关于进一步提高蒙古语语料库质量的建议

艳花

（内蒙古大学 蒙古学学院蒙语所，内蒙古 呼和浩特 010021）

摘要：本文回顾了蒙古语语料库建设情况，并比较参照了蒙古语语料库与其它语种的大型语料库之间的差距，就进一步提高蒙古语语料库质量方面提出了几点建议。

关键词：大型语料库；蒙古语语料库；比较；建议

中图分类号码：H212　　**文献标识码：**A

一、引言

通过几代蒙古语文工作者的努力，我们在语料库建设方面取得了一些可喜的成就。中世纪蒙古语语料库（1984 年）、100 万词级现代蒙古语文语料库（1991 年）、500 万词级现代蒙古语文语料库（1998 年）、契丹小字语料库（2000 年）、八思巴字文献语料库（2001 年）、蒙古语口语材料语料库（2004 年）等相继问世。这对蒙古文信息处理工作带来了福音。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工作提供基本信息的能力。但是由于蒙古文信息处理工作开展的较晚、基础研究和应用技术等方面比较薄弱、人手不足、资金短缺等种种原因，跟英语、汉语等其它语种的语料库建设相比，我们的语料库还有很多不足之处，还有待于改进。该论文就提高蒙古语语料库质量方面，提出了以下几点建议，予以参考。

二、几点建议

（一）必须进一步扩大语料库库容。

随着计算机技术的发展和应用，语料库的建库变得越来越容易。许多语言资料都可以用万能扫描仪来录入。现在，语料库的规模越来越大。有 Brown 语料库、COBUILD 语料库、Longman 语料库、英国国家语料库 BNC、国际英语语料库 ICE、LOB 语料库等有几千万甚至几亿词级规模的语料库已被建立。语料库的建设是利用语料库进行语言研究的基础。虽然语料库规模的大小不是衡量语料库质量的唯一标准，但是语料库的建设是进行语言研究工作的基础。我们现在虽然已经建设了 500 万词级的现代蒙古语文语料库，并正计划扩充到 1000 万词级，但这远远不能满足蒙古文信息处理工作的需求。我们可以通过下面的简单表格就能看出我们的语料库有很多不足之处。所以我们必须加紧进行语料库扩充工作，与国际接轨。

表 1：蒙古语语料库与其它语料库规模比较

语料库	规模
Brown 语料库	100 万词级
COBUILD 语料库	5 亿词级
Longman 语料库	2800 万词级
英语国家语料库 BNC	1 亿词级

国际英语语料库	2000 万词级
现代汉语研究语料库	2200 万词级
现代蒙古语文语料库	500 万词级

（二）增强语料库的多样性。

一种语言的语料库反映的是本语言在人们日常生活当中的真实情况。语料库的多元化，对语料库的质量有举足轻重的作用。一个广泛、全面、均衡地收集各方面材料的语料库对语言的研究和分析是非常有利的。下面我们简单比较一下蒙古语文语料库和英语语料库的语料分布情况。

我们已经建立的 100 万词级现代蒙古语文语料库的语料分布：蒙古语文教材有 50 万词左右，占 50.3%；政治类有 20 万词左右，占 20.3%；文学类有 20 万词，占 19.6%；报纸新闻类有 10 万词，占 9.8%。

90 年代建立的“英语国家语料库”（简称 BNC）包含一亿个词。其中有 9000 万词的书面语文本和 1000 万词的口语文本。书面语文本有两大类：信息性部分和想象性部分。信息性部分从主题范围、文本类型和层次三个方面规定了百分比。按主题范围划分：自然科学 5%，社会和社区 15%，商业和金融 10%，信仰和思想 5%，应用科学 5%，国际事务 15%，艺术 10%，休闲 10%。按文本类型划分：书籍 55－65%，报刊 20－30%，其它各种印刷出版的文本 5－10%，各种非出版的文本 5－10%，用于口述的文本 2－7%。按层次分：专业性文本 30%，非专业性文本 50%，普及性文本 20%。想象性文本占书面语文本的 20%－30%。按文本的层次分为三类：纯文学性的、中间的和通俗的各占 30%。从上面的数据不难看出，我们的语料库覆盖面太窄了。

我们不如再降低一下标准，和同样规模的英语语料库参照以下。20 世纪 60 年代由布朗大学建立的 Brown 语料库的规模也是 100 万词级。我们可以通过下面简单的表格来了解一下它的语料库分布情况。我们会更深刻地感觉到必须扩充蒙古语文语料库多样性的紧迫性。

表二：Brown 语料库的语料分布情况①

编号	语料体裁	样本数
A	报刊：新闻报道	44
B	报刊：社论	27
C	报刊：评论	17
D	宗教	17
E	技能和爱好	36
F	流行的传说	48
G	纯文学的、传记和自传	75
H	混杂的	30
J	教学	80
K	一般小说	29
L	侦探小说	24

M	科幻小说	6
N	冒险和西部小说	29
P	浪漫爱情故事	29
R	幽默	9
总数		500

（三）语料库的内容也要与时俱进。

世界上的任何事物无时无刻都在运动、变化着，语言也不例外。我们生活当中每时每刻都有新词出现，旧词被淘汰。如果我们不加紧改进语料库内容，会对在语料库基础上进行的研究工作带来不便。比如，我们要对语料库进行词频统计，80 年代末 90 年代初建立的语料库当中词频最高的肯定是有关改革、市场经济方面词。而如果对现今的语料库进行词频统计，词频最高的可能会变成知识经济、网络等方面的词。所以我们必须时刻更换语料库内容，要不然在此基础上进行的查询、统计数据等将不能精确地反映该语言的应用现状。因此我们必须建立一个动态性的语料库。

（四）应该建设蒙古语语料库平台。

从 1983 年研制的《元朝秘史》文件检索系统来计算，我们至今也有了十几个蒙古语语料库。但到现在我们还没有系统地归纳这些语料库。各语料库都分散在不同的文本中。所以我想我们应该借鉴其它语种的大型语料库的建设经验，建立一个蒙古语语料库平台。这将对文件检索、统计分析、提取信息等工作带来很大的便利。如果我们在网上查看 Brown 语料库，我们会从该语料库的网站上可以查到关于 Brown 语料库很多信息。有 Contents；Versions of The Corpus；Coding Procedure of From A ； The Tagged Version；List of Tags； Copyright Restrictions；Basic Technical Information；The Individual Samples；List of Samples 等目录。各目录也都有更详细的信息。这些信息包括语料库建设说明，语料库内容，该语料库的不同的版本，标注版本，标注语料库排序，技术信息，配套的检索软件，语料库分类信息，每个语料文本的作者、主题、题材、年代等信息，语料采样原则等等很多信息。这些成功的经验应该成为我们的指路灯。

（五）我们应该充分利用现有的资源。

自现代蒙古语语料库诞生以来，我们主要是利用它来检索，统计蒙古语词在真实文本中的使用情况。虽然我们建立了 10、20 万词级测试语料库来进行了词类标注，建立了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，还进行了短语切分和标注工作，有一些论文形式的成果接连出现，编纂了蒙古语词频词典。但这些工作做得还不够。毕竟语料库是一个用之不尽的资源，我们可以利用它开展更多的研究工作的。这方面还是英语比较处于领先地位。如，COBUILD 英语语料库自建立以来，在该语料库的基础上已经编辑了 The Collins COBUILD English Language Dictionary, The Collins COBUILD English Grammar, The Collins COBUILD Student's Grammar，The Collins COBUILD English Language Dictionary，The Collins COBUILD Dictionary of Phrasal Verbs，A School Dictionary，Collins COBUILD English Guides，Collins COBUILD English Usage 等七部词典。目前，还正在编辑同义词大全和各种语法书。并且他们已经对全部语料库进行了词性标注，还对近两亿词的语料已经进行了句法分析。此外，该语料库的建设者也为语言学家及用户提供了很多复杂的软件。这些软件可以完成搜索特定词的组合模式、查找一个词的词频、找出词的使用实例并进行分析等等任务。看看别人，再想想自己。我们还有很多的工作要做。

（六）要加紧现代蒙古语口语语料库的建设。

纵观其它语种的大型语料库，我们会发现它们都是有书面语语料库和口语语料库两部分组成。像 BNC 英语国家语料库、现代汉语研究语料库、国际英语语料库等。现在，我们虽说 2004 年已经建

立了蒙古语口语语料库，但我觉得那不是真正的口语语料库。只不过是一次尝试性试验。我们自己也有演讲、辩论、广播、课堂讲授、讨论、采访等等建立口语语料库的选材空间，为什么不建立一个完完全全的蒙古语口语语料库呢？

（七）要进一步规范各种标记符号、不规范的书写形式。

建库之后，录入人员以及研究人员对 100 万词级现代蒙古语文语料库进行了七、八次乃至十几次的校对，但我们的语料库还存在不少问题。还有零零星星的录入错误，有些符号不统一，对人名、地名的标记不符和规则等等问题，影响着我们语料库的质量。所以我们必须规范这些细节内容，不能让它们影响语料库的质量。

（八）要尽快统一蒙古语词类标记集。

对语料库进行词法、句法、语义、语用标注将会大大提高语料库的可用性。目前，我们在词类标注、句法标注等方面做了一些工作，还没有涉足语义标注、语用标注等研究。其中在词类标注方面，至今蒙古文信息处理工作中使用过的词类标记集当中最有影响力的有两种。即蒙古语词类标注系统 AYIMAG 所使用的蒙古语词类标记集和面向信息处理的蒙古语词语分类及标记集。现在，虽然大家普遍承认面向信息处理的蒙古语词语分类及标记集简单明了、跟国际上通用的标记吻合、对非词语也制定了相应的标记等优点。但目前为止，蒙古语词类标记集还没有最后统一起来。标记集中的某些细节问题还有待于讨论。

（九）现代蒙古语语料库对文当、文本编辑、检索软件没有一个统一的管理系统。

今后我们应该建立一个数据库管理程序，对蒙古语语料库进行维护是必要的。

三、小结

总之，今后语料库语言学在语言信息处理工作中会扮演越来越重要的角色。对我们蒙古语而言，建立一个广泛的、全面的、多样化的语料库不仅功在当代。而且对我们蒙古文化的信息化、国际化都会有推动作用。所以我们应该不惜投入人力、财力、物力来提高蒙古语语料库质量。我们将翘首等待蒙古语语料库越来越完备。

注释

①黄昌宁，李涓子.语料库语言学[M].北京：商务印书馆，2002.51-52.

参考文献

[1]黄昌宁,李涓子.语料库语言学[M].北京：商务印书馆，2002.

[2]华沙宝.关于蒙古语语料库建设[M].中国，北京。2004，4.

[3]王建新.介绍当代三个英语语料库[J].外语教学与研究,1996，(3).

[4]雪艳，文化，那顺乌日图.蒙古语语料库综述[R].首届全国少数民族青年自然语言处理学术研讨会，2004 年，呼和浩特.

[5]网站：Brown Corpus：http://khnt.hit.unib.no/icame/manuals/brown/index.htm

[6]网站：The Bank of English：http://titania.cobuild.collins.co.uk/boe-info.html

[7]网站：British National Corpus：http://info.ox.ac.uk./bnc

YAN Hua

(Academy of Mongolian studies, Inner Mongolia University, Huhhot 010021, China)

Abstract: The paper briefly reviewed the situation of constructing the Mongolian corpus，Moreover, comparatively consulted and followed the disparity between else languages large corpus together with the Mongolian corpus, and author proposed some advise to improve the Mongolian corpus quality.

Key words: large corpus ；Mongolian corpus ；compare；Proposition

收稿日期: 2006-12-15;

作者简介: 艳花（1979－）,女,蒙古族，内蒙古科左中旗人。内门古大学蒙古学学院硕士研究生,主要从事蒙古文信息处理方面的学习与研究。