蒙古语语料库综述

雪艳[1]，文化[2]，那顺乌日图[3]
（内蒙古大学 蒙古学学院,内蒙古 呼和浩特 010021）

摘要： 内蒙古大学蒙古语文研究所，自1983年以来先后建立了各种语料库10余个，这些语料库是蒙古语研究的重要资源，尤其是经过浅加工的100万词级现代蒙古语语料库自建成到现在一直都是现代蒙古语研究和蒙古文信息处理研究的重要数据来源。直至今天，蒙古语语料库的建设和研究工作整整经历了20年。我们有必要及时回顾和整理这些年的工作，将蒙古语语料库建设的经验和最新的语料库建设、加工、管理技术结合起来，提高蒙古语语料库的建设质量和开发效率，使它具有更广泛的利用价值。

关键词： 蒙古语语料库；利用状况

中图分类号码： H085.6　　**文献标识码：** A

一、　引言

20世纪80年代初，内蒙古大学蒙古语文研究所与内蒙古计算中心合作，将“蒙古秘史”输入计算机，并匹配了一些检索、分析软件。在此基础上，内蒙古大学蒙古语文研究所建立了“中世纪蒙古语语料库”（1984年），开了蒙古语文语料库建设的先河。自1984年至1990年，内蒙古大学蒙古语文研究所在国家社科基金资助下，又建立了100万词级的“现代蒙古语文数据库”。该数据库于1990年通过内蒙古自治区科学技术委员会的组织鉴定，被誉为“世界上第一个现代蒙古语文数据库”。完成100万词级的“现代蒙古语文数据库”之时，国内外很多语料库均已达到或超过500万词规模以上。由此，紧接着将语料库扩展到了500万词。在国家教委人文社科基金资助下，500万词级《现代蒙古语文数据库》于1998年建成，增加了一部分数理化教材和医学、法律等方面的内容，但该语料库是未经加工的生语料库。

20世纪90年代后，内蒙古大学蒙古语文研究所又先后建立了契丹小字语料库（2000年）、八思巴字文献语料库（2001年）和蒙古语口语材料语料库（2004年）等。

2000年，结合汉蒙机器翻译系统，建立了近20万词的汉蒙对照政府文献语料库。为了研制基于实例的汉蒙机器翻译引擎，目前正在建立150万词级的汉蒙双语对齐语料库。

上述蒙古语语料库，各自都有不同的目标和特征，它们的规模、层次、用途和加工程度都不一样。其中，建设重点是现代蒙古语语料库。经过十多年的建设，现代蒙古语语料库已初具规模，具备了为蒙古文信息处理提供基本信息的能力。

建立现代蒙古语语料库初期，我国颁布、使用的蒙古文编码国家标准是基于字形的码，无法表现（或存储）蒙古文中诸多同形异音字母的区别。因此建立语料库时采用了蒙古文拉丁转写方式，同时制定了一套蒙古文拉丁转写规则。建立语料库初期，语料库中有些特殊词语形式、各类附加成分、人名、地名和复合词等都采用了人工标记的方式。

在语料库加工方面，也做了一些探试性的工作。如，为了解决键盘输入语料库的文本校对问题，曾编写了面向拉丁转写文本的现代蒙古语自动校对程序。此后，又进行了蒙古语词干、词根、词尾的自动切分和复合词的自动识别、蒙古语语料库的词类标注等一系列基础性的工作。目前正在对蒙古语语料库进行短语标注。蒙古语传统研究的有关理论以及近几年所做的面向信息处理的蒙古语词语语法属性、语义属性研究以及蒙古语语料库词类标注研究等成果，为我们进一步对蒙古语语料库加工提供了一定的知识、技术支持。近年正在进行面向信息处理的蒙古语语义研究成果将被运用到语料库语义标记方面。

蒙古语语料库在蒙古语语料库语言学的兴起中起了非常重要的作用：（1）提供了真实语料；（2）提供了统计数据；（3）验证现行的理论以及构建新的理论方面提供了依据。

直至今天，蒙古语语料库的建设和研究工作整整经历了 20 年。我们有必要及时回顾和整理这些年的工作，将蒙古语语料库建设的经验和最新的语料库建设、加工、管理技术结合起来，提高蒙古语料库的建设质量和开发效率，使它具有更广泛的利用价值。

二、蒙古语语料库种类、加工情况以及相关的加工检索统计软件

我们把 10 余个语料库分为蒙古语单语语料库和汉蒙双语语料库两大类。又将蒙古语单语语料库按照语料来源的不同蒙古文字种类，分为若干不同类型的语料库。分类情况如图 1 所示：

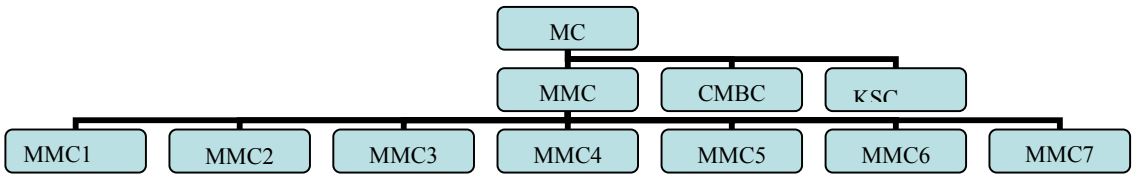

图 1

MC(Mongolian Corpus):蒙古语语料库

MMC(Mongolian Monolingual Corpus):蒙古语单语语料库

MMC1: 汉字标音蒙古文语料库（《元朝秘史》语料库）

MMC2: 回鹘体蒙古文语料库

MMC3: 八思巴文语料库

MMC4: 托忒文语料库

MMC5: 口语材料语料库

MMC6: 17 世纪满蒙关系书信语料库

MMC7: 现代蒙古文语料库

CMBC(Chinese-Mongolian Bilingual corpus):汉蒙双语语料库

KSC(Kidan Script Corpus): 契丹小字语料库

内蒙古大学蒙古语文研究所针对研究契丹小字的需求，建立了契丹小字语料库，该库以碑文（金石资料）作为语料来源，利用方正蒙文版 7.30 的内码录入原语料。另外，还专门开发了一套检索系统，该系统具有查询原字出现频率、出现位置以及进行原字索引的功能。

下面对其他语料库的加工情况以及相关的加工检索软件做详细介绍。

（一）蒙古语单语语料库（MMC）

1、汉字标音蒙古文语料库（《元朝秘史》语料库）（MMC1）

《元朝秘史》是极为重要的中世纪蒙古语文献。内蒙古大学蒙古语文研究所的语料库建设工作正是以它作为起点，一步步向前迈进的。《元朝秘史》语料库的建设共经历了三个阶段，详情如下：

（1）第一代《元朝秘史》语料库（又名：《元朝秘史》文件索引）

建成时间：1983 年
语料来源：李盖提拉丁转写本
建库时的软硬件环境：Alpha-1000 微机
文字录入方法：参照李盖提拉丁转写体例，用 ASCII 码转写蒙古文
目前已废弃。

（2）第二代《元朝秘史》语料库

建成时间：1988-1989
语料来源：按《四部丛刊》本，录入了汉字标音部分和旁译，还参照李盖提拉丁转写体例录入了拉丁转写部分。
建库时的软硬件环境：PC 机、DOS 操作系统
文字录入方法：拉丁转写、汉字
加工情况：有词尾切分、断句标记
检索功能：具有以汉字标音、旁译、拉丁转写对词根、附加成分进行查询的功能；可用上述三项（标音、旁译、拉丁转写）中的任一项作为索引对象，对另两项进行查询（即用三种格式进行检索）；还可根据要求进行断句查询
目前已废弃。

（3）第三代《元朝秘史》语料库

建成时间：部分建立
语料来源：参照李盖提拉丁转写体例录入了拉丁转写部分（对李盖提拉丁转写体例中的不当之处作了详细修改）。其中第一章按《四部丛刊》本录入了汉字标音部分和旁译
文字录入方法：拉丁转写、汉字
加工情况：有词尾切分、断句、引语、人名、地名标记
检索功能：直接利用 WORD 编辑器的检索功能

2．回鹘体蒙古文语料库（MMC2）

（1） “回鹘体蒙古文文献（道布）”
（2） 黄金史

建成时间：1984 年
文字录入方法：蒙古文拉丁文转写
加工情况：有词干词尾切分标记；复合词标记断句、引语标记
相关的加工检索统计软件：mdb（现代蒙古文语料库检索系统，后面有关于 mdb 的详细说明）

3．八思巴字文献语料库（MMC3）

建成时间：2001 年、2004 年
语料来源：八思巴字蒙古语文献

文字录入方法：方正蒙文书版 6.0
相关加工检索统计软件（检索功能）：利用读音或转写符号或拉丁转写形式对单词、附加成分、音节以及任意字符串进行带有例句（音节和任意字符串的查询结果中不带例句）的查询；还可以建立词表。

4．托忒文语料库（MMC4）

建成时间：1999 年、2000 年
语料来源：《江格尔 1》、《江格尔 2》（中国民间文学艺术研究会新疆分会整理，新疆人民出版社，出版年月 1985、1987）
文字录入方法：托忒蒙古文拉丁转写
加工情况：有词干词尾切分标记、复合词标记
相关加工检索软件：mdb

5．口语材料语料库（MMC5）

建成时间：2004 年
语料来源：国际音标标音语料、录音材料
录入方法：方正蒙文书版 6.02，国际音标
加工情况：有词干词缀切分标记
相关加工检索软件：mdb

6.《17 世纪满蒙关系书信》语料库（MMC6）

建成时间：部分建立
语料来源：蒙古王公与满清统治者之间来往的书信档案（中国第一历史档案馆提供）
文字录入方法：蒙语部分，蒙古文拉丁文转写；满语部分，尚未录入，录入方法有待确定。
加工情况：有词干词缀切分标记；词性标记
相关加工检索软件：mdb

7．现代蒙古文语料库（MMC7）

内蒙古大学蒙古语文研究所于 1991 年建立了 100 万词级现代蒙古文语料库后，又经过几年的努力，到 1998 年的时候将这个语料库扩充到了 500 万词级。就语料类型和加工程度而言，前 100 万词级语料和后扩充的 400 万词级语料有很大的区别，所以分两部分进行介绍。现代蒙古文语料库中除了有以上提及的综合型通用语料以外，还包括一个纯文学语料库。现代蒙古文语料库以蒙古文拉丁转写形式录入，共使用一套加工、检索统计系统 mdb（后面要专门讨论）。

（1）100 万词级《现代蒙古语数据库》

建成时间：1991
语料分布：1.蒙语文教材有 50 万词左右，占 50．3%；2.政治类有 20 万词左右，占 20．3 %；3.文学类有 20 万词，占 19．6%；4.报纸新闻类有 10 万词，占 9．8%。
文字录入方法：蒙古文拉丁转写
加工情况：经过 5 次人工校对；有词干词缀切分标记、复合词标记和人名、地名标记；其中 30 万词语料带有词性标注

（2）400 万词级《现代蒙古语数据库》

建成时间：1998 年

语料分布：新增加的约 460 多万词的语料中，1.文科教材类有 34 万多词，占 7.3%；2.理科教材类 45 万多词，占 9.6%；3.文学类 79 万多词，占 16.9%；4.新闻类 81 万多词，占 17.4%；5.政治类 72 万多词，占 15.4%；6.社会科学类 129 万多词，占 27.6%；7.自然科学类 26 万多词，占 5.6%;6.口语类约 4 万词，占 0.9%。比原计划多收集了 70 来万词的语料是为最后对语料的种类进行调整作准备的。
文字录入方法：蒙古文拉丁转写
加工情况：语料的加工情况，深浅不一
除此，还有一部分补充语料，该语料以文学为主，有词干词缀切分标记、复合词标记和人名、地名标记。

（二）汉蒙双语对照语料库（CMBC）

2000 年，为检验汉蒙机器翻译结果，我们曾经做过一些汉蒙对照政府文献语料库。2003 年，内蒙古大学蒙古语文研究所承担了一项 863 项目，即基于实例的汉蒙机器翻译系统的研制和开发，汉蒙双语语料库的建设工作也随即正式开始。这对汉蒙双语语料库对蒙古文信息处理、语言教学、汉蒙双语研究和汉蒙双语词典的编纂等工作都有着极为深远的意义。目前该库的建设情况如下：
建成时间：正在建设中
文字录入方法：蒙语部分用蒙古文拉丁转写形式录入
加工情况：以句对齐形式录入，对其中 350 个汉蒙句对中的蒙古语句进行了词干词缀切分标注以及词性标注。
相关加工软件：初步建成了新的蒙古语词性标注软件和汉蒙词语自动对齐软件（有待进一步完善）

（三）蒙古语语料库加工检索软件简介

蒙古语语料库的一系列加工检索软件是在 DOS 环境下实现的。这套工具主要用于回鹘体蒙古文语料库、托忒文语料库、口语材料语料库、《17 世纪满蒙关系书信》语料库以及现代蒙古文语料库的加工和检索。由于这套系统的功能相对较多，通用性也较强，因此在这里做单独介绍。

1．加工软件

蒙古语语料库的加工程序主要有以下几种：附加成分的自动切分程序、自动校对程序(MHAHP)、词类标注系统 (AYIMAG)、复合词的自动识别程序构词附加成分自动切分标记、从方正蒙文编码到 ASCII 码的转写程序 (MTOG)、新的蒙古语词性标注系统、汉蒙双语语料库词语自动对其程序等。

我们在后来的汉蒙双语语料库的建设过程中，发现有必要对蒙古语词做更深层次的语法标注，即不仅对某一词的整体词性做出识别，还要对构成词的词干和附加成份进行分别标注。这样才更有利于下一步的汉蒙词语对齐的实现。根据深加工的需求，我们重新修订了蒙古语词类标注体系。又在此基础上，实现了新的蒙古语词性标注系统。

2．语料库统计检索工具 mdb（面向蒙古语文研究者的应用软件）

mdb 系统主要包含以下几项功能（目录）：（1）语料编辑；（2）正字法校对；（3）查询分析；（4）音节；（5）句子统计；（6）建立词库；（7）统计；（8）浏览语料；（9）打印；（10）排版

预处理；（11）结束。

三、 蒙古语语料库的利用状况

语料的加工情况和相关检索软件的成熟程度决定了其利用率的高低。下面主要以中世纪蒙古语料库和100万词级现代蒙古语浅加工语料库为例，介绍近20年来蒙古语语料库的利用情况。

（一）蒙古语语料库在传统蒙古语言研究方面的利用情况

“中世纪蒙古语语料库”的问世，不仅为蒙古语族语言的历史比较研究和以《元朝秘史》为主的中世纪蒙古语研究提供了易于查询的语言材料，而且为以语料库语言学方法研究蒙古语打下了良好的基础。如:《〈蒙古秘史〉语言的数范畴》(确精扎布)、《关于〈蒙古秘史〉中〈ede,tede〉两个词》(莫·恩和巴图)、《〈蒙古秘史〉语言的-da/-de~-ta/-te~-a/-e~-tur/tor~-dur/-dor附加成分》(哈斯巴特尔)等论文，是在利用《元朝秘史》文件检索系统（第一代《元朝秘史》语料库），对《秘史》某些语言现象进行研究的基础上撰写而成的。此后语料库语言学方法在蒙古语言的研究中被广泛采用。内蒙古大学蒙古语文研究所的很多研究生的学位论文也都是利用《元朝秘史》文件检索系统来完成的。如:《内蒙古大学蒙古语文研究所的研究生毕业论文（第二集)》（1991年5月)。

《100 万词级现代蒙古语语料库》建立初期，内蒙古大学的学者们首先利用它对蒙古语字母的使用度、蒙古语数范畴的内容、名词与形容词在句子中的作用等进行分析，同时提出了不少新的见解。如：确精扎布著《蒙语语法研究》（第一册）。

（二）蒙古语语料库在面向蒙古文信息处理的基础研究方面的利用情况

《100 万词级现代蒙古语语料库》不仅对传统语言学研究开辟了新的途径，而且为面向蒙古文信息处理的各项基础研究也提供了重要的数据依据。如：那顺乌日图《“蒙古语语法信息词典”框架设计》、青格乐图的《面向信息处理的蒙古语固定词组研究》、巴达玛敖德斯尔的《语文词典中无词类标注词的处理研究》、扎·伊兰的《关于现代蒙古语前接词与后接词》、那仁通拉嘎的《面向语文信息处理的蒙古语词变化形式分类》、达胡白乙拉的《面向信息处理的现代蒙古语名词短词结构规则研究》、姜迎春的《面向信息处理的蒙古语形容词语义研究》、额尔敦朝鲁的《面向信息处理的蒙古语动词语义研究》、吉仁花的《面向信息处理的蒙古语形容词短语结构研究》、巴·萨日娜的《“蒙古语语法信息词典”中动词语法属性字段的填写》等学位论文都具有代表性。

（三）蒙古语语料库在应用系统开发方面的利用情况

除了论文形式的成果之外，蒙古语词频词典的编纂、蒙古语兼类词统计矩阵模型的建立、蒙古语语法信息词典中的诸多属性字段的设置以及蒙古语人名识别规则集的建立等等应用成果也是都在《100万词级现代蒙古语语料库》库提供的数据基础上取得的。

从以上语料库的利用情况看，蒙古语语料库虽然规模较小、加工粗浅，存在着很多不完善的地方，但它奠定了蒙古语文信息处理研究的基础，为人们提供了较为可靠的语言数据，促进蒙古语言学从传统的“定性研究”逐步转向“定量研究”的过程。

四、结语

蒙古语语料库已初具规模，但是在语料本身以及加工检索系统的管理方面还存在很大的漏洞。这种现象不仅不利于语料库的扩充和进一步的加工，而且对语料库用户也带来了很大的不便。

就后续工作而言，一方面要继续扩大语料库的规模，使统计规律更贴近语言实际情况。但是这里存在一个语料收录原则的问题，如何制定合理的收录原则，建立大规模蒙古语平衡语料库等将是我们下一步工作中重点考虑的问题之一。

另一方面，要制定合理的、严格的加工标注规范，对语料进行语法、语义方面的深加工，使语库成为真正意义上的语言知识库的有机组成部分。

另外，还要加强语料库数据的管理和维护。

参考文献

[1]华沙宝,巴达玛敖德斯尔.蒙古语语料库建设现状分析和完善策略[M].北京:清华大学出版社,2003.

[2]那顺乌日图.蒙古文信息处理[M].呼和浩特 :内蒙古科学技术出版社,1998.

[3]那顺乌日图.关于在蒙古语文研究中运用统计学方法的问题[J].民族语文,1993,(5).

[4]五百万词级现代蒙古语文数据库[M].研制报告.

[5]现代蒙古语文数据库[M].使用说明书.

[6]华沙宝.关于蒙古语语料库建设[M].北京:2004,4,11-12.

[7]黄昌宁,李涓子.语料库语言学[M].北京:商务引书馆出版社,2002.

[8]姚天顺,等.自然语言理解[M].北京:清华大学出版社,2002,第二版.

[9]华沙宝.实现500万词级“现代蒙古语文数据库”的主要措施[A]。语文学术论文集（七）[C].呼和浩特:内蒙古大学蒙古语文研究所,1984.

[10]那顺乌日图.蒙古文词根、词干、词尾的自动切分系统[J].内蒙古大学学报,1997,(2).

[11]内蒙古自治区电子计算中心蒙文信息处理组.运用电子计算机分析《蒙古秘史》语言的情况[A].语文学术论文集（七）[C].呼和浩特:内蒙古大学蒙古语文研究所,1984.

[12]内蒙古大学蒙古语文已研究所.《蒙古秘史》词典[A].语文学术论文集（七）[C],1984.

[13]那顺乌日图,刘群,巴达玛敖德斯尔. 关于“汉蒙机器辅助翻译系统”[A]。ALTAI HAKPO （JOURNAL OF THE ALTAI SOCIETV OF KOREA）2001,11.

[14]华沙宝. 现代蒙古语文数据库软件[J].内蒙古大学学报,1992,(2).

[15]蒙古语文已研究所计算机室.关于现代蒙古语文数据库[J].内蒙古大学学报,1992,(1).

[16]确精扎布.关于《蒙古秘史》的复数附加成分[J].内蒙古大学学报（蒙）,1990,(3).

General Introduction of Mongolian Corpus

Xue Yan[1],Wen Hua[2],Nasun-urt[3]

(The Institute of Mongolian Studies, Inner Mongolia University, Hohhot 010021,China)

Abstract: The researchers in the Institute for Mongolian Language studies have established 10 more kinds of corpra since 1983,and they are very important resources for the Mongolian language studies. It's definitely necessary to review the 20 years' history of the Monglian corpus since we are now trying to improve the scale and quality of the corpus.

Key word: Mongolian corpus; usage;

收稿日期: 2004-9-13;

基金项目: 国家 863 计划项目(2003AA115510); 国家自然科学基金项目(36963005)

作者简介: 1、雪艳: 内蒙古大学蒙古学学院 2003 级博士生，主要研究方向是蒙古文信息处理;

2、文化: 内蒙古大学蒙古学学院 2003 级博士生，主要研究方向是实验语音学; 3、那顺乌日图: 内蒙古大学教授，主要研究方向是蒙古文信息处理。