蒙古语语料库建设现状分析和完善策略[1]

华沙宝　巴达玛敖德斯尔

内蒙古大学蒙古学学院 呼和浩特 010021

mghsb@imu.edu.cn　mgbmos@imu.edu.cn

摘要：本文对现代蒙古语语料库的语料做了分析，指出了语料的种类、规模、各类标记和标注加工等方面存在的问题，提出了将要采取的完善策略和近期达到的建设目标。重点建设蒙古语单语语料库，还要建立汉蒙并行语料库。

关键词：现代蒙古语，语料库，标注加工

On Analyses and Perfection for Mongolian Corpus

Huashabao　Badma-odsar

The Mongolian College, Inner Mongolia University, Hohhot 010021

E-mail: mghsb@imu.edu.cn　mgbmos@imu.edu.cn

Abstract: In this article, the authors analyze contemporary Mongolian Corpus and conclude the problems such as classification for language materials and size of the corpus, all kinds of tagging systems and tagging processing for the corpus. As a result, the purpose of constructing and the tactics of perfecting Mongolian corpus have been concluded. The purpose of construction is to build up Mongolian corpus mainly and bilingual corpus for Mongolian and Chinese.

Keywords: Contemporary Mongolian, Corpus, Tagging processing

在国家自然科学基金资助下，我们完成了对现代蒙古语语料库的词性标注研究。在这个基础上，我们正在开展对蒙古语语料库的短语标注研究。大家知道，短语标注研究，是整个语料库加工过程中的一个承前启后的关键环节。实现蒙古语短语标注系统，我们已经有了拟采用的方法和技术路线。这些方法和技术，都和现有语料库的实际情况有着直接或间接的依赖关系。因此，认真分析蒙古语语料库的现状，找出存在的问题，明确完善的策略，不仅对

[1] 本项研究获国家自然科学金（60263001）和国家社会科学基金的资助（02BYY036）。

顺利实现蒙古语短语标注系统有必要，对促进整个蒙古语语料库建设的发展也是很有意义的。

1 蒙古语语料库建设概况

1983 年，我们研制《元朝秘史》文件检索系统， 有了第一个蒙古文电子文本。随后又添加了《黄金史》、《回鹘蒙古文文献集》等语料，编写了一套蒙古语文研究专用软件，建立了中世纪蒙古文语料库。“7·5”期间，在内蒙古自治区哲学社会科学规划办和国家社会科学基金的资助下，我们建立了 100 万词级的《现代蒙古语文数据库》。后来得到教育部的资助，把《现代蒙古语文数据库》的规模扩展到 500 万词级。十几年来，经过国家自然科学基金资助项目一《蒙古文附加成分的自动切分和复合词的自动识别》、《对蒙古语语料库的词性标注与统计》等研究，《现代蒙古语文数据库》，即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多方面的基本信息的能力。但是，其中还有不少不完善的或者还没有启动的工作，我们今后继续建设的重点仍然还是现代蒙古语语料库。

2 现代蒙古语语料库现状分析

a) 拉丁转写规则

蒙古文是拼音文字，但字母不属于拉丁文。20 世纪 80 年代初，世界上众多文字还没有解决自己的输入输出问题，蒙古文也不例外。1987 年，我国颁布了蒙古文编码国家标准，但它的字符集只包含了蒙古文的一些基本字形，没有体现同形异音字母的区别。因此，从 1983 年第一次输入的《蒙古秘史》到后来建立的《现代蒙古语文数据库》，我们一直采用蒙古文拉丁转写方式。除了同形异音字母问题，在蒙古文拼写过程中，还需要解决各个字母以多种不同变形形式出现、相同字母组合以不相同的变形形式出现的问题。因此 ，我们不仅把蒙文基本字母（即字母原形）与拉丁字母对应起来，还选用了一些有形的 ASCII 符号，赋予特定的含义，构造了一套蒙古文拉丁转写规则。从中世纪蒙古语语料库到现代蒙古语语料库以及相关的一些词典，我们都是用这套拉丁转写规则来完成的。刚开始建立语料库时，我们并没有什么智能化工具，对于象真实文本中有些单词的不规范书写形式、动词附加成分、人名、地名和复合词等都用人工方式做了特殊记号。

b) 结构和选材

我们建立的 100 万词级《现代蒙古语文数据库》，语料的种类包括小说、语文、报刊、政治等四个方面，依次占语料库总量的 19.6%、50.3%、9.8%和 22.9%，录用建国后发表的文章作为选取语料的时间限制。扩充到 500 万词级时，增加了一部分数理化教材和医学、法律等方面的内容。

c) 自动校对

语料库的所有文本输入工作都是通过键盘输入法来完成的。因此，随之而来的问题自然是令人头痛的文本校对问题。为省事省时，我们编写的第一个工具软件就是面向拉丁转写文本的现代蒙古语自动校对程序。这个程序，以蒙古语正字法作为基本依据，利用词干词典、附加成分列表以及相关的规则来校验文本中各单词的拼写合理性，能够指出词干、动词附加成分、分写附加成分、阴阳元音同现等四个方面的拼写错误。有了自动校对工具软件，人工校对的侧重点自然就转到检查文章内容是否遗漏、重复等比较粗的方面，劳动强度得到了大幅度减轻。

d) 动词附加成分的自动切分和复合词的自动识别

蒙古语的文本处理，没有象汉语那样做词切分的要求，但要求识别词干与附加成分的界限、复合词与普通单词的界限。蒙古语是黏着语言，文本中大量出现词干加附加成分构成的词。人工标记词干与附加成分的界限，不仅工作量大，而且很容易出现遗漏、不统一等问题。于是我们开发了第二个工具软件—面向拉丁转写文本的蒙古语词干、词根、词尾的自动切分和复合词的自动识别程序。这个程序，对动词附加成分、动词干、连结元音的识别率较高，但对复合词的识别能力较低。

e) 词类标注

词类标注，是我们为使生语料变成熟语料而做的实质性加工。对蒙古语语料库的词类标注系统，借鉴了英语语料库和汉语语料库的词性标注算法。我们选择 10 万词语料作为词类标注训练集，建立了词类同现矩阵，结合规则方法和统计方法（CLAWS 算法）解决了兼类词歧义问题，标注准确率达到 95%左右。

f) 短语切分和标注

我们正在研究对蒙古语语料库进行短语标注问题。蒙古语传统研究的有关理论以及近几年我们所做的面向信息处理的蒙古语固定词组研究、蒙古语电子词典框架研究和对蒙古语语料库的词类标注研究等成果，为我们进一步对蒙古语语料库做短语标注提供了前提条件。蒙古语短语标注研究已经获得国家社会科学基金和自然科学基金的资助，准备在 2005 年年底完成。

到目前为止，我们还没涉及有关语义、语用标注研究。但是，我们积累了不少有关语料库建设方面的基本经验，对国内外语料库语言学研究情况有了比较清晰的了解。现在，完全可以利用我们所建立的现代蒙古语语料库来获得词频统计、词类统计、附加成分统计、音节统计、句子统计结果。

3 存在的问题

我们在建立蒙古语语料方面虽然做了不少工作，取得了一些成绩，但是，与国内外其他语种的语料库建设相比，我们的差距还是很明显的。除了语义、语用标注等大的建设内容还没有启动外，与现代蒙古语语料库建设相关的问题大概可以提以下几点：

a) 蒙古语语料库文本还没有用蒙古文编码国际标准来建立。这对研究人员虽然没有多少影响，但对蒙古文直接用户带来了一定的不便。

b) 目前所使用的蒙古文拉丁转写规则不符合蒙古文拉丁转写国际标准。这对信息交换、资源共享不利。

c) 录入员的负担比较重。就是说，因为蒙古语相关的基础研究还没有跟上，录入员还在承担对附加成分、动词干、人名、地名、不规范书写形式和复合词等做标记的任务。

d) 语料的覆盖面不够充分。不仅缺少自然科学、口语方面的语料，有关哲学、政治、经济、法学等领域的语料还需要补充。

e) 自动校对程序缺乏语法校对功能，即还没有词语搭配识别功能。

f) 附加成分的自动切分和复合词的自动识别程序对复合词的识别率低、识别范围小，大量的复合词仍需要通过人工方式处理。

g) 现代蒙古语语料库对词语的划分是粗线条的，它虽然把蒙古语词类划分为 14 类，但没有对这些词类再进一步进行子类划分。对词义歧义处理、句法分析等，词语的多层次分类是不可缺少的。我们在研制开发《汉蒙机器翻译系统》和《蒙古语语法信息词典》的过程中，根据信息处理多方面的需要，初步确定了一个面向自然语言处理的蒙古语词语分类体系及其标记集。其中，包含基本词类 15 个，附加类别 6 个。对基本词类，还做了子类划分的尝试。但这些分类还没有用到现代蒙古语语料库中。

h) 标记附加成分、不规范书写形式、复合词和词类的符号使用情况需要进一步规范化。

4 未来的目标和准备采取的策略

a) 至今为止，我们还没有得到存在比 500 万词级更大规模的现代蒙古语语料库的消息，可其它语种语料库规模达到几千万、上亿个词次的报道很多。我们从蒙古语语料库的应用情况也感觉到 500 万词级的语料库确实有些不够用。尤其是基于统计方法来归纳某个方面的规律或处理某个方面的歧义问题时，这种感觉更为突出。我们准备在“10·5”和“11·5”期间，扩大现代蒙古语语料库的规模和覆盖面，完善词类标注、句法标注的同时要启动语义标注、语用标注等深加工研究，使得现有的现代蒙古语语料库上升为我国国家级蒙古语语料库。这样做，不仅有它的学术价值，而且对使用蒙古语言文字的地区的信息化、现代化建设产生巨大的推动作用，还会提高我国在国际蒙古学研究领域内的地位，产生良好的社会效益。

b) 为实现建立国家级现代蒙古语语料的目标，我们将扎扎实实地开展面向蒙古语信息处理的基础研究，采取行之有效的技术措施，从录入到各类标注加工，提高各个环节上的形式化、规范化、自动化、智能化程度，研究开发多种多样的工具软件和应用系统，为从多视角研究蒙古语提供知识资源和技术资源。其中，从蒙古语本身出发，做好基础研究是最根本的。我们正在承担两个国家课题：面向信息处理的蒙古语语义研究和蒙古语语料库短语标注研究。完成这两项研究，会在较大程度上推进蒙古语语料库的加工处理进程。从研究方法的

角度来说，我们将继续采取理性主义和经验主义方法相结合的方式。这种途径，对蒙古文信息处理是非常合适的，蒙古语语料库词类标注系统的成功就是一个很有说服力的实例。

c) 还有，为了面向国际、面向蒙古文直接用户，我们将按照蒙古文拉丁转写国际标准改写我们的蒙古文拉丁转写规则，实现从拉丁转写到蒙古文编码国际标准的转换。蒙古文编码国际标准已经得到ISO的批准，蒙古文拉丁转写国际标准方案，估计今明年也会得到ISO的批准。有了这些条件，在网络环境下共享蒙古语语料库资源就有了保障。现在，我们已经编写了一个文本转换程序，实现了转换现有蒙古文拉丁化文本为蒙古文国际标准编码文本。

d) 建立汉蒙并行语料库，为实现基于实例的汉蒙机器翻译系统打基础。汉蒙并行语料库，自然会包含极其丰富的汉蒙对照信息，它将成为在多领域内开展汉蒙双语并行处理研究所必需的基本资源。实现基于实例的汉蒙机器翻译系统，更是需要较大规模的汉蒙并行语料库为它做后盾。早日实现实用性汉蒙机器翻译系统，对于同时用蒙汉两种文字及时传播党和国家的方针政策，加快实施我国西部大开发战略，促进民族地区的跨越式发展都具有十分重要的意义。

总之，语料库语言学在语言信息处理过程中扮演越来越重要的角色。可我们知道，语料库建设是一项基础建设，需要付出艰辛的劳动。从我们的国情出发，特别是根据我们蒙古民族居住生活的西部地区实际出发，建立一个具有相当规模和功能完善的蒙古语语料库，除了认真考虑人力、物力、财力方面的因素之外，更重要的是我们要考虑如何从蒙古语的自身出发，开展与信息处理相关的基础研究，以便满足来自蒙古文信息处理实际的各种形式化、数字化需求。可以说，这是蒙古文信息处理成败的关键所在。

参 考 文 献

1. 内蒙古大学蒙古语文研究所计算机室，关于《现代蒙古语文数据库》，内蒙古大学学报，1 992 年第 1 期。
2. 华沙宝，《现代蒙古语文数据库》程序设计，内蒙古大学学报，1992 年第 2 期。
3. Craeme Kennedy，语料库语言学入门，外语教学与研究出版社，2000 年。
4. 陈小荷，现代汉语自动分析，北京语言文化大学出版社，2000 年。
5. 冯志伟，计算语言学探索，黑龙江教育出版社，2001 年 • 哈尔滨
6. 黄昌宁，李涓子，语料库语言学，商务印书馆，2002 年 • 北京