

ЦИФРОВЫЕ ТЕХНОЛОГИИ В ГУМАНИТАРНЫХ ИССЛЕДОВАНИЯХ

DOI 10.30792/2304-1838-2019-91-105

УДК 81.347.7

ББК 81.1

Kuribayashi Hitoshi

ON THE DEVELOPMENT AND UTILIZATION OF WEB-DICTIONARY OF MONGOLIAN TRADITIONAL DICTIONARIES

The article discusses the experience and the results of international cooperation between The Center for Northeast Asian Studies of Tohoku University (Japan) and the School of Mongolian Studies of Inner Mongolian University (China) on creating an electronic Mongolian-Chinese dictionary. It was compiled on the basis of the "Mongolian-Chinese dictionary: revised and expanded edition" (Publishing house of the University of Inner Mongolia, 1999), which includes more than 57 thousand dictionary entries. The "Mongolian Universal White (Mongolian Universal-White.ttf)" and "Moukan Gothic12 (MOUKAN.tiff)" fonts have been developed for adequate reflecting the letters of the Mongolian alphabet. Cyrillic spelling of Mongolian words, pronunciation of the word (Chakhar dialect), correlation of the word with Mongolian dictionaries, addressing to the page of the printed edition were added to the vocabulary units of the dictionary. The article discusses in detail the theoretical issues of software development of the online dictionary, the difficulties encountered in its preparation, and proposes concrete advices on the practical use of the dictionary.

Keywords: Mongolian written language, on-line dictionary, Mongolian-Chinese Dictionary, Unicode Standard.

Курибаяши Хитоши

О РАЗРАБОТКЕ И ПРИМЕНЕНИИ ВЕБ-СЛОВАРЯ ТРАДИЦИОННЫХ МОНГОЛЬСКИХ СЛОВАРЕЙ

В статье рассмотрен опыт международного сотрудничества между университетом Тохоку (Япония) и Институтом монголоведения Университета Внутренней Монголии (Китай) по созданию электронного монгольско-китайского словаря. За основу был взят «Монгольско-китайский словарь: переработанное и дополненное издание» (Изд-во Университета Внутренняя Монголия, 1999), включающий более 57 тыс.

словарных статей. Для адекватного отражения букв монгольского алфавита был применен шрифт MongolianUniversalWhite.ttf. К словарным единицам языка были добавлены: кириллическое написание монгольских слов, произношение слова (чахарский диалект), соотнесение слова с монгольскими словарями, адресация к странице печатного издания.

Ключевые слова: монгольский письменный язык, веб-словари, монгольско-китайский словарь, кодировка Unicode.

The Center for Northeast Asian Studies (CNEAS) of Tohoku University, Japan and the School of Mongolian Studies of Inner Mongolia University, China have been a part of the Joint Project on compiling of an online “Mongolian-Chinese Dictionary” within the frames of Academic Exchange Agreement signed in 2008. The goal of the Project was to computerize one of the Mongolian-Chinese dictionaries, namely the “Mongolian-Chinese Dictionary: Revised and Enlarged Edition” (Inner Mongolia University Press, 1999) so that we can use it on our own computer and on the Internet. As the result of this Project, the “Mongolian-Chinese Dictionary: Web version” was released in January 2015 on the website of the CNEAS. Since then the Web version of “Mongolian-Chinese Dictionary” has been in service and offered for “everyone” to use at “any time” in “any place” without any account (Figure 1).

The paper edition of “Mongolian-Chinese Dictionary” is a dictionary of modern written Mongolian, which is spelled by traditional Mongolian script, now being in use by Mongolian people in Inner Mongolia and other parts of the People's Republic of China. Its first edition was published from Inner Mongolia People Publishing House in 1976 and the number of the entry words amounted to approximately 50,000. The abundance of the entry words and reliability of its content made the dictionary quite popular and highly regarded not only among the Mongols but also among students of the Mongolian language in the world. Thus from the beginning of its publication,

the dictionary has been acknowledged as the standard of modern written Mongolian.

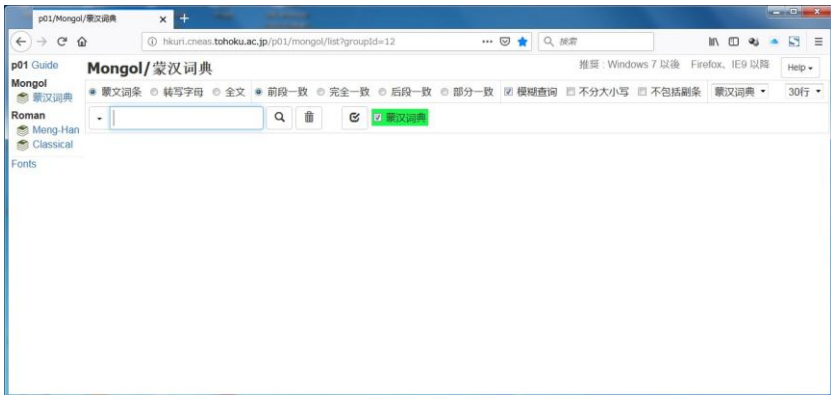


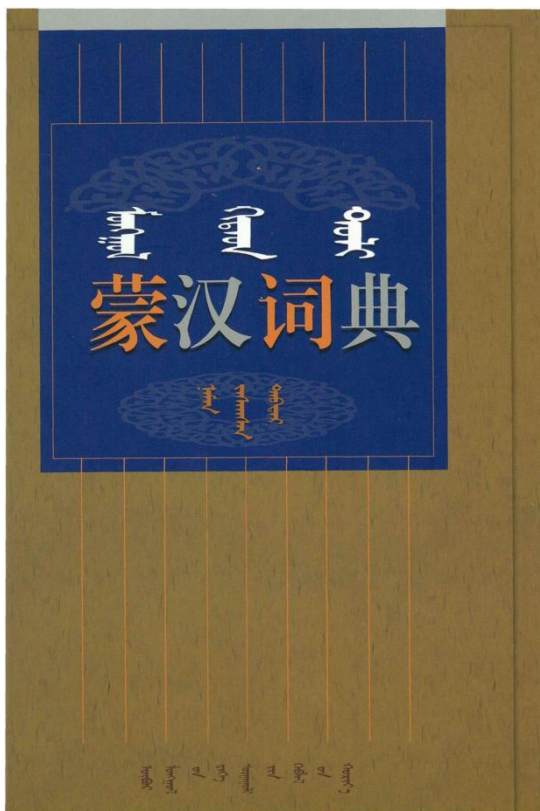
Figure 1. <http://hkuri.cneas.tohoku.ac.jp/p01/>

In 1999 its new version or “Revised and Enlarged Edition” was published, which contained more than 53,000 entry words, including around 26,000 headwords (main entry words) and around 27,000 compound words and idioms [1] (Figure 2).

The new edition of the “Mongolian-Chinese Dictionary” was groundbreaking from the fact that the phonetic denotation has been added to each headword by means of the International Phonetic Alphabet (IPA). It is especially significant because written Mongolian is basically based on the spelling system of 13th century Mongolian and its spelling is quite different from any dialect of modern colloquial Mongolian.

Furthermore, there are various dialects among the Mongolian speakers in China, and the differences between them are huge. It was only at the end of 1970's that one of the dialects of Inner Mongolia or the Chakhar dialect of Shilingol Aimag was decided to be the basis of the standard pronunciation for the Mongolian language in China, but it is still far from being popular and widespread even in 2010's. Since the phonetic transcription of the dictionary is based on the

standard pronunciation of colloquial Mongolian, the dictionary has come to serve as the standard of spoken Mongolian as well as written Mongolian.



*Figure 2. Mongolian-Chinese Dictionary /
Revised and Enlarged Edition (1999).*

From the “Mongolian-Chinese Dictionary: Revised and Enlarged Edition”, we used the data to develop our online dictionary. We have got every single piece of information on the online version such as: 1) entry words and examples in Mongolian script, 2) their Romanized transcriptions, 3) phonetic representations in IPA, 4) parts of speech, and 5) Chinese equivalents (translations). In addition, we added such

new data as 6) audio or sound data of the pronunciation of each headword, 7) Cyrillic Mongolian equivalent, 8) a link to online dictionaries of Cyrillic Mongolian, and 9) graphical or image data of the page of the original paper dictionary.

In the process of the development of the Web version of the “Mongolian-Chinese Dictionary” we encountered not a few problems both in theory and in practice. Let us see some issues from our experience for further discussion.

1. On the platform of the program

In order to use the dictionary on a computer, it is desirable to place the program on “common” platform, which means that it should run regardless of the kind of OSs or computers. We chose the program to be in the shape of “Web” or “online” dictionary because the Internet is nowadays so popular and Web browsers are available on any OS and computer in any country. We decided to put the “online” dictionary on the World Wide Web (WWW) so that everyone can use it by Web browsers on his or her own computer.

Since the specifications of Web browsers are not the same in every detail, we optimized our online dictionary to Firefox and Internet Explorer (IE 9 or later) on Windows 7 or later. This is the recommended platform of our online dictionary. Otherwise one might have some inconvenience in displaying traditional Mongolian script properly, owing to the specifications of Web browsers.

2. On encoding traditional Mongolian script

The traditional Mongolian character set has been assigned in Unicode Character Code U1800-U18AF. It contains the code points of punctuational signs, format controls (indicating free variations and vowel separator), digits, and Mongolian letters as well. Each Mongolian character have several shapes according to its position in a word and it is a function of a font set that connects characters in appropri-

ate shapes to form a word. Therefore a font set is essential to display Mongolian words in traditional Mongolian script.

In 2011 the “rule of Mongolian presentation forms character set” and “use rules of controlling characters” and other related rules were issued as the standard in the People's Republic of China [2], although they are not in the Unicode Standard. It would not be realistic just to wait and do nothing until all the issues of traditional Mongolian are resolved in the Unicode Standard. So we decided to adopt a font set of “Mongolian Universal White (MongolianUniversalWhite.ttf)” which has been developed by ALMAS Inc., Japan in strict conformity with the China standard of 2011. “Mongolian Baiti (monbaiti.ttf)” is a well-known font of Microsoft Windows, but it is compliant with China standard GB18030-2000 and cannot meet the demand of the China standard of 2011.

In order to display Mongolian words properly in the online dictionary, it is necessary to install the font “Mongolian Universal White (Mongolian Universal-White.ttf)”, which can be freely downloaded from the website of the CNEAS below:

<http://hkuri.cneas.tohoku.ac.jp/p01/font>

Another font “Moukan Goth12 (MOUKAN.ttf)” is recommended to be downloaded from the website above to be installed so as to display “Roman transcription” of traditional Mongolian headwords according to the original notation.

3. On Input Method of traditional Mongolian script and its keyboard layout

We chose that the input method of traditional Mongolian should be realized in online dictionary itself, which means that the users need not prepare nor install IME program of traditional Mongolian for themselves. All we have to do is to download and install the font “Mongolian Universal White (MongolianUniversalWhite.ttf)” stated above. Thus, input from a keyboard will be directly converted into Mongolian traditional script automatically.

The fact is that keyboard layout has nothing to do with the Unicode Standard, so we could choose any keyboard layout as we like. WPS Office and its keyboard layout are most popular among the Mongols in the PRC, but they are far from familiar to the students of the Mongolian language outside of the PRC. We adopted our keyboard layout (Figure 3) for the convenience of both Mongolian people and students of Mongolian.

The Roman letters of Figure 3 coincide the imprinted letters on an English keyboard. Such a Roman letter on a keytop will be, when typed, automatically converted into traditional Mongolian script in the search field.

ᠠ	a	ᠨ	n	ᠰ	S
ᠡ	e	ᠪ	b	ᠲ	t
ᠢ	i	ᠫ	p	ᠳ	d
ᠣ	o	ᠬ	h	ᠴ	c
ᠤ	u	ᠭ	g	ᠵ	j
ᠥ	O	ᠮ	m	ᠶ	y
ᠦ	U	ᠯ	l	ᠷ	r
		ᠰ	s	ᠠ	w

Figure 3. Keyboard layout for inputting traditional Mongolian script

Merits of the Web version of the “Mongolian-Chinese Dictionary”

Taking advantage of the computerized dictionary, the online dictionary also supports the function of various searching methods such as “Forward match” (default value), “Backward match”, “Exact match”, and “Partial match”. Using the “Full text search” option, we can search the whole text of the dictionary including not only headwords and idioms but also Chinese translations and example sentences of Mongolian. When we look up Chinese words through “Full text search”, then the “Mongolian-Chinese Dictionary” will serve as a “Chinese-Mongolian Dictionary”.

“Fuzzy search” is a function that makes it possible to search even if one doesn’t know the correct pronunciation of the target word. This is especially important because of the fact that in the traditional Mongolian alphabet, there are many letters which appear exactly the same shape, but differ in pronunciation. Such letters are called “monograph letters”. For example, the consonant letters <t> and <d> have the same shape, and <k> and <g> are also the same. Besides, the vowel letters <o> (the 4th vowel) and <u> (the 5th vowel) have the same shape, <ö> (the 6th vowel) and <ü> (the 7th vowel) are also same, and the vowel letters <a> and <e> take the same shape in non-initial position of a word.

The existence of “monograph letters” in traditional Mongolian makes it hard to learn Mongolian. If we were trying to find any word written in traditional Mongolian letters, and the word has any “monograph letters” in it, it means that there might be more than one way of reading them. A word, which has one pair of monograph letters, might be read in two ways. If there are two pairs of monograph letters, it might be read in four ways, and if there are three pairs, it increases to eight ways. Then, we have to comb through the dictionary to find over each of them. It is only after we have found the word in the dictionary, that we can know the correct reading (pronunciation)

of the target word. That is why a learner needs a lot of time and effort to “read” traditional Mongolian scripts using paper dictionaries.

On the other hand, a good many Mongolian native speakers have another problem. Among the Mongolian speakers in China, where traditional Mongolian is used, there are various dialects. And it is normally that they read the Mongolian scripts with the pronunciation of their own dialect. There are many variants of pronunciation for the words which include “monograph letters” depending on the dialects, and the gaps of them are huge. It is not uncommon that Mongolian speakers in China get the pronunciations of [t] and [d], [k] and [g], [o] and [u], and [ö] and [ü] mixed up. Those pairs have the same in shape when they are spelled by traditional Mongolian script.

Because of the situation stated above, searching for words in the dictionary is not an easy task not only for learners of the Mongolian language but also for Mongolian speakers themselves. The monograph letters are usually classified and ordered based on the pronunciation of the Standard Mongolian. Therefore, the large part of the Mongolian people, those who haven't acquired the Standard Pronunciation can hardly find the word in a dictionary, no matter how they search the words on dictionary by their own dialects.

For the purpose of mitigating this difficulty of searching for words including “monograph letters”, we have implemented the function of “Fuzzy Search” as default value in the "Mongolian-Chinese Dictionary: Web version".

If we use the “Fuzzy search”, the pairs of the “monograph letters” are regarded as the same letters. In other words, it equates the consonant letters: <t> and <d>; <k> and <g>; the vowel letters: <o> and <u>; <ö> and <ü>, and <a> and <e> at non-initial position of a word. Even though we don't know how to read the word correctly, we only have to input either one of those pair letters, then the “Fuzzy search” will find every word that has the same appearance and shows the words including “correct” pronunciations. We can say that the

“Fuzzy Search” produces an identical effect to searching the word by its shape or appearance in traditional Mongolian letters.

In the “Mongolian-Chinese Dictionary: Web version”, audio or sound data has been added. Each headword, more than 26,000 in total, has audio data, recorded by the standard pronunciation. Clicking on the speaker icon, we can listen to the pronunciation of the word. As most of the general users of the dictionary, either a Mongolian or a foreigner, might not be familiar with the phonetic symbols, we have given extra consideration to adding a listening function to the pronunciation directory.

Numbers have been added at the end of each item in the search result which indicate the location or the page of the original dictionary. Each page number, ranging from 1 to 1414, is linked to the image on the original page, and on clicking on any number, an image of the original page pops up. This function would increase the dependability of the data of the online dictionary. We can move the image to any place and enlarge or reduce it by clicking on it. If we put our mouse on the image and click once, it will be enlarged according to its movement. When we click the image again, the image will be restored to its original size.

One more data which we have added to the online dictionary is Cyrillic equivalent of every headword and a link to Cyrillic Mongolian dictionaries. Cyrillic Mongolian is another written language of Mongolian people, which has been used as an official language in the country of Mongolia since 1940's. Cyrillic equivalent will be useful for those who are familiar with Cyrillic Mongolian both in Mongolia and in other countries. On the other hand, it would be regarded as one colloquial form of the traditional Mongolian headword, since Cyrillic spelling is based on the pronunciation of the modern *Khal-kha* dialect.

If we click on the Cyrillic equivalent, it will lead us to the website of “dictionaries of Cyrillic Mongolian” and the word will be in no time searched in it automatically.

Online dictionaries of Cyrillic Mongolian and others

Online dictionaries of Cyrillic Mongolian have been published on the website of CNEAS simultaneously with the Web version of “Mongolian-Chinese Dictionary”. The URL of the website is <http://hkuri.cneas.tohoku.ac.jp/p02/> (Figure 4).

There 3 dictionaries of Cyrillic Mongolian are available such as: (1) Mongolian explanatory dictionary (МОНГОЛ ХЭЛНИЙ ТОВЧ ТАЙЛБАР ТОЛЬ, 1966); (2) Mongolian-Chinese dictionary (ШИНЭ МОНГОЛ ХЯТАД ТОЛЬ, 1999); (3) Orthographical dictionary (МОНГОЛ ХЭЛНИЙ ЗӨВ БИЧИХ ДҮРМИЙН ЖУРАМЛАСАН ТОЛЬ, 2018).

Like the online dictionary of the “Mongolian-Chinese Dictionary”, we have put such information as: 1) entry words, idioms and examples in Cyrillic Mongolian script, 2) Mongolian explanations or Chinese translations, 3) page numbers, which are linked to the image data of the original page, and 4) traditional Mongolian equivalents, which are linked to the web version of “Mongolian-Chinese Dictionary”.

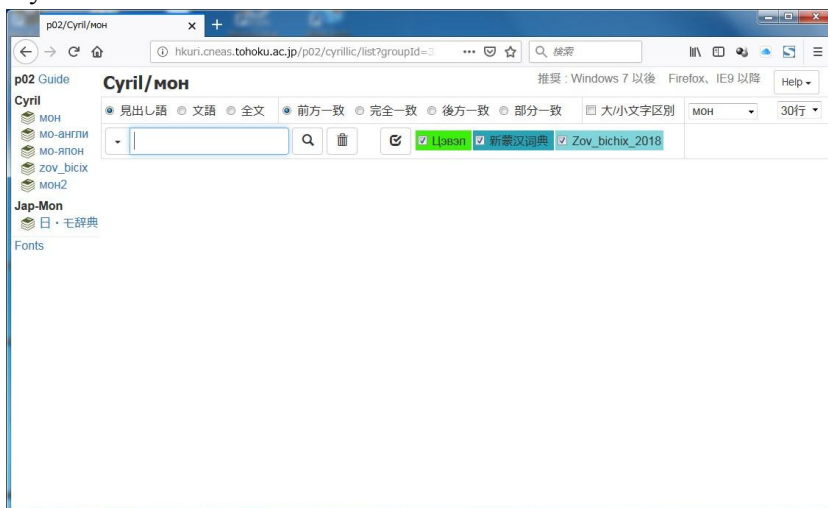


Figure 4. <http://hkuri.cneas.tohoku.ac.jp/p02/>

To every Cyrillic headword of the dictionaries, traditional Mongolian equivalent has been added, which is linked to “Mongolian-Chinese Dictionary: Web Version”. Clicking on it, the website of “Mongolian-Chinese Dictionary” opens and begins searching for the word in traditional Mongolian script. Using this function, we can search traditional Mongolian word through Cyrillic Mongolian dictionary.

When we search a word in this website, the IME of Cyrillic Mongolian is not necessary because typed letters are directly converted into Cyrillic letters in search field.

The keyboard layout for inputting Cyrillic Mongolian script is as follows:

а	а	л	l	х	h
б	b	м	m	ц	с
в	v	н	n	ч	х
г	g	о	o	ш	w
д	d	ө	@	щ	\
е	.	п	p	ъ	^
ё	/	р	r	ы	:
ж	j	с	s	ь]
з	z	т	t	э	e
и	i	у	u	ю	,
й	;	ү	y	я	q
к	k	ф	f		

Figure 5. Keyboard layout for inputting Cyrillic Mongolian script

The Cyrillic scripts are in conformity with the Unicode Standard, although the keyboard layout has nothing to do with it.

The Roman letters of Figure 5 coincide the imprinted letters on an English keyboard. Such a Roman letter on a keytop will be, when typed, automatically converted into Cyrillic Mongolian script in the search field.



Figure 6. <http://hkuri.cneas.tohoku.ac.jp/>

On the basis of our experience of developing the Web version of the “Mongolian- Chinese Dictionary”, we have made and published more than thirty online dictionaries such as Mongolian, Oirat, Mon-

guor, Dunshang, Dagur, Manchu, and so forth on the website of CNEAS. They are published together with the Web version of the «Mongolian-Chinese Dictionary» in 2015.

1. Cyrillic Mongolian dictionaries (<http://hkuri.cneas.tohoku.ac.jp/p02/>)

Besides 3 dictionaries of Cyrillic Mongolian stated above, such dictionaries as Mongolian-English, Mongolian-Japanese, Japanese-Mongolian are also available.

2. Oirat dictionaries (<http://hkuri.cneas.tohoku.ac.jp/p03/>)

Written Oirat words are spelled by Todo script.

3. Monguor dictionaries (<http://hkuri.cneas.tohoku.ac.jp/p04/>)

Monguor-Chinese dictionaries. Monguor words are spelled by Roman script.

4. Dagur dictionaries (<http://hkuri.cneas.tohoku.ac.jp/p05/>)

Dagur-Chinese dictionaries. Dagur words are spelled by Roman script.

5. Manchu dictionaries (<http://hkuri.cneas.tohoku.ac.jp/p06/>)

Manchu-Chinese dictionaries. Manchu words are spelled by Manchu script.

Online dictionaries and materials of the Mongolian (traditional, Cyrillic, Oirat, Monguor, Dagur, etc.) and Manchu languages are accessible from the website below: <http://hkuri.cneas.tohoku.ac.jp/> (Figure 6).

Notes

[1] The original title of the dictionary is as follows:

《蒙古语词典》
“mongyol kitad toli, nemen jasaysan debter”.
1999.

“mongyol kitad toli, nemen jasaysan debter”.
1999.

内蒙古大学蒙古学研究院蒙古语文研究所编《蒙汉词典（增订本）》内蒙古大学出版社，1999年

[2] The documents of Information technology are as follows:

1) 中华人民共和国国家标准 GB/T 26226-2010 信息技术 蒙古文变形显现字符集和控制字符使用规则 2011-01-14 发布, 2011-05-01 实施 (Information technology—Mongolian presentation forms character set and use rules of controlling characters. GB/T 26226-2010) 2011-01-14 (issued), 2011-05-01 (executed). P. 108.

2) 中华人民共和国国家标准 GB/T 25914-2010 信息技术 传统蒙古文名义字符、变形显现字符和控制字符使用规则 2011-01-10 发布, 2011-11-01 实施 (Information technology—Traditional Mongolian nominal characters, presentation characters and use rules of controlling characters. GB/T 25914-2010) 2011-01-10 (issued), 2011-11-01 (executed). P. 37.

3) 中华人民共和国国家标准 GB/T 26235-2010 信息技术 信息处理用蒙古文词语标记 2011-01-14 发布, 2011-05-01 实施 (Information technology—Mongolian word and expression marks for information processing GB/T 26235-2010) 2011-01-14 (issued), 2011-05-01 (executed). P. 13.