Definition of the Mongolian Character Codesets Enabling Multilingual Text Manipulation

Tomoko I. Kataoka*, Yutaka Kataoka*, Kazutomo Uezono†, Takeo Tatsumi*, Jun-ichi Yoshida‡, Katsuhiko Kakehi†, and Hiroyoshi Ohara†

* Centre for Informatics, Waseda University
† School of Science and Engineering, Waseda University
‡ School of Letters, Waseda University

Abstract

Mongolian related languages have been transcribed in different types of scripts which have complicated orthographies – such complexities have inhibited encoding them into codesets and text manipulation on computer systems. And it is essential to mix those scripts in both vertical and horizontal directions particularly for historic reasons. Analyses of the scripts in the world have clarified the definition of one *character* and its constructive information, which made it possible to assign the optimal character codeset(s) for Mongolian scripts having complicated orthographies. As a result, mixed texts of any scripts including Mongolian ones were able to be given generalized I/O and Text Manipulation as internationalization.

片岡 朋子*、片岡 裕*、上園 一知†、辰巳 丈夫*、吉田 順一‡、筧 捷彦†、小原 啓義†
* 早稲田大学 情報科学研究教育センター
† 早稲田大学 理工学部
‡ 早稲田大学 文学部

概要

蒙古(関連)語は、複数の異なるタイプの文字群で記述されてきた。これらの文字群は、極めて複雑な正書法を持つため、文字コード化及び計算機処理化が進んでいなかった。特にこれらの文字群は、歴史的経緯により、縦書きと横書きでの混在処理が必要とされる。全世界の文字を分析し、一文字の定義と文字を構成する情報を得た。これにより複雑な正書法をもつ蒙古文字に最適な文字コードを決定することができた。この結果、国際化の一環として、蒙古文字を含む全ての文字の混在文書の、単一的機構による一般化した入出力・文書処理が可能となった。

1. Introduction

True Internationalization (I18N) should not be equal to a mere collection of *Locales*. It means the simultaneous *consistent* mixture of all the scripts in the world – some of them are written horizontally from left to right, some others from right to left, and there even exists such a script written vertically from the bottom. Not only input and output but also text manipulation and communication must be provided for such international mixed script handlings. The Internationalized Multilingual I/O and TM/C Project of Waseda University, has been devoted for preparing the necessary environments for its realization.

To implement our system, all the world scripts with their writing conventions were analyzed, and essential information was discovered for defining one character of any type and assigning it a proper final glyph. Also the common necessary sets of writing conventions bound to specific scripts were clarified. Most parts of a writing convention is not language-specific (hard-coded) as believed, so the extent of possible language-independent mixed text handling was defined instead. Thus, clearly the definition of I18N became possible.

Here are reported the results of the research and implementation of the system done for numbers of scripts which had been used or are still used to write Mongolian and its related or neighboring languages to testify the further plausibility of our research. And certain codeset designs, as both character definable and glyph definable, are proposed for recommendation for the proper text manipulation and communication. Mongolian scripts have been believed impossible to extract a math rule set from, for its complicated orthography and script-sound ambiguities (See below). However, by the separation of information to define a glyph from the *Position dependency*, the problems were solved.

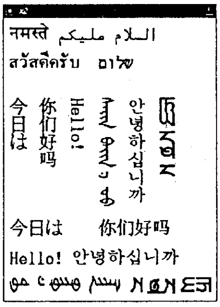
Scripts for writing Mongolian spread over vast areas even to Russia and to East Europe in the age of the Mongolian Empire. Not only their first script *Uighur-Mongolian* and its reformed *Mongolian* – still read and written today, but *Paspa*, *Todo* or *Soyombo* had been used since then. *Manchu*, the official literal language in the Ching Dynasty, and its descendent *Sibo* have their base on *Mongolian script*. As a result, very large amount of invaluable literature in those Mongolian related scripts (they are often written in several of those scripts mixed) is found almost anywhere in the world. That is why the appropriate encoding of those scripts is highly required.

The research covered the following scripts: Paspa, Soyombo, Uighur, Manchu, Sibo, Cyrillic (three types), Mongolian scripts used in Mongolia and in Inner Mongolia, China. All of them were analyzed and proposed suitable codesets, only some of which are discussed here on account of limited space. Great heritage of literature written in Paspa or Mongolian, and also in Manchu became to be handled properly by the system based on this extensive research.

2. Internationalization

As noted above, I18N is a simultaneous mixing of any number of scripts, which may be quite different from one another in such an aspect as writing direction or internal structure of a *syllabic*, with I/O and text handling consistencies. It must not be taken as *Multilingualism*, mixing various languages which necessarily involves language-specific information. Rather, I18N can be defined as the common basis for multilingualism of the highest level, handling any numbers, i.e., all, languages, and so must be provided with language-independent properties.

Practically, mixing of the scripts written horizontally and those written vertically should be realized without any inconsistencies. Language-independent, true international meta-writing convention, so to speak, the mechanism of putting characters in two dimensional plane of coordinates must be realized.



नमस्ते السلام عليكم

สวัสดีครับ שלום

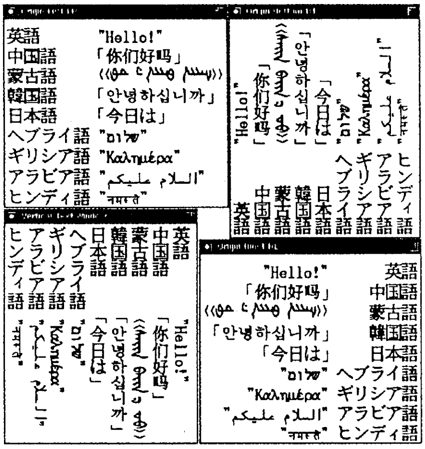
今日は 你们好吗 Hello! 안녕하십니까

今日は 你们好吗

Hello! 안녕하십니까

Figure 1: Mixing Different Writing Directions

When a character is actually written/displayed, it can vary its form according to the writing direction and its position in a word. Notice the shapes of the punctuation marks, *Direction-dependent* characters, in Figure 2. Even a direction-independent one can vary when mixed with the scripts in different directions. And the scripts also change depending on their position in a word, i.e., *Initial*, *Medial*, *Final* and *Independent form* as in *Perso-Arabic scripts*. Some scripts like Latin happen to have the same form for any position. Thus, all characters are defined as being both 1) *Direction dependent* and 2) *Position dependent* [1]. In other words, there is no predefined default origin: once origin/physical direction, etc. are specified by a user, other processes for drawing automatically follow [Figure 1].



英語　"Hello!"
中国語　「你们好吗」
蒙古語　《ᠰᠠᠶᠢᠨ ᠪᠠᠶᠢᠨ᠎ᠠ ᠤᠤ》
韓国語　「안녕하십니까」
日本語　「今日は」
ヘブライ語　"שלום"
ギリシア語　"Καλημέρα"
アラビア語　"السلام عليكم"
ヒンディ語　"नमस्ते"

Figure 2: Internationalized Writing System

Interprocess communication at any level should also be ensured. ISO 2022 [2] must be fully supported with necessary *extension methods* [1,3]. For proper basic text manipulation, character codeset must be able to define unambiguously one *character* and also its final *glyph* (Refer to the next section). Considering the repeated processes of character handling and communicating, it is a prerequisite that a mb (multibyte) codepoint string must be uniquely mapped into a WC (wide character) codepoint, ensured to be one character, and vice versa. Thus, a set of WC must be 'one' set of characters including all mb codesets without overlapping with themselves. Where the truly language-dependent information must be handled, the TMC (Text Manipulation Code) is available to play the role [1,4]. Each codepoint in each mb/WC/TMCs must be uniquely convertable with codepoint extension methods among mb, WC and TMCs.

3. Codeset Designs

Characters to be processed on a computer system are listed in character codesets, though one codepoint does not always correspond to one character – Non ISO extensions. TIS 620-2533:1990 [5] is one that defines parts of a character as distinct codepoints. Thus, a codeset should have rules for code extensions and is classified as *Character specifiable* or not. Also it must have ways to define a final glyph and is classified as*Glyph specifiable* or not.

GB 8045:87 for Mongolian scripts [6] defines both glyphs and glyph parts, together with character/glyph names. More than one characters which happen to have the same glyph are assigned the same codepoint. A certain character is generated by 3/9, or by the combination of 2/10 and 2/9 as well. It is a typical uncomputable codeset for one character, so it is impossible to utilize this for proper text manipulation.

Since a Glyph shape varies according to directions and positions, for a codeset to be ensured to be an I18n codeset, it must be *Name defined* instead of *Glyph defined.*

4. Mongolian Scripts and Mongolian Language

Mongolians have used varieties of scripts to transcribe their language, an agglutinating (containing roots which take a complex range of grammatical suffixes), basically SOV word-order language as is Japanese. In the early 13c. BC., they borrowed the Uighur script, originally written right to left, to establish their national vertical script under the influence of Chinese culture. This classic Uighur-Mongolian script, refined and some characters added later to be the so-called Mongolian script of the present form, has been favored over other more appropriate scripts designed to transcribe the language: Paspa, Todo or Soyombo.

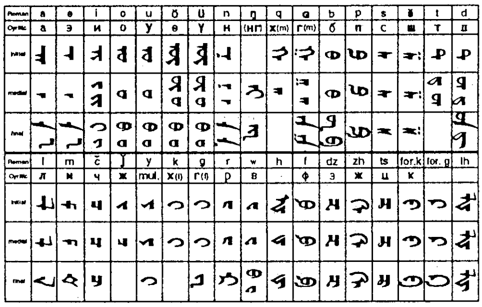

Roman
a e i o u ŏ ŭ n ŋ q ɑ b p s š t d
Cyrillic
а э и о у ө ү н (нг) х(m) г(m) б п с ш т д
initial
medial
final
Roman
l m č ǰ y k g r w h f dz zh ts for.k for. g lh
Cyrillic
л м ч ж mul. х(l) г(l) р в ф з ж ц к
initial
medial
final

Figure 3: Mongolian Script for Mongolia

Mongolian script is *phonemic* and *Position-dependent*, with basic initial/medial/final forms for each character. Certain endings are written separately as final forms after a short space, whose preceding character is also realized as a final, not a medial form. Notice that more than one characters happen to have the same glyph form. Thus, the medial forms for 'a', 'e' and sometimes for 'n' can cause ambiguities. However, native Mongolian syllable structure is V(owel) | C(onsonant)V | VC |

CVC, and a consonant character and a vowel character tend to appear alternatively in most syllables. It probably was enough for reading and writing to distinguish only when consonant sequences or vowel sequences occur, given certain vocabularies.

Alternative forms in medial and final rows are defined according to the information as, whether the preceding/ following character is a consonant/vowel or whether it is a member of a certain set of consonants which influence the shape of the following character, whether the word is monosyllabic, whether it is a foreign word, whether the suffix is in a certain set of case endings, etc. Which characters should be included was decided by the historical and practical factors: earlier additions of characters to transcribe Tibetan, Sanskrit or Chinese sounds as ‘p’ or ‘f’ have become familiar and necessary, and other characters added later and considered often used today are listed after ‘f’ in the table. As for the *Galik script* for Buddhism should be given a separate code from this, because its characters are not in daily use.

The proposed codeset is a Character definable one for text manipulation. The character ‘a’ and ‘e’, for example, are given distinct codepoints, although they happen to have the same glyph in the medial position. There are general rules; however, there are so many context and user dependent exceptions that it would be inefficient to write such rules for the automaton. Rather, style variation selectors can identify the forms for each character: where no form selection code is used, the upper form – if there is only one medial form, the form itself – is selected. F1 is used to select the initial form. Likewise, F2 is for the alternative medial form, F3 for the first final form variant, F4 for the second, F5 for the third, and F6 for the fourth final form, respectively.

Mongolian script has no ligatures in the true sense as Perso-Arabic. The so-called ligatures including ‘l’ or ‘m’ are that just a part of those characters extends to the preceding/following character space. And those with ‘b/k/g/f/p + vowel’ sequence do change shapes, while they have no ‘no-composed’ counterpart meaningful sequence unlike Arabic. Thus, there should be no codepoint for making ligatures. Intelligent Output Mechanisms (OMs) with composition functions can automatically draw these forms. Notice that the composition restriction codepoint is prepared for users who do not want to use composed forms. Input Mechanism (IM) is responsible for such identifications without taking a user’s time and patience.

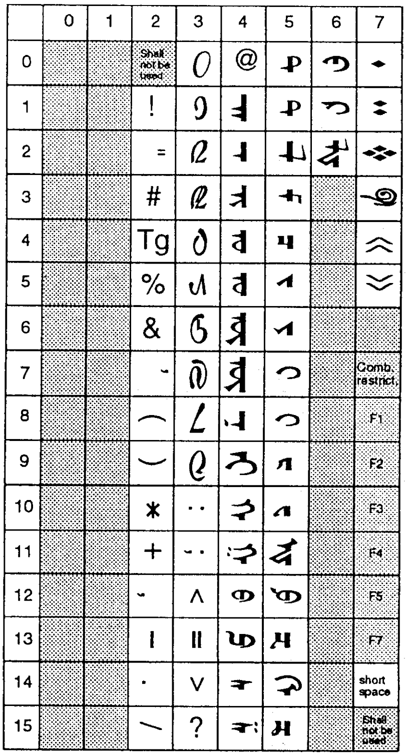

0
1
2
3
4
5
6
7
Shall not be used
Tg
Comb. restrict.
F1
F2
F3
F4
F5
F7
short space
Shall not be used

Figure 4: Codeset for Mongolia

On the other hand, people in Inner Mongolia have favored the pre-classic style, not the xylographic style often seen in outer-Mongolia and remained in Japan. They use the different writing convention from that in Mongolia: when spelling the sequences ‘n/G’(final) and final ‘a/e’, vowel script carries the dot(s) instead of the consonant one. Scripts for ‘t’ and ‘d’ are distinguished in different ways for the two, and also the numbers of the characters.

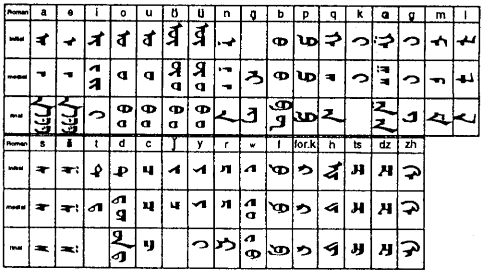

Roman
a
e
i
o
u
ö
ü
n
ŋ
b
p
q
k
ɑ
g
m
l
s
š
t
d
c
ǰ
y
r
w
f
for.k
h
ts
dz
zh

Figure 5: Script for Inner Mongolia

5. The Pre-classic Uighur-Mongolian Script

The Uighur script, phonemic, Position-dependent, with no distinct vowel characters – vowels were transcribed using 'dots', was borrowed from the Sogdians. When introduced for Mongolians, the script, originally written horizontally from right to left, turned 90 degrees to the left for writing vertically from the top. In this script the Chingis Khan Stone was inscribed and *The Secret History* was originally written.

6. The Todo (Oirat) Script

In 1648 Zaya Pandita reformed the Mongolian script to make it intelligible to the Oirats. The new alphabet gives distinct glyph shapes to the formerly ambiguous characters, such as 'a' and 'e', or 'o' and 'u', or 'x' and 'G'. It is still in use in Alashan and in Sinkiang.

7. The Manchu Script and the Sibo Script

Later in the Ching Dynasty, Manchu script was made based on Mongolian script and became the official literary language of China. Like the Todo script, it can differentiate every consonant and vowel characters using dots and shape change reforms. Note that it is for Manchu language, not for Mongolian. The script with six vowel characters cannot transcribe Mongolian completely.

8. The Cyrillic Scripts

The *Cyrillic script* has been used in Mongolia since 1946. It does not have enough number of characters to transcribe Khalkha-Mongolian, and Mongolians added two characters for rounded vowels. But still, velar /g/ and uvular /G/ are transcribed by the same Cyrillic character 'г'. Cyrillic script nor has a distinct 'y' character but five syllabics including [j] sound instead for seven 'y + vowel phonemic' sequences in Mongolian: another ambiguity. Even where one to one correspondence is kept between the Mongolian sounds and Cyrillic characters, some Mongolian sounds do not exactly match the Russian sounds. Therefore, this script is not ideal to transcribe the Mongolian phonological system.

There are two other Cyrillic systems used for dialects of Mongolian language: Kalmuck (Oirat) and Buriat. The numbers of characters, and even some sound values of characters differ among the three systems. Thus, it is impossible to unify these systems.

Mongolia: 35 characters

А а a	Б б b	В в b,(w)	Г г g, ɢ	Д д d	Е е je, jö	Ё ё jo	Ж ж dž	З з dz	И и i
Й й ĭ	К к (k)	Л л l	М м m	Н н n, ŋ	О о o	Ө ө ö	П п p	Р р r	С с s
Т т t	У у u	Ү ү ü	Ф ф (f)	Х х x	Ц ц ts	Ч ч tš	Ш ш š	Щ щ (štš)	Ъ ъ
Ы ы i	Ь ь	Э э e	Ю ю ju, jü	Я я ja					

Buriat: 36 characters

А а a	Б б b	В в (v)	Г г g	Д д d	Е е je	Ё ё jo	Ж ж ž	З з z	И и i
Й й ĭ	К к (k)	Л л l	М м m	Н н n	О о o	Ө ө ö	П п p	Р р r	С с s
Т т t	У у u	Ү ү ü	Ф ф (f)	Х х x	Һ һ h	Ц ц (ts)	Ч ч tš	Ш ш š	Щ щ (štš)
Ъ ъ	Ы ы ï	Ь ь	Э э e	Ю ю ju, jü	Я я ja				

Kalmuck (Oirat): 39 characters

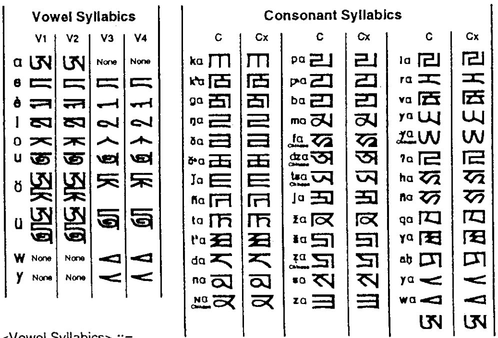
А а a	Ә ә ä	Б б b	В в w	Г г g	Һ һ ɣ	Д д d	Е е e	Ё ё (jo)	Ж ж (ž)
Җ җ dž	З з z	И и i	Й й j	К к k	Л л l	М м m	Н н n	Ң ң ŋ	О о o
Ө ө ö	П п p	Р р r	С с s	Т т t	У у u	Ү ү ü	Ф ф (f)	Х х x	Ц ц ts
Ч ч tš	Ш ш š	Щ щ (štš)	Ъ ъ	Ы ы ï	Ь ь	Э э e	Ю ю ju	Я я ja	

characters enclosed by '(' and ')' are for foreign words.

Figure 6: Three Cyrillic Systems

9. The Paspa Script

The *Paspa script* is the one that the Khubilai Khan of the Yuan Dynasty ordered Paspa to make as the national/international script. He designed the script based on the *Tibetan script* (*conjunct-syllabic*) to make it viable as an I18N script enable to transcribe foreign languages. It has enough number of characters to represent the sounds and syllable structures even of the other languages than Mongolian, and still in use in Buddhist temples in Mongolia, Inner Mongolia and Tibet. The Tibetan script shows the syllable boundaries by Tseg's, while Paspa by the connections of characters.


Vowel Syllabics
V1
V2
V3
V4
a
e
ė
i
o
u
ö
ü
w
y
None
Consonant Syllabics
C
Cx
ka
kʰa
ga
ŋa
ča
čʰa
ǰa
ña
ta
tʰa
da
na
NA Chinese
pa
pʰa
ba
ma
fa Chinese
dza Chinese
tsa Chinese
ja
ža
ša
ẓa Chinese
sa
za
la
ra
va
ya
ɣa Chinese
ʔa
ha
ña
qa
γa
ab
ya
wa

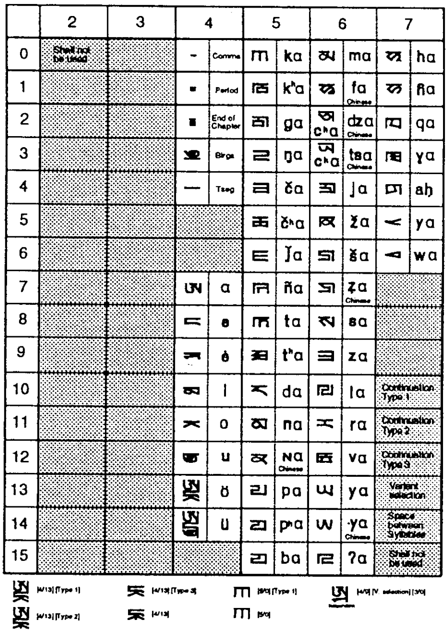
<Vowel Syllabics> ::=
<V1> | <V2> | <V3> | <V4>

<Consonant Syllabics> ::= <C> | <Cx>

<Basic Syllabics> ::= <V1> | <C> | <V2> <C> | <Cx> <V3> | <Cx> <V4> <C>

Figure 7: Structure of Paspa

In Paspa, unlike Tibetan, the connection conventions of glyphs within a syllable vary according to languages, for it is used to transcribe mixed texts of different languages. No boundary symbols between syllables can define the glyph shape automatically. Thus, it shows the four variants of glyphs after the codepoint, with a connection type combined. It also has the codepoint to define which variant to select.



	2	3	4		5		6		7	
0	Shall not be used		-	Comma		ka		ma		ha
1			■	Period		kʰa		fa Chinese		ɦa
2			■	End of Chapter		ga	cʰa	dza Chinese		qa
3				Birga		ŋa	cʰa	tsa Chinese		ɣa
4			—	Tseg		ča		ɟa		ah
5						čʰa		ža		ya
6						ǰa		ša		wa
7				a		ña		ẓa Chinese		
8				e		ta		sa		
9				ė		tʰa		za		
10				i		da		la	Continuation Type 1	
11				o		na		ra	Continuation Type 2	
12				u		Na Chinese		va	Continuation Type 3	
13				ö		pa		ya	Variant selection	
14				ü		pʰa		·ya Chinese	Space between Syllables	
15						ba		ʔa	Shall not be used	

[4/13] [Type 1] [4/13] [Type 3] [5/0] [Type 1] [4/0] [V. selection] [3/0]

[4/13] [Type 2] [4/13] [5/0]

Figure 8: Codeset for Paspa

10. Summary

All Mongolian related scripts were researched in the historic order and were encoded into character codesets. The principled distinction of information for specifying a character itself and that for specifying its glyph articulated the roles of conversion from mb to WC and that of OM which selects a glyph. By these analyses, writing conventions which are not related to specific languages were determined to mix all scripts as internationalization – Mongolian related scripts need to be mixed, as found in Buddhist and other texts.

The researches above also brought large information, with clearer view of those Mongolian related languages provided, to process them as natural language processing. Especially, influences among languages and among scripts are important for language education, making sorting orders and database accessing codes. We started making codesets for all Brahmi derived scripts and older Perso-Arabic scripts.

11. Acknowledgements

We thank so many people who helped us use those scripts. Special thanks go to Professor H. Kuribayashi.

References

[1] Kataoka, Y. et al. Codeset Independent Full Multilingual Operating System: Principle, Model and Optimal Architecture, SIG System Software & OS, IPSJ, 68-4, March 1995, pp. 25–32.

[2] ISO/IEC 2022: 1986, Information processing – 7-bit and 8-bit coded character sets – Code extension techniques.

[3] Uezono, K. et al. The Worldwide Multilingual Computing (2): Functions, Model, Design and Architecture of Multilingual I/O TM/C System, Proceedings of the 51th General Meeting of IPSJ, Vol. 3, September 1995, pp. 247–248.

[4] Kataoka, T. et al. The Worldwide Multilingual Computing (4): Essentials for the Multilingual Text Manipulation, Proceedings of the 51th General Meeting of IPSJ, Vol. 3, September 1995, pp. 251–252.

[5] TIS 620-2533 (1990), Thai Character Codes for Computers, Thai Industrial Standards Institute, Ministry of Industry, Thailand.

[6] GB 8045-87 (1987), Mongolian 7-bit and 8-bit coded graphic character sets for information processing interchange.

[7] Kataoka, Y. et al. The Worldwide Multilingual Computing (1): Essentials, Principles and Scope Covering All Characters in the World, Proceedings of the 51th General Meeting of IPSJ, Vol. 3, September 1995, pp. 245–246.

[8] Tanaka, T. et al. The Worldwide Multilingual Computing (3): An Implementation of the Multilingual I/O TM/C System and Waseda X11, Proceedings of the 51th General Meeting of IPSJ, Vol. 3, September 1995, pp. 249–250.

[9] Oya, Y. et al. The Worldwide Multilingual Computing (5): Multilingual Text Manipulation and Text Widget, Proceedings of the 51th General Meeting of IPSJ, Vol. 3, September 1995, pp. 253–254.

[10] Daikokuya, H. et al. The Worldwide Multilingual Computing (6): Multilingual Text Interprocess Communication and Input Mechanism, Proceedings of the 51th General Meeting of IPSJ, Vol. 3, September 1995, pp. 255–256.