实现500万词级《现代蒙古语文数据库》*的主要措施

华沙宝

（内蒙古大学蒙古语文研究所　呼和浩特 010021）

摘　要： 本文着重介绍了用ASCII转写蒙古文方正编码文本的软件－MTOA。迄今为止，所有蒙古文编码系统都没有以字母作为唯一的收字单位，同形字母都没有分开编码。MTOA找出了方正文本中每一个蒙文单词的字母间隔并确认每一个字母的读音和书写特征。它不仅解决了500万词级《现代蒙古语文数据库》的语料来源和获得途径，而且为使电子出版业专用的大量的蒙文文本变成信息处理多种用途可以共享的资源提供了条件。

关键词： 蒙古文　语料库　字母独立间隔　转写

A main measure to build the Mongolian corpus including 5 million words

Huashabao

(Mongolian Language Institute Inner Mongolia University Huhhot 010021, China)

Abstract: This paper will introduce MTOA, a software which transcribes a text recorded under Founder Mongolian encoded into ASCII text. Up-to-date, no one Mongolian encoding system has used only the letter as coding unit and no one has given different codes for letters which have the same glyph. The MTOA has found out the boundary between letters in a word included in the Founder Mongolian text and determined pronunciation and orthography of the letter. So, a method to get resources for the Mongolian corpus is supplied by the MTOA and the changing of a lot of proper texts for Mongolian electronic printing system into resources of Mongolian information processing shared by various purposes becomes possible.

Keyword: Mongolian corpus, character boundary, transcription

一、引言

1991年12月，内蒙古自治区科委和社会科学规划委员会对我们在“七五”期间建立的100万词级《现代蒙古语文数据库》做了技术鉴定，充分肯定了《现代蒙古语文数据库》的意

* 该项目获国家教委资助。

义和作用。然而，作为整个蒙文信息处理工作的一项基础建设，100 万词级的语料显得不够充分，它不能充分表现蒙古语在各种领域内的实用情况。蒙文信息处理，不仅需要归纳、提取蒙古文的表层形式结构规则，还需要对蒙古文深层结构进行研究。满足这种需要的语料库，似乎数量级越高、代表性越广、用起来就越好。但具体实现，还要受限于人力、财力、时间等多方面的因素。我们根据自己的现有条件，计划第一步把《现代蒙古语文数据库》从 100 万词级扩充到 500 万词级。1993 年国家教委为此立项资助，并要求我们在 1995 年 12 月前实现。

按照本项目的实施计划，我们在原有 100 万词级《现代蒙古语文数据库》的基础上增加了 460 万单词的语料，使之成为当今世界上最大的现代蒙古语文语料库。新增加的语料中：

1. 文科教材类有 34 万多单词、占 7.3%；
2. 理科教材类 45 万多单词、占 9.6%；
3. 文学类 79 万多单词、占 16.9%；
4. 政治类 72 万多单词、占 15.4%；
5. 新闻类 81 万多单词、占 17.4%；
6. 口语类 4 万多单词、占 0.9%；
7. 社会科学类 129 万多单词、占 27.6%；
8. 自然科学类 26 万多单词、占 5.6%。

新设计的软件有：

1. 从蒙古文方正编码到 ASCII 码的转写
2. 蒙古文常用复合词自动识别
3. 蒙古文词根、词干、词尾的自动切分

并改进了蒙古文自动校对软件。

1997 年 7 月，内蒙古自治区教育厅组织验收了 500 万词级《现代蒙古语文数据库》，肯定了语料的代表性、广泛性和各类语料比例的合理性，尤其对从蒙古文方正编码到 ASCII 码的转写部分的意义和技术处理给予了高度评价。

二、建立 500 万词级《现代蒙古语文数据库》的主要问题

在建立 100 万词级语料库的五年期间，我们曾经为录入、校对、修改语料库文本而投入了艰苦的劳动，耗去了大量的时间，付出了极其昂贵的代价。在暂短的三年时间内完成 400 万单词的录入、校对和其它文本加工任务，仍然是建立 500 万词级《现代蒙古语文数据库》的主要困难所在。然而，电子出版系统的普及为我们提供了丰富的语料资源。利用电子出版系统的文本来补充现代蒙古语文数据库的语料自然是一种妥善解决录入、校对问题的良好途径。因为这样做，可以避免录入、校对过程中的重复劳动，可以节省大量的人力、物力、财力和更多的时间。

蒙文是一种拼音文字，它的最小独立单位是字母。作为蒙古文信息处理的大本营，现代蒙古语文数据库要求被收入的语料文本一定要保证字母的独立间隔，即语料中的每一个蒙文单词一定是由若干个只表示一个独立字母的字符和表示特殊情况的区别符构成。然而，由

于历史的种种原因，从 1987 年公布的我国蒙古文国家标准编码到后来各种演变出台的蒙古文编码，都未能适应蒙古文信息事业深入发展的需要。它们的共同缺点就是编码收字原则没有以字母为单位。例如，现行蒙古编码字符集中的“ᠠ”，它至少可以对应蒙古文三个不同字母 ᠠ [ɑ]，ᠡ [ə]，ᠨ [n]的词中形式，“ᠣ”可以对应四个字母 ᠣ [ɔ]，ᠤ [ʊ]，ᠥ [o]，ᠦ [u]的词中形式，“ᠢ”可以对应字母 ᠵ [dʒ]的词首形式，也可以对应字母 ᠢ [i]或字母 ᠶ [j]的词中形式。字符集中还有 ᠪᠠ ，ᠪᠣ ，ᠪᠢ … 等由一个辅音字母加一个元音字母而成的音节字符，有由两个辅音字母构成的 ᠪᠯ ，ᠺᠯ ，ᠮᠯ … 和由两个辅音字母加一个元音字母构成的 ᠮᠯᠢ 等连字符。不但有 ᠠ，ᠠ 等只作为字素的字符，象上面提到的 ᠠ ，ᠣ ，ᠢ 等字符，还要充当如同字母 ᠪ 的词中形式 ᠪ 和字母 ᠥ 及其词首形式 ᠥ 等一些独立字母的部分字素（笔划）角色。这就是说，由于现行蒙文编码都没有保障码位用途的唯一性，无论用哪一种蒙文编码系统建立的文本，都不能直接纳入现代蒙古语文数据库作为语料文本。借助蒙古文电子出版系统的文本来作为现代蒙古语文数据库的语料，首先要解决的问题就是明确文本中每一个蒙文编码在当前单词中所起的作用，即找出每一个单词的字母间隔并确认每一个字母的读音和书写特征。

为了更好地反映字母独立间隔，由于以字母为单位的蒙文编码至今还没有问世，我们暂时采取用拉丁字母和其它一些 ASCII 码字符来转写蒙文字母的办法－MTOA，以此作为保障语料库文本字母独立间隔的一种过度性措施。在研制蒙古文国际标准编码时，我们充分注意到了保障字母独立间隔的重要性，在编码原则中明确规定“蒙古文国际标准编码只对蒙古文规范字符编码，大于一个字母的连字和小于一个字母的字素都不予编码”。不久的将来，蒙古文国际标准编码会得到国际标准化组织的批准并系统实现。那时，我们将会很轻松地把 ASCII 码文本再转回由蒙古文国际标准编码构成的文本。用现行蒙古文编码建立的大量电子出版系统的文本，如果仅仅满足印刷需要而告结使命，那将是一个非常遗憾的事情。它虽然目前还不能够直接服务于蒙文信息处理研究工作，但它毕竟是众人付出大量辛勤劳动的结果，是不该丢弃的珍贵的信息资源。如果把它们改造成具有字母独立间隔特性，并用 ASCII 码转写下来，那么它就可以变成用现有的条件和通讯联网设备，在国内、国际间可以交换的数据。从而它可以变成在广泛的领域内，在各种信息处理用途中可以共享的资源，变成整个蒙古文信息处理事业的宝贵财富。我们研制转写软件 MTOA 的目的和意义，从狭义上讲，它能够为我们建立 500 万词级语料库的工作解决语料来源和获得途径，节省我们大量的精力和时间。从广义上讲，它能够使大量的电子出版系统的文本变成蒙文信息处理多种用途可以共享的资源。这一措施的实现，是本课题研究突破的最主要的技术难题。另外，考虑到北大方正电子出版系统蒙文版在国内外蒙古文使用和研究领域中的普及率最高，MTOA 选择方正蒙文编码作为转写的对象。至于其它现行蒙古文编码系统，它们和方正蒙文编码系统之间的差别不是很大，只要做一些相关的改动就可以实现从其它蒙古文编码系统到 ASCII 码的转写系统。

三、用 ASCII 码转写蒙文字母的约定

为了使转写结果仍具有较好的可读性，用 ASCII 码转写蒙文字母时我们规定了尽可能选用读音相近的拉丁字母来转写的原则。具体转写对应关系如下表所示：

蒙文字母	ASCII 码	蒙文字母	ASCII 码
ᠠ	A	ᠰ	S
ᠡ	E	ᠱ	$
ᠢ	e	ᠲ	T(t)
ᠣ	I	ᠳ	D(d)
ᠤ	0	ᠴ	C
ᠥ	V	ᠵ	J
ᠦ	O	ᠶ	Y(y)
ᠧ	U	ᠷ	R
ᠨ	N(n)	ᠸ	W
ᠪ	B	ᠹ	F
ᠫ	P	ᠺ	K
ᠬ	H	ᠬᠤ	h
ᠭ (ᠺ)	G(g)	ᠽ	c
ᠯ	L	ᠵᠢ	Z
ᠮ	M	ᠷᠢ	r

补充说明：

①“%”表示它的相邻字符来自蒙文字母词中形式。

②转写字母 ᠠ 、ᠡ 的分写形式时，在 A 、E 前加“_”。 //底划线

③转写字母 ᠩ(ᠺ)，用 NG 转写。

④ᠨ 字母在词中辅音前以 ᠨᠢ 形式出现时，用 n 转写。

⑤ᠭ 字母在中性词中出现时，用 g 转写。

⑥ᠲ 字母在词中以 ᠲᠠ 形式出现时，用 t 转写。

⑦ᠳ 字母在词首以 ᠳᠠ 形式出现时，用 d 转写。

⑧词中出现的 ᠶᠠ ，用 YI 转写。

⑨词中出现的 ᠶᠠ ，用 yI 转写。

⑩转写静词类的分写附加成分，前加“－”。 //减号

⑪格附加成分 ᠤᠨ 、ᠡᠨ 转写成－IYEN（－IYAN），－IYER（－IYAR）。

⑫ᠵᠢ 、ᠴᠢ 用 ZHI 、CHI 转写。

⑬类似 ᠣᠨ 、ᠤᠨ 等应转写成 O’ 、SU’ 等等。

四、MTOA 的软件结构

MTOA 由数据库、知识库和控制模块三部分组成。

（一）数据库

数据库包括一部词典管理模块和八部规范化的词典数据、一组蒙古文变形附加成份、一组蒙古文构词附加成份组合。另外，还有名词的变动词干部分。词典管理模块提供文件编辑软件、排序软件和词典格式规范化软件。依次运行这些软件，就可以完成增补、删除、修改词条等词典编辑工作，可以对各种词典进行排序，最终形成供 MTOA 控制模块调用的规范化的数据。

根据各词典的单词编码类型，这八部词典可分为蒙文码词典和 ASCII 码词典，两类各含四部词典。根据蒙文码词典中单词的书写形式，进一步划分整词词典和词干词典两种。所谓整词的意思是指词条所含单词的书写形式以规范的词首形式开始以规范的词尾形式结尾。

当然，词中出现的字母也以规范的词中形式书写。词干词典与整词词典所不同之处是词干词典中的单词都以词中书写形式结尾。四部蒙文码词典中一部是整词词典，其它三部为词干词典。这三部词干词典分别为动词词干词典、名词变动词干词典和名词类词干词典。四部ASCII码词典分别与这四部蒙文码词典对应，它们的词条只是一串ASCII码，表示所对应蒙文码词典中的某一词条所含蒙文单词的ASCII码转写结果。

蒙文码词典与ASCII码词典之间，通过一个索引表相对应。建立这一索引表是有必要的。因为随着对MTOA的各词典进行编辑操作，词条的数目、位置都有可能改变。在MTOA中，词典编辑操作在各ASCII码词典上进行，而查找工作是在各蒙文码词典上进行的。词典查找方法有很多种，其中我们选择了查找速度比较快的二分法。因为二分法只有在有序数据上才能正确运行，MTOA要求各蒙文码词典必须有序，而ASCII码词典的序列要求对MTOA来说不是必要的。既便排了序，ASCII码词典中的词条和与之对应的蒙文码词典中的词条之间不会保持水平方向上的对应关系。

由于这些词典的容量都比较大，直接采用分配数组调入内存的方法已突破了DOS系统访问内存的常规。MTOA的主体操作是词典查找，如果把频繁进行的主体操作都要通过硬盘访问途径来实现，那么MTOA的运行速度就会受到很大影响，整体效率会明显地降低。为解决这一矛盾，我们采取了在扩展内存中建立虚拟盘的措施，把MTOA用到的词典都置入虚拟盘内。使用虚拟盘的技术已经很成熟普及，效率也很高，不需要什么特殊的软件来支持。因此，采取建立虚拟盘的措施也便于MTOA的普及。我们用访问硬盘和访问虚拟盘的两种途径，用MTOA转写了含62314单词的一个文件。结果，访问硬盘用了36分55秒，而访问虚拟盘用了16分47秒。

附加成分量不大，可以直接调入数组，放置内存即可。

(二)知识库

识别字母独立间隔并确认字母，要涉及到许多语音、语法和蒙古文正字法知识。如蒙古文元音和谐律、名词的变动词干与格附加成分的搭配规律、辅音字母结合规律、连结元音规律等等。另外，外来语和分写附加成分也都是具有独特规律的转写对象。

我们共编排了二十多个独立函数，建立了MTOA的专用知识库。这些函数可以分别判断当前单词的阴阳性、是否属于外来语、是否具有词干加附加成分结构和当前码的各种属性等等。控制模块可以随时调用这些函数、做出对当前情况的各种判断，正确选择下一步操作。

(三)控制模块

控制模块是MTOA的核心，即它控制MTOA的全部操作过程。MTOA的操作，大体上分以下七个方面。

1. 整词查找　这是MTOA中最基本的、最单调的一种操作。说最基本的是因为所有被转写的单词都首先走过这一关，说最单调的是因为这不需增加任何辅助操作，只是在蒙文码整词词典上做一次二分法查找。查找成功，则通过蒙文码整词词典索引表，从ASCII码整词词典中索取转写结果。查找不成功，直接进入下一个操作。

2. 词干加附加成分　这一步主要是识别当前单词是否具有动词加动词变形附加成分结构。先判断该单词是否以某一个动词变形附加成分结尾。判断成立，切去与动词变形附加成分匹配部分，再判断剩余的词干部分是否在动词词干词典内。两步判断均成立，则这

次操作成功。如果第一步判断成功而第二步判断不成功，那么有可能不成功的原因由辅音字母结合规律和连结元音规律所致。所以，这需要进一步判断匹配动词变形附加成分的部分是否以使变辅音开头。是，则判断剩余的词干部分是否以“ᠣ”或“ᠥ”或“ᠦ”等结尾。如果这一判断也成功，那么根据具体成功的条目，还需要进一步判断词干尾部的辅音情况。如果词干尾部是可变辅音，而且它与当前动词变形附加成分的连接变化恰好和当前单词中所出现的情形一致，那么，根据实际匹配情况，对词干做相应的修正后再从动词词干词典中查找。查找成功，则这一操作以成功结束。在以上判断、查找过程中，只要有一处不成功，则这一操作以失败结束。这一操作遍历全部动词变形附加成分。

3. 词干加附加成分组合　　这一步操作过程与上述词干加附成分的处理过程基本相同。所不同的是这里所说的附加成分组合是动词构词附加成分的组合，这些组合有些在词中出现，有些在词尾出现。在词中出现的情形，归操作 4 处理。另外，动词构词附加成分 ᠯᠠ ᠷᠠ …等是接在名词、数词和形容词后派生动词的。所以在它们前面出现的词干部分，只有在名词类词干词典中出现，才能够使操作成功。操作 2 的词典查找只涉及动词词干词典，而操作 3 则涉及动词词干词典和名词类词干词典。

4. 词干加附加成分组合再加附加成分　　这一操作是 MTOA 最为复杂的一步。因为它不仅要遍历全部以词中形式出现的动词构词附加成分组合（内循环），还要遍历全部动词变形附加成分（外循环），判别动词构词附加成分组合与词干部分衔接处的辅音结合情况，判别动词构词附加成分组合与尾部动词变形附加成分之间辅音结合情况，词典查找涉及动词词干词典和名词类词干词典。MTOA 在这一步花费的时间也最多。

5. 外来语　　蒙文字母中有一些字母是专门用来记外来语的。例如，ᠫ，ᠹ，ᠽ，ᠼ…等。这些字母的出现，提供当前单词属外来语的信息。在外来语中，有些字母的转写变得有规律可寻。例如：ᠣ 在阳性词中为 [ɔ]、在阴性词中为 [u]，没有表示[ʊ]和[o]的可能，ᠳ 只表示[d]、ᠲ 只表示[t]等等。

6. 其余单词　　这些单词，还有在操作 5 中提到的外来语，是 MTOA 用 1 — 4 的操作不能转写的单词。这类词的大部分是外来语或人名地名等专用名词。它们的拼写一般不符合蒙文拼写常规，人们很难给出识别它们字母独立间隔的条件。MTOA 转写这些单词的结果是模糊的，因为在这里它主观地参照了蒙文拼写常规和外来语转写特点。为降低这种模糊性，我们搜集了一些拼写人名、地名的单词，作为 MTOA 词典的补充。

7. 分写附加成分　　静词类的分写附加成分是蒙语非常活跃的成分，它们的使用度很高。在方正蒙文编码中，如同以上所述的 ᠣ、ᠢ、ᠢ 等码能够代表的角色很多，ᠲ、ᠳ 等在不同环境中代表不同字母的码也不少，并且它们的出现频率都比较高。这不仅对识别字母独立间隔带来了相当多的困难，而且对 MTOA 的整体效率也起到举足轻重的作用。在格附加成分中，它们的出现频同样很高。但有幸的是这些码所代表的角色，在格附加成分这一限定环境中，变得比较单纯。例如，在格附加成分中，ᠲ 永远是[t]，ᠳ 永远是[d]，ᠢ 永远是元音，ᠥ 只能是[ʊ]或[u]（在非格附加成分的场合，它还可以是[ɔ]，[o]）。利用格附加成分的这些方便之处，MTOA 专门设定了转写格附加成分的独立模块。这一模块简化节省了许多判断步骤，对提高 MTOA 的整体效率和速度起到了积极的作用。

五、后记

词典是MTOA的根本基础。MTOA虽然安排了一系列推导措施，但这些推导追根究底，最后还是取决于词典。一部好的词典*，是能够复盖动态书面语的绝大部分，但永远不会是全部。所以，在转写过程中遇到一些词典没有收进来的词语是无可非议的事。再说，蒙文中不同字母使用同一个"符号"书写的现象很多。例如，方正蒙文编码中的 ᠥ 可以表示[h]+[ə]，还可以表示[g]+[ə]。类似这样的情形，导致了一些形同音不同词语的产生。词义的识别已超出MTOA的功能范围。由于这些原故，对MTOA的转写结果还需要做一部分人工校对工作。人工校对的对象是：①冠以"+"号的词。"+"表示该单词可能有其它转写形式。②冠以"*"号的词。"*"表示该单词的转写可能有错误。例如，MTOA转写实 ᠨᠠᠷᠠᠨ 一词的结果是+NEREN，因为它还可以转写为NARAN。转写 ᠬᠤᠸᠠᠰᠤᠨ 一词的结果是*H0WA $ AB00。这是一个很少碰到的人名，词典不会收进去。MTOA只是主观地参照常规转写规则和转写外来语的一些限制条件把它转写成这样，正确的转写结果应该是HVWA $ ABVV才对。

作为对MTOA的测试，我们转写了从100万词级《现代蒙古语文数据库》中随机选择的一部分语料。这部分语料含794个单词，转写结果中"+"号出现了77次、"*"号出现了32次，准确率达86.27%。

建立500万词级《现代蒙古语文数据库》是由确精扎布教授主持的项目，参加本项目研究工作的还有华沙宝、吉仁尼格、那顺乌日图、白音门德、娜仁通拉嘎等同志。MTOA是完成该项目的主要措施，用它转写460万词的语料，收入到500万词级《现代蒙古语文数据库》中。在研讨MTOA的过程中，我们遇到了一些传统研究中从未碰到的问题。从计算语言学的角度去观察和研究这些问题是很有趣的。

参考文献

〔1〕冯志伟，"中文动词词组型科技术语潜在歧义结构的实例化"，《计算语言学研究与应用》，北京语言学院出版社，1993年9月。

〔2〕俞士汶，"语言信息处理研究的意义与方法"，《中国计算机报》，1991年第18期。

〔3〕苑春法，黄昌宁，孙致宇，赵强，"新一代语料库的建设与管理"，《计算语言学研究与应用》，北京语言学院出版社，1993年9月。

〔4〕刘开英，郭炳炎，《自然语言处理》，科学出版社，1991年。

〔5〕确精扎布，那顺乌日图，"关于蒙古文编码(上、下)"，《内蒙古大学学报》，1995年第1、2期。

〔6〕华沙宝，"现代蒙古语文数据库的程序设计"，《内蒙古大学学报》，1991年第2期(蒙文版)。

〔7〕华沙宝，"蒙古文自动校对系统"，《内蒙古大学学报》，1997年第4期。

* MTOA的词典取源于内蒙古大学蒙古语文研究室编纂的《蒙汉辞典》。