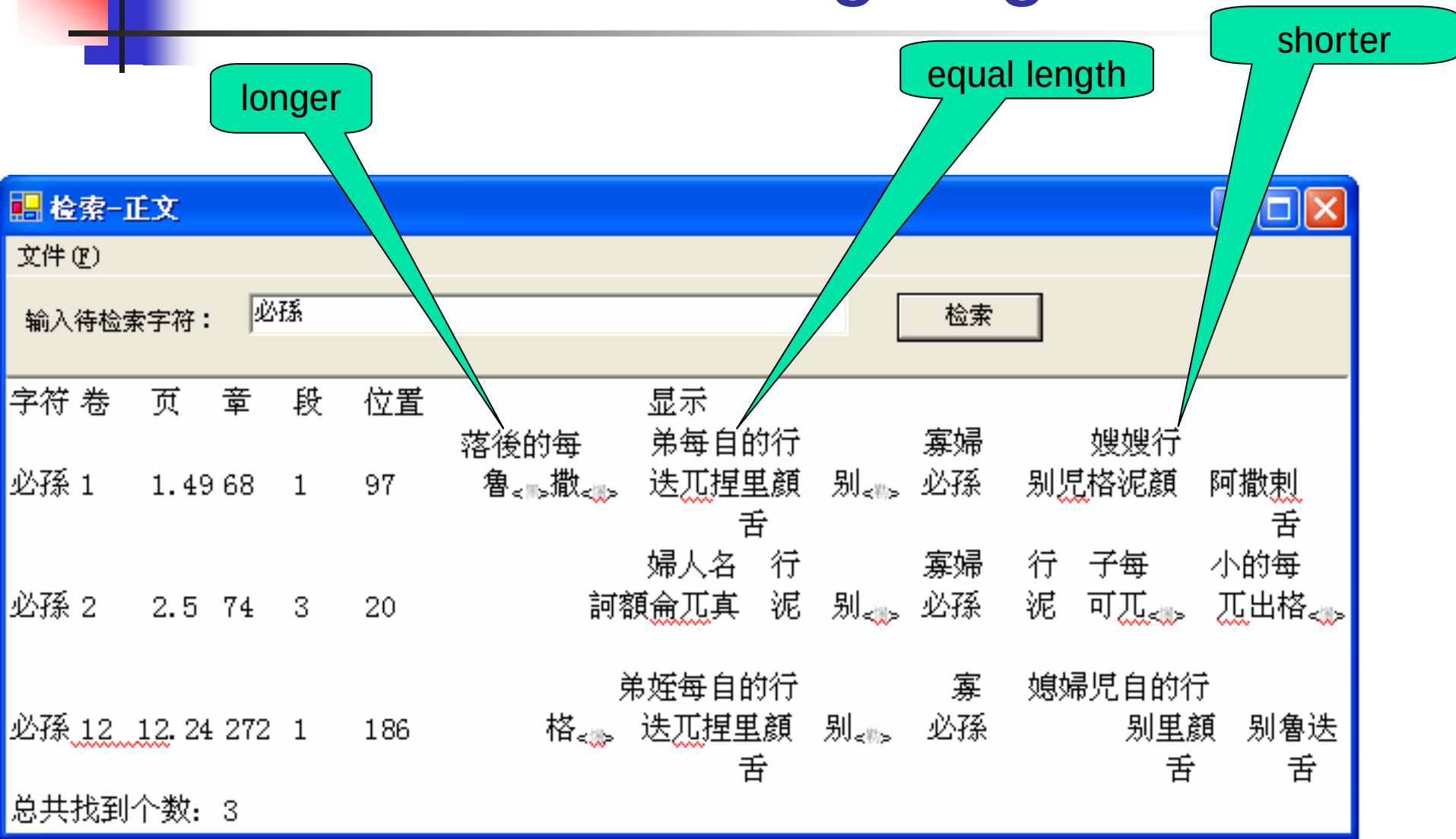
Browse the original format



Retrieval by alignment

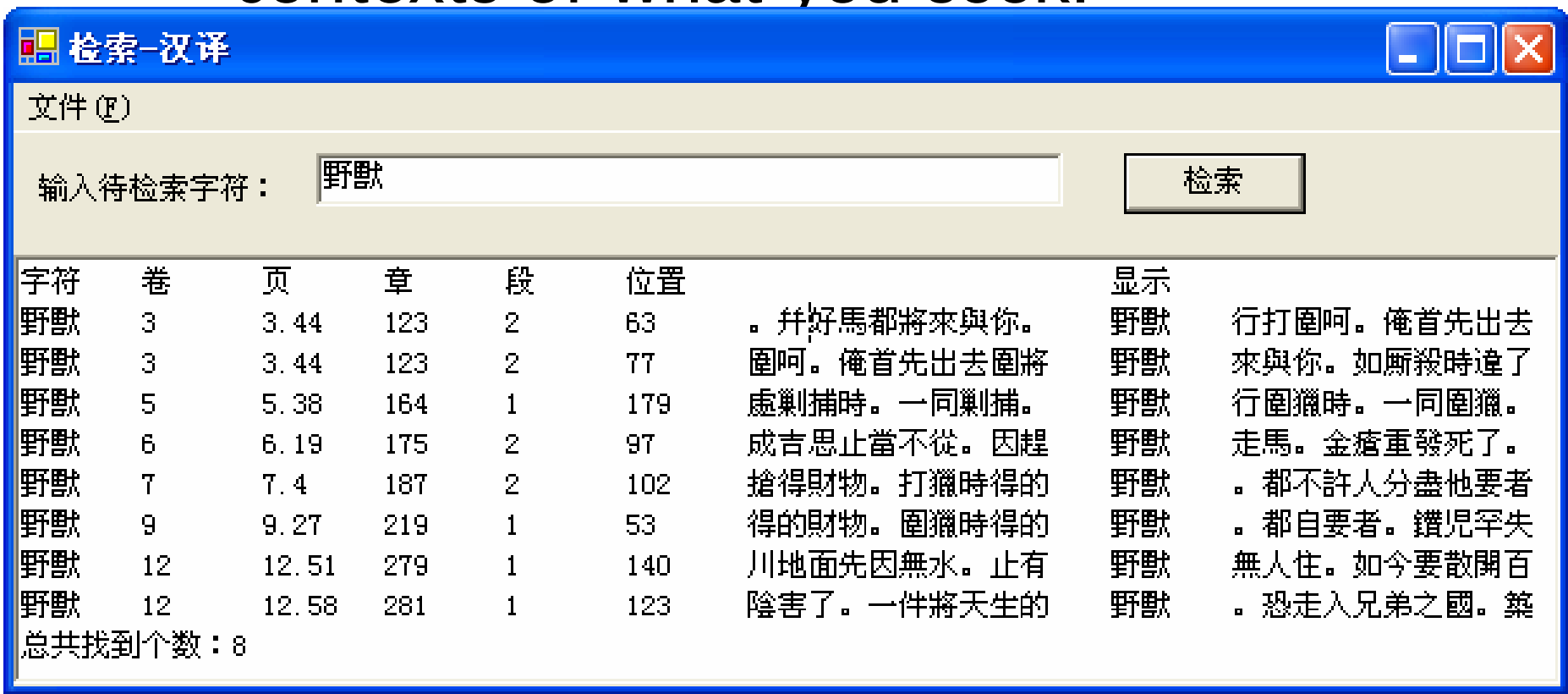
- To realize one content with three lines, the technology of interlinearization (alignment) has been used, which judges the browsing position of the three chunks according to what line is longer between 1st lines, 2nd lines, and 3rd lines.

The result of aligning retrieval



Retrieval of translation by concordance

- A method of concordance may tell the contexts of what you seek.



What may the electronic data tell us? (for historians)

- Who appears in SHM?
- Where did the named persons go?
- What events happened to them?
- How can you find the secrets of original texts from Chinese characters?
  - Seek names of persons
  - Seek names of places
  - Seek some event keywords

检索-汉译

文件(F)

输入待检索字符： 斡難河 检索

字符	卷	页	章	段	位置		显示	
斡難河	1	1.16	24	1	63	梁瘡禿尾子的馬。順着	斡難河	去。到巴諄阿剌名字的
斡難河	1	1.19	30	1	18	不忽合塔吉。後來順着	斡難河	去尋他。　行到統格黎
斡難河	1	1.31	50	4	61	一子。名不里孛闊。於	斡難河	邊筵會時。將太祖的弟
斡難河	1	1.34	54	3	15	父。也速該把阿禿児在	斡難河	放鷹。見篾児乞氏的人
斡難河	1	1.36	55	2	160	了。即便打着馬。逆着	斡難河	走了。
斡難河	1	1.38	56	2	116	怎生般艱難。哭的聲將	斡難河	的水幷川裏林木都振動
斡難河	1	1.41	59	1	49	児的妻訶額倫正懷孕於	斡難河	邊迭里溫孛荅山下生了
斡難河	2	2.7	75	2	38	親上頭。將針做鉤児於	斡難河	裏釣魚。又結網捕魚。
斡難河	2	2.18	81	1	52	十六日。泰亦赤兀每於	斡難河	岸上做筵會。日頭落時
斡難河	2	2.18	81	1	114	於頭上打倒走了。走到	斡難河	邊林内臥着恐怕人見。
斡難河	2	2.18	81	1	129	内臥着恐怕人見。又入	斡難河	水的溜道裏仰臥着。身
斡難河	2	2.19	82	2	45	來着。白日般月明裏。	斡難河	邊樹林裏挨排着尋。帖
斡難河	2	2.22	84	2	116	。必救了我。所以順着	斡難河	尋鑽児罕失剌去了。
斡難河	3	3.9	106	2	107	軍。共二萬軍上馬逆着	斡難河	來。孛脫罕孛斡児只的
斡難河	3	3.13	108	1	24	字那裏相合了。起去往	斡難河	源頭。原相約會處。孛
斡難河	3	3.27	116	1	66	時。帖木真十一歲。於	斡難河	冰上打髀石時。札木合
斡難河	4	4.5	129	2	134	被札木合推動。退著於	斡難河	哲列捏地面狹處屯札了
斡難河	4	4.7	130	2	81	百姓來了。喜歡。着於	斡難河	邊林裏做筵席。先於訶
斡難河	4	4.38	144	2	89	落。將百姓起了。渡過	斡難河	整治軍馬。候成吉思來
斡難河	8	8.27	202	2	21	捕了。至是虎児年。於	斡難河	源頭。建九腳白旄纛做
斡難河	9	9.6	211	2	31	風匣自不峭罕山來。於	斡難河	迭里溫孛荅地面生我時

总共找到个数：21

A sample (for grammarians)

检索-汉译

文件(F)

输入待检索字符： 被 检索

字符	卷	页	章	段	位置		显示	
被	1	1.6	9	2	30	敦地面貂鼠青鼠野物。	被	自火裏禁約。不得打捕
被	1	1.34	53	1	5	親自送去。	被	塔塔児人拿了。送與大
被	1	1.34	53	1	97	人。為親送女兒上頭。	被	人拿了。今後以我為戒
被	1	1.38	56	2	88	說。我的丈夫頭髮不曾	被	風吹。肚腹不曾忍餓。
被	1	1.39	57	2	6	因俺巴孩合罕	被	拿時。將合荅安忽圖剌
被	1	1.48	67	2	58	乞顏來了。因記起舊日	被	擄得冤雠。暗地裏和了
被	1	1.49	68	2	52	帖木真去做女婿回時。	被	塔塔児家暗地毒害了。
被	2	2.3	71	1	38	俺巴孩皇帝死了麼道。	被	訶額侖這般說。
被	2	2.5	73	2	5	察剌合老人	被	傷在家裏臥時。帖木真
被	2	2.5	73	2	36	收的幷俺衆人的百姓。	被	他將去。因勸他的時分
被	2	2.5	73	2	47	他將去。因勸他的時分	被	他傷了。帖木真哭將出
被	2	2.8	76	2	80	。我釣得一箇金色魚。	被	別帖児別古台奪了。他
被	2	2.10	77	2	31	我昨前射得箇雀兒。也	被	他奪了。今遍釣得箇魚
被	2	2.10	77	2	45	。今遍釣得箇魚兒。又	被	他奪了。似這般呵。一
被	2	2.10	77	2	115	。将箭抽着要射他時。	被	別帖兒児了說。泰亦赤
被	2	2.14	79	2	174	。上馬走入山林裏去。	被	泰亦赤兀看見了。隨後
被	2	2.16	80	2	189	割開。牽着馬下山來。	被	泰亦赤兀守的人拿將去
被	2	2.24	85	1	41	箇名字的兒子說。雀児	被	龍多児趕入草叢去呵。
被	2	2.30	90	2	14	真的慘白騧馬八疋在家	被	賊刼将去了。又有一箇
被	2	2.31	90	1	25	來。帖木真說。我的馬	被	人劫去了。說了。別古
被	2	2.51	103	2	99	躲得過。我的小性命。	被	不児罕山遮救了。這山
被	3	3.3	104	2	43	王罕處去到了說。不想	被	三種篾児乞每。將我妻
被	3	3.6	105	3	39	。教對他說。我的妻子	被	篾児乞每擄要了。咱每
被	3	3.7	105	1	34	也聽得帖木真安荅的妻	被	人擄了。我心上好生痛
被	3	3.19	111	2	56	弟赤列都。他妻訶額侖	被	也速該奪要了。因此來
被	3	3.29	117	1	135	下。做了筵席。夜晚共	被	宿有來。
被	4	4.5	129	2	124	巴主地面對陣。成吉思	被	札木合推動。退著於斡
被	4	4.8	130	1	76	坤太子。死了的上頭。	被	人這般打。說著大聲哭
被	4	4.10	131	1	3	韁繩。	被	別古台拿住。不里孛闊

A sample (for corpus linguists)

- Statistics for Chinese characters in SHM

Char. types	Amount	Token	Rate
1, Characters of translation	1669	47907	21.3354
Initials	2	1302	
Endings	9	608	
Character	1665	40674	
Full stop	1	5323	
2, Characters of alignment	1567	61893	27.5641
3, Annotative characters	10	1582	0.7045
4, Characters of transliteration	546	113160	50.0396
Single words	540	90457	
Word initials	2	15533	
Word endings	15	7170	
Total Chinese characters	2099	224542	100.00

Thank you for your attention!

- The electronic data of SHM will be useful for
  - Historians
  - Philologists
  - Linguists
  - Geographers
  - Military scientists
  - Students

Types of characters in SHM

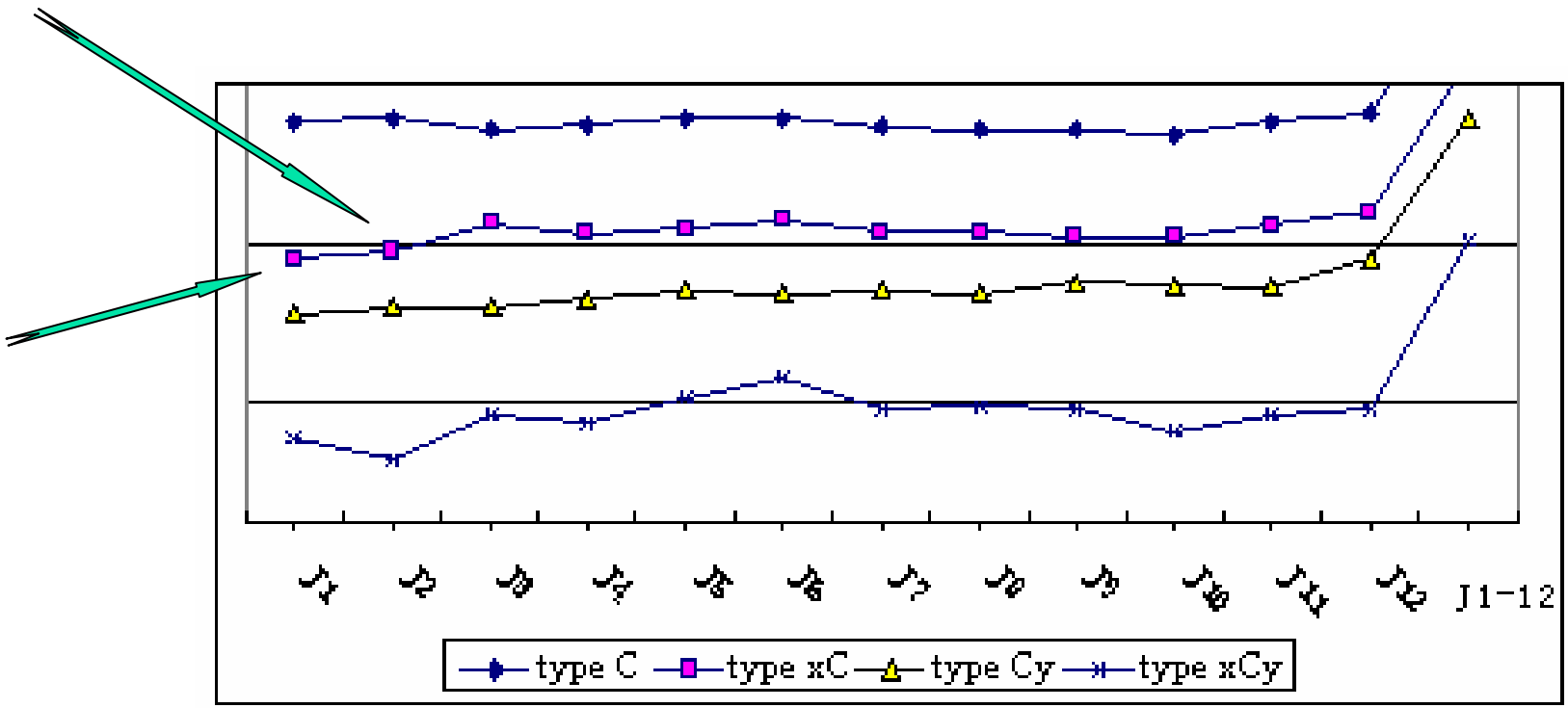
- Type one (C): single character.
  - “児”
- Type two (xC): character with initials.
  - “舌児”
- Type three (Cy): character with endings.
  - “阿勒”
- Type four (xCy):character with initials and endings.
  - “舌魯黑”

Statistics for types of characters in SHM

- Pay attention to tokens of Type xC inV.1 & v.2

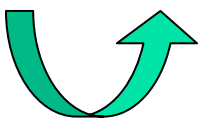
	C	token	rate	xC	token	rate	Cy	token	rate	xCy	token	rate
V.1	250	5949	6.5766	37	820	0.9065	111	367	0.4057	19	62	0.0685
V.2	236	6241	6.8994	34	907	1.0026	113	396	0.4377	13	45	0.0497
V.3	253	5309	5.8690	37	1354	1.4968	115	406	0.4488	21	87	0.0961
V.4	236	5670	6.2681	32	1175	1.2989	114	465	0.5140	25	76	0.0840
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

The reason: An interesting case



See how the character types influence the whole text

- “児”(and “舌児”) is a character in the most high frequency in the whole document.
- There are two forms of “児”, one is single “児”，another is “舌児” with initials.
  - “児” in vol. 1 is 362, “舌児” is 0
  - “児” in vol. 2 is 478, “舌児” is 8
  - “児” in vol. 3 is 17, “舌児” is 391
  - “児” in vol. 4 is 57, “舌児” is 298
  - “児” in vol. 5 is 16, “舌児” is 376
  - ……
  - “児” in vol. 1~12 is 1116, “舌児” is 3707



- Now we know the transliterators made some errors in vol.1 and 2. All forms of the character should be “舌児”.