形码方案的分析及现有编码方案的改进

Badral Sanlig badral@bolorsoft.com
Jamiyansuren Togoobat onoltmn@yahoo.com
Bolorsoft LLC
http://www.bolorsoft.com

蒙古：乌兰巴托

摘要：关于形码方案和音码方案，学界都有相关研究。每个方案都有缺点和优点。本文主要包含三部分。第一部分，介绍蒙古语的历史信息；第二部分，介绍分析纯粹的形码方案；第三部分，提出音码的改进方案。了解语言和文字信息对我们做出决定有重大影响，因此第一部分即导言会介绍蒙文的基本情况。现有的形码方案并非纯粹的形码方案，它包含一些音码方案的因素，因此第二部分介绍分析纯粹的形码方案。对现行的音码方案提出或多或少的改进升级意见。

1. 导言

13 世纪开始，蒙古文就作为官方文字通行于世。迄今我们所知的最早的蒙古文碑刻写于 1224 或 1225 年。最早的蒙古语法书可以追溯到 13 世纪，如元代高僧搠思吉斡节儿(Saskya Pandita Kun dga’ rgyal mtshan，1182—1251)写成的《蒙文启蒙》（一译《心箍》）(Jirüken-ü tolta)和却吉奥斯爾写成的《蒙文启蒙》(Choiji-Odsar，1307-1321)。但比较遗憾的是，这些作品都没能流传至今。比较幸运的是，18 世纪及之后的对《蒙文启蒙》（即《心箍》）的相关的注释文本流传于世，其中最早最著名的是 18 世纪的丹赞达格巴于 1723-736 年所作《蒙文启蒙诠释-清除错字之苍穹玛尼经》（Jirüken-ü tolta-yintayilburi üsüg-ün endegürel-i arilγaγci Oγtarγui-yin mani，英译 The Space Jewel for Eliminating of Letter Ambiguity: Commentary on the Heart Essence）。丹赞达格巴是乌珠穆沁蒙古一位著名的转世喇嘛。通过这些文献，我们可以了解体现蒙古正字法的蒙古文是如何起作用的。

在此，我将在丹赞达格巴《蒙文启蒙诠释》的基础上介绍蒙古文正字法的专业术语。其中一个重要原因就是我将此书作为解构蒙古文字的重要来源。

我将以《蒙文启蒙诠释》作为研究对象，扼要论述蒙古文语法的重要组成部分——元音、辅音、音节以及闭音节的五个特点。

历史上，蒙古的字母表发生了几次变化。现通用的蒙文字母包含七个元音和 28 个辅音。七个元音分别是 a, e, i, o, u, ö, ü。基本辅音有 na, ba, qa (ke), γa (ge), ja, ya, ta, da, ma, ča, ra, sa, ša, la, wa , pa, ang and lha。以上辅音用于书写标准蒙古文。还有一些辅音 fa, za, ca, ža, ka, ha, Zhi, Chi 是在书写外语时候用的。在蒙古文中 o 和 u; ö 和 ü 是一个意思的不同写法。但是《蒙文启蒙》中把它们列为完全不同的字母，这在丹赞达格巴《蒙文启蒙诠释》可以体现出来。丹赞达格巴认为 o 和 u;

ö 和ü就是完全不同的字母。下文我将举出三个例子来证明此观点。

1）“……a 生成 o 和 u, e 生成 ö 和ü。Na 生成 no 和 nu，ne 生成 nö 和 nü。Ba 生成 bo 和 bu，be 生成 bö 和 bü”。（丹赞达格巴，P.6r）。文中，他尝试解释蒙文“肚子（gedesü）”或者“腹部”，这个单词有两个不同的字母 o 和 u。但是他为了表达这个单词，写了两边 O。在解释 ö 和 ü 的用法时，他也是这么做的。O 和 u 与 ö 和 ü的不同之处就在于前者与后元音一致，后者只与前元音一致。然而，蒙文“肚子”（gedesü or belly）同样使用字母 O 来表示 ö 和 ü，不过我们可以通过前元音和后元音的不同来分辨不同的字母。

2）“……元音有 a, e, i, o, u, ö 和 ü。由于它们在拼写的时候放在辅音的前面，因此被称为首字母 aq-a üsüg 或者母字母 eke üsüg”（丹赞达格巴，P.7a）。这里，丹赞达格巴用七个字母表示七个元音，而不是五个。为了表达出来自己的意思，他不得不使用四个 O。

3）“……字母 š 与元音一起出现，构成 ša, še, ši, šo, šu, šö 和 šü”（（丹赞达格巴，P.6r）。此处，他展示了辅音字母 š 与七个元音不同的组合。如果他真的考虑 o 和 u, and ö 和 ü 是独立的字母，他会只描述五个音节（如ša, še, ši, šu, šü）。但是他用了七个音节，表明他认为即使这些元音使用一个字母来显示，它们代表的也是各自独立不同的字母。

《蒙文启蒙》强调的另外一个蒙语的规律，那就是元音和谐律。丹赞达格巴认为，蒙古语中存在三种类型的元音。

1）代表后元音的阳性元音（er-e 或 čingγ-a egesig）—— a,o,u。

2）代表前元音的阴性元音（em-e 或 köndei egesig）——e, ö, ü。

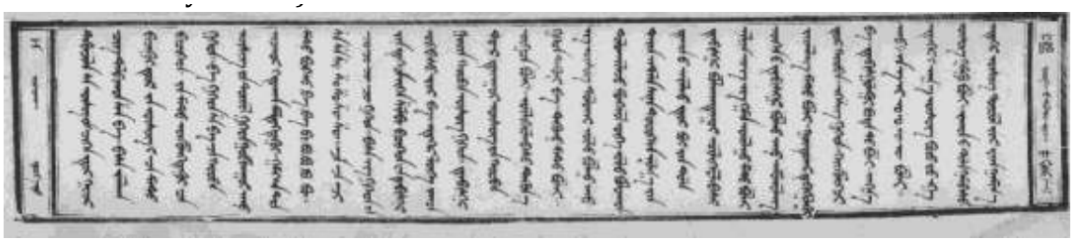
3）中性元音（saγarmaγ egesig）—— i。

简单的说，元音和谐律指得是一个词语要么包含后元音（a,o,u）要么包含前元音（e, ö, ü），但是不可能同时包含两种，即一个词语的后缀一定会跟派生词的元音和谐。元音 i 是中性，所以前元音词和后元音词都可以出现 i,但是当 i 出现在词中得每一个音节中时，这个词就被认定为前元音词。元音和谐律还影响其他字母，γ/q 和 g/k，前者只出现在后元音词中而后者只出现在前元音词中。

辅音一般被分为两类。闭音节辅音，位于词中位置后面可以再跟一个辅音或者出现在词尾位置。还有一种是元音后辅音。据丹赞达格巴统计，音节尾缀辅音有 11 个。

如，《蒙文启蒙诠释》记载“阳性闭音节的辅音总共有 11 个，分别是 an, ab, aγ, am, al, ar, as, ad, ay, aw 和 ang.……阴性闭音节的辅音有 en, eb, eg, em, el, er, es, ed, ey 和 ew”（p. 7r）。现代蒙文中，这些闭音节辅音中的 w 和 y 转换成了元音 u 和 i。一个典型包含 y 的词 nayma （八）。不同于蒙文标准双元音写法(ᠤᠢ)，这是一种比较独特的写法，它只写一个长齿或者斜线(ᠢ)。其他与闭音节 w 相关的词也有例子如 keüked （孩子），taulai (兔子)，teüke （历史）等等。

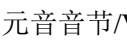
闭音节 ng 比较特殊，它不出现在词首的位置（也就是后面不接元音）。

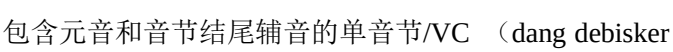


丹赞达格巴解释道“ng 字母不出现在词首位置，只出现在描写新生儿儿哭泣的声音的词中。因此，它被认为是一个闭音节辅音。如，宝宝呜呜/哇哇大哭 ing ng (ᠢᠩ ᠩ)”。

这个特点并不会出现在 pdf 文档蒙文辅音元音序列-Weizhe171209.pdf 中，因为这里面 ng 辅音后面接了元音。看一眼，我们就知道此文作者不希望在他的方案中出现连写模式。但是，这个方案因为 ng 辅音的特性而失败了。

《蒙文启蒙》中全面详细描述了音节结构，比如，

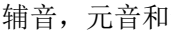
元音音节/V：  (a, e, i, o, u, ö, ü)。

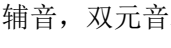
包含元音和音节结尾辅音的单音节/VC （dang debisker üy-e): 

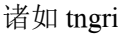
 。

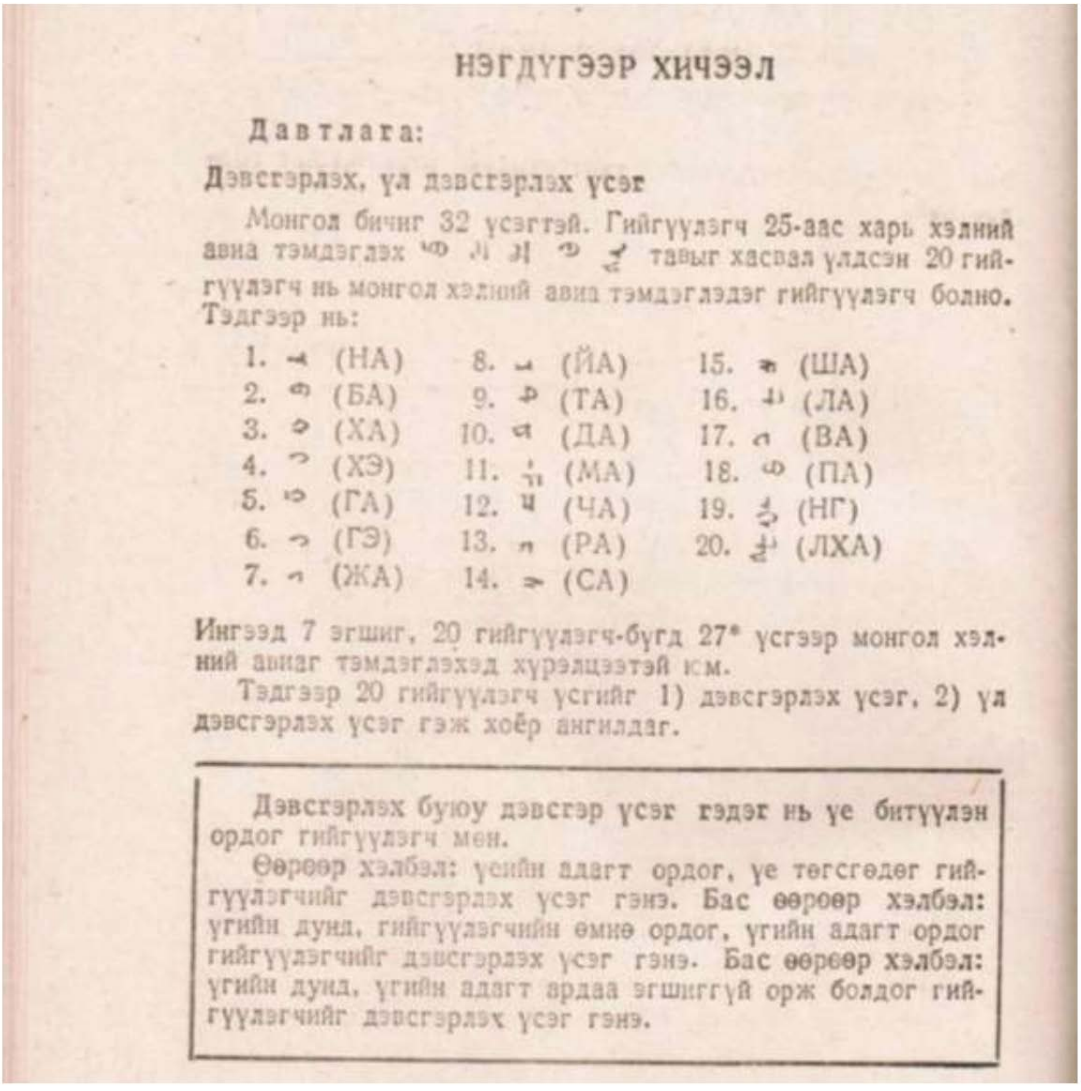
双元音音节和音节尾缀辅音（dabqur debisker）/ VYC： 

辅音和元音/CV:

辅音，元音和音节尾缀辅音/CVC： 

辅音，双元音和音节尾缀辅音/CVYC： 

诸如 tngri  （天空）和*gšan* (时刻)不包含在音节结构之中。



НЭГДҮГЭЭР ХИЧЭЭЛ

Давтлага:

Дэвсгэрлэх, үл дэвсгэрлэх үсэг

Монгол бичиг 32 үсэгтэй. Гийгүүлэгч 25-аас харь хэлний авиа тэмдэглэх тавьж хасвал үлдсэн 20 гийгүүлэгч нь монгол хэлний авиа тэмдэглэдэг гийгүүлэгч болно. Тэдгээр нь:

1. (НА)	8. (ЙА)	15. (ША)
2. (БА)	9. (ТА)	16. (ЛА)
3. (ХА)	10. (ДА)	17. (ВА)
4. (ХЭ)	11. (МА)	18. (ПА)
5. (ГА)	12. (ЧА)	19. (НГ)
6. (ГЭ)	13. (РА)	20. (ЛХА)
7. (ЖА)	14. (СА)	

Ингээд 7 эгшиг, 20 гийгүүлэгч-бүгд 27* үсгээр монгол хэлний авиаг тэмдэглэхэд хүрэлцээтэй юм.

Тэдгээр 20 гийгүүлэгч үсгийг 1) дэвсгэрлэх үсэг, 2) үл дэвсгэрлэх үсэг гэж хоёр ангилдаг.

Дэвсгэрлэх буюу дэвсгэр үсэг гэдэг нь үе бүтүүлэн ордог гийгүүлэгч юм.

Өөрөөр хэлбэл: үеийн адаг ордог, үе төгсгөлөг гийгүүлэгчийг дэвсгэрлэх үсэг гэнэ. Бас өөрөөр хэлбэл: үгийн дунд, гийгүүлэгчийн өмнө ордог, үгийн адаг ордог гийгүүлэгчийг дэвсгэрлэх үсэг гэнэ. Бас өөрөөр хэлбэл: үгийн дунд, үгийн адаг ардаа эгшиггүй орж болдог гийгүүлэгчийг дэвсгэрлэх үсэг гэнэ.

乌兰巴托 1986 年的蒙古文学校教材中把 Ga,Qa,Ge,He 列为各不相同的字母。

2. 形码的分析

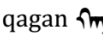
N4889 上提交的形码方案并非一个纯粹的形码方案。因此，我们把它称为半形码方案。

半形码方案

在 N4882 会议上，我们详细列举了半形码方案的好处：

1. 更清晰，文本没出现二义性

但其实文本中还是有二义性的。

 qagan  qannan

 /aagad/, /aannad/, /aagaon/, /aannaon/

 /aaagagur/, /aqgagur/, /aaagnannor/

2. 更简便的字体安装，但是必须符合当地上下文规则。

我们承认这个问题。不过我们只要对现行方案稍加改造就可解决此问题。

3. 现代蒙文不需要变体序列 （除了为反向兼容性支持旧蒙文）。

此问题的确存在。我们可以大量减少现有方案中的 FVS 数量。一个 FVS 字符也是够用的。

4. 更多简单的用户体验；看见什么字就能打什么字。

这是不正确的。如到   。除非避免位置变体，否则这个问题会一直存在。

5. 更简单的搜索功能。

这个问题无法避免除非能消除视觉上的二义性。这个看得见的二义性不像音码方案的那种，它是因为方案中的混合编码形成的。如，详情可见点 1。

6. 支持国际域名。

国际域名安全问题依旧存在。

7. 比当前方案更安全，安全隐患大大减弱。

安全隐患依旧存在。

这个方案的缺点及不足之处：

1. 混合（字形和字母）编码引发更多复杂问题修补。一个蒙文用户既可以拼对蒙文，也可以想到纯粹的字母组成。如，“öndör”是蒙文“高”（ᠥᠨᠳᠥᠷ）的正确拼法，但是可能会因为编码被拼成“aueadur”。几乎所有的单词都会拼错。

2. 存在形码方案中其他误用情况。

移动问题

形码方案中的移动问题在 N4890 中反映过。编码移动的问题从来都不简单。共存状况不能够消失。蒙文中西里尔字母一直存在很大问题。在统一码之前，我们使用 win1251 给蒙古西里尔字母编程。两个字母每个字母有两个变体，比俄文多。2000 年至今，共存问题依旧存在。美国国家标准块和西里尔块(0400-04FF)依然存在很多不正确的字体，有 Ө,Y。

综上，半形码方案是不可接受的。

以下，我们分析纯粹的形码方案。

纯粹的形码方案

图 1 所示的老打字机是我们理想中的纯粹的形码方案。

图 1

这种形码方案直到 90 年代中期还在用。

每个蒙文都有固定的笔画（不是字形）。

图 1 是传统蒙文打字系统。

字母成分

根据蒙古文语言学书籍，字母的各个成分有专门的术语。这些术语经常与字母的组成部分或者字形的位置变体有关。这些教学专业术语分成以下三组：字母成分或字母组成部分，闭音节辅音和附属字。

字母基本有十一种成分(mo. ᠵᠦᠷᠠᠮ zuram)。所有这些名字都是由形状和成分位置派生出来的。

- ᠠᠲᠰᠠᠭ Atsag / 前面， 牙

A, E, NA, GA, DA, NG, MA, LA，水平画下来的笔画，一般被称为字牙。

- ᠲᠢᠲᠢᠮ Titim / 头

通常，字母分为有字头与无字头两种（衬线与无衬线类比）。字头风格是可辨识的，通过七个元音，一些辅音还有后缀的起始部分。过去曾有弱化辅音的作用。这一成分不包含任何语音信息。19 世纪初，有字头概念，表示辅音弱化。

- ᠱᠢᠯᠪᠢ Shilbi / 长齿

字母 I,JA,YA 的斜线是一个长条，也是阴性元音的标志。

- ᠡᠪᠡᠷ ᠭᠡᠵᠡᠭ Ever, Gezeg / 角

字母 MA,CA,ZA,LA 中使用的发音符号。

- ᠰᠤᠤᠯ Suul /尾

长笔画或者短笔画，经常在单词词尾。

- ᠭᠡᠳᠡᠰ Gedes / 边

用一条弧线形成内部空间，如 O, U, OE, UE, BA, PA。

- ᠨᠤᠮ Num / 弓

一条开放的弧线，如在 KE, GE, NG, BA, PA 和 FA 中。

- Nuruu /干

这个术语有两个不同的意思。第一个，整个单词都依靠这一条线贯穿。第二个，嵌入两个字母中间的一个空间。

- Zavj /

一条开角线，如 SA 和 SHA 就有。

- Zartig /

耳形的一部分。

- Tseg, dusal / 点

变音符号的一点，N, G, Sh 中使用。

- Shavj / 终

终结。这个术语是印刷业专门发明，以方便字体使用。

有些 Zuram 可以被称为字形的位置外形。比如长尾 A 是 A, E 和 N 的词尾外形。字头 A 是 E 的起始外形。

为了方便认识蒙文字母的字形，以上选择的是学校已经普及的。

Lubsanbaldan, 1972, x. 209–219; Лувсанбалдан, 1975; Кара, 1972, x. 41; Мижиддорж, 1976; Шагдарсүрэн, 1975; 1984; 1987]. Нэр томъёог ерөнхийд нь зурлагын нэр, авиа дуудлагын нэр, дагавар болон сул үгийн нэр гэсэн гурван том бүлэгт хувааж болно.

Зурлагын нэр

1. *ačuγ* (ᠠ) — ачаг
2. *sidü* (ᠰ) — шүд
3. *örgesü* (ᠥ) — өргөс
4. *niruγu* (—) — нуруу
5. *γoul* (—) — гол
6. *silbi* (ᠢ) — шилбэ
7. *em-ün temdeg* (ᠦ) — эмийн тэмдэг
8. *urtu sidü* (ᠤ) — урт шүд
9. *segül* (ᠦ) — сүүл (ᡐ үсэгт)
10. *degegsi ebertei silbi* (ᠡ) — дээш эвэртэй шилбэ
11. *doγuγsi ebertü silbi* (ᠣ) — доош эвэртэй шилбэ
12. *eteger silbi* (ᠡ) — этгэр шилбэ
13. *erteger silbi* (ᠡ) — эртгэр шилбэ
14. *yatuγar silbi* (ᠠ) — ятгар шилбэ
15. *erbegelǰn silbi* (ᠣ) — эрвээлжит шилбэ
16. *gedesü* (ᡤ) — гэдэс
17. *oduγudu* (ᡩ) — ходоод
18. *baγa oduγudu* (ᡩ) — бага ходоод
19. *yeke oduγudu* (ᡩ) — их ходоод
20. *bitegü* (ᡤ) — битүү
21. *γoγčuγa* (ᡤ) — гогцоо
22. *geǰige* (ᡤ) — гэзэг
23. *eber* (ᡝ, ᡞ) — эвэр
24. *degegsi eber* (ᡞ) — дээш эвэр
25. *doγuγsi eber* (ᡞ) — доош эвэр
26. *titem* (ᡨ) — титэм
27. *qayarqai toluγai* (ᡴ, ᡴ) — хагархай толгой
28. *angarqai toluγai* — ангархай толгой
29. *aγsabur* (ᠶ) — агсвар
30. *numu* (ᠨ) — нум
31. *ǰabaǰi* (ᠵ, ᠵ) — завьж
32. *eretü* (ᠷ, ᠷ)— эрүү
33. *aru-yin sa* (ᠠ, ᠠ) — арын са[1]
34. *öbür-ün sa* (ᠥ, ᠥ) — өврийн са[2]

[1] Ихэвчлэн бурнад уламжлалд хэрэглэдэг нэр.
[2] Ихэвчлэн бурнад уламжлалд хэрэглэдэг нэр.

图 2

若要全部蒙文字母包括阿礼嘎礼字母都能数字化，纯粹的形码方案只要需要 100 个

字符/成分显示在键盘上。

还有二义性的问题，我们得出的结论是对于蒙古文字，我们不能抛弃二义性。

我们不能用此方案储存并传输蒙古语言文字信息，在加上可用性和文本处理问题，我们认为纯粹的形码方案会很快地被淘汰。

因此，纯粹的形码方案也是不可行的。

3. 对现行方案的改进

音码方案的分析

正如标题一提到的，蒙文毫无疑问是用音码方案。13 世纪蒙文语言学家却吉奥斯爾(Choiji Odsar)写成的《蒙文启蒙》，为当代蒙文的起源，其借鉴的就是回鹘文字。虽然《蒙文启蒙》没有流传于世，但是 18 世纪的丹赞达格巴于 1723-736 年所作《蒙文启蒙诠释-清除错字之苍穹玛尼经》（Jirüken-ü tolta-yintayilburi üsüg-ün endegürel-i arilγaγci Oγtarγui-yin mani，英译 The Space Jewel for Eliminating of Letter Ambiguity: Commentary on the Heart Essence）依旧流传于世。通过此书，我们发现当时的蒙古文就存在音和形方面的问题，这一问题也困扰着现有的音码方案。

因此，我们继续研究音码方案的问题，对此，我们依旧在让步。

为什么说在不久的将来，音码方案几乎是难处理的？

1. 不正确且 FVS 功能太多，这会导致不正常的用户体验；
2. 不均匀，无明确说明的字体安装；
3. 混乱的字体规则和一些缺失的字符。

因此，消除在 N4882 上提到的现行方案的所有缺点是非常有必要的。

编码问题

我们仔细检查了蒙文板块，从最初的标准到现在的统一码 10.0 标准，还查看了技术报告 170，来判断统一码 3.0 中提到的最初方案是否是错误的。

我们发现一些关键问题或者错误，如对在蒙文中重点控制元音和谐律的蒙文字母 QA,GA 的错误编码。现在，我们朝完全相反的方向努力，即我们经常尝试通过字体规则决定文本中的这些字母。事实证明，这是不可能的。即使我们定了很多规则，现在我们依然无法确定这些字母是阴性还是阳性。

现在，我们可以通过严格区分蒙文字母 QA - QE, GA – GE 来大大减少规则的复杂性及其文本相关性。

- 窄式不换行空格并无准确的定义且并无达到预期。见 L2/17-036 方案。
- 在现有标准的版本中，风格字符和频率字符都有编码。这些都必须清除。
- 蒙文板块有一些缺失的字符
- 阿礼嘎礼蒙文板块缺失字符表
- 托忒蒙文板块缺失字符。

自由变体选择符

对用户来说，在输入法中认清控制字符让人非常迷惑且不好打字。我认为，我们只需在输入法中设置一个自由变体选择符。关于字符变体选择，我们有一个很好的例子，九十年代，有一个很好的打字软件（磁盘操作系统中），叫 Sudarch。这个编辑器上有一个自由变体选择符，可应用于所有的变体。

如，类比现在的统一码，FVS1 就用一个 FVS，FVS2 就是用两个连续的 FVS，FVS3 就是用三个连续的 FVS。优点有：

a) 用户无需在键盘上到处找 FVS 键的位置。

b) 用户可通过键入 FVS 或回格键就可直观的看到变体使用的是哪个变体键（1，2，3 等等）。

窄式不换行空格

窄式不换行空格引发了许多问题。首先，它无后缀连接符功能。因为除了微软，几乎所有平台都不支持这个字符。大部分系统中，窄式不换行空格被空格 SPACE（0020）取代。这个很容易证明。比如，你可以在脸书中输入或者尝试将文本中窄式不换行空格复制粘贴到脸书上。

现在，用户可使用空格和词中连接符（nirugu）来正确显示后缀。

缺失的字符

引号（Qutationmarks）

引号，我们使用拉丁字符《》，因为我们不使用中日韩统一表意文字(CJK Unified Ideographs)字体。可是问题是《》这个符号不美观，且不集中。所以我们需要加上这个标点。其他的标点我们也使用拉丁字符来表示，只是它们的位置不集中在中间。难道我们要把所有需要的拉丁字形都重新绘制一遍吗？设计字体，我们需要明确的指导方案！

第一个字母的缩写符

增加一个连接的停止符，在其后可以写一些缩写字母，如人的姓等等。

复合词加入符

我们需要多个控制符来拼写诸如 Batumungkhe (), Ueruntuyaga ()和 GereltOd ()这样的复合词。这种方式是为了防止出现 GerelTod ()这样的困扰。之前，我们使用一个长词中连接符来区分复合词。现在，使用两个词中连接符和几个 FVS 就可能实现，但是从用户角度来看，这样操作很麻烦。如果有一个词中连接符可以充当一个有效的音节字

符的话，还是可能的。

输入设备

我们需要向用户展示，如果要将一些隐形字符如 FVS 键放到输入设备上，我们需要多次键入？

现在，用户不知道他们输入隐形字符需要点击多少次。

这将在会议上向大家展示。

字体

现行编码的主要问题是字体规则太多导致编码不稳定。我们在现行编码方案中可以找到很多这样的例子。字母QA,GA的计算机辅助语言教学规则，不稳定的FVS如 /bicigVFS1/,  /bicig/。

稳定 FVS 最好的办法就是我们为字体和绘制制定一个标准的规则。

详情见附件。

改进的音码方案

推荐现行音码使用的升级可以通过不同方式体现。

小改进：

1.修补了位置不合的问题

2. 在输入设备上自由变体选择符的数量减少至 1 个。

3. 消除格式和频率字符，重组阿礼嘎礼蒙文板块的一些变体。

4. 对蒙古字母 QA 和 GA 分别编码，配合 FVS1 使用。

5. 混杂字体规则的标准化。

6. 绘制规则的标准化。

7.不做彻底改变。

大升级：

小改进之外的大升级也是必要的。

1.对蒙文字母 QA（182C）按照阴性和阳性分别编码。

2.对蒙文字母 GA（182D）按照阴性和阳性分别编码。

对于能满足我们要求的技术，我们必须考虑成本效益，效率和寿命长短再做决定。我们是考虑下一个 20 年还是 200 年的技术呢？

我们是储藏和传输物理真实语言信息还是高度概括显形信息？

综上所述，我们认为现行的方案是统一码 Unicode 最出彩的编码方案。