

蒙古语语料库 COMS 设计

马丽¹，宋建斌²，敖其尔³

(内蒙古大学 计算机学院，内蒙古 呼和浩特 010021)

摘要：本论文介绍了语料库的基本特点和类型，分析了蒙古语语料库系统 COMS(Corpus Of Mongolian Systrm)建设的必要性。给出了 COMS 的词表结构和结构模型，这些结构模型包括：文本模型、数据库模型和 FDB 模型。接着，介绍了 COMS 在统计分析、语音合成和语音分析等方面的应用。最后，对本课题开展的工作做了总结。

关键字：蒙古语；语料库；结构；应用

中图分类号：H085.6

文献标识码：A

“语料库语言学(Corpus Linguistics)是 80 年代才崭露头角的一门计算语言学的新的分支学科。它研究机器可读的自然语言文本的采集、存储、检索、统计、语法标注、句法语义分析，以及具有上述功能的语料库在语言定量分析、词典编纂、作品风格分析、自然语言理解和机器翻译等领域中的应用”。[1]语料库语言学研究的基础是机器可读的大容量语料库和一种易于实现的统计处理模型，两者是相辅相成、缺一不可的。

蒙古语是应用比较广泛的语言，它在发音、韵律，词汇及其词的形态变化，组织句子结构，文章的构成等许多方面都具有独特的特性。蒙古语语料库系统 COMS (Corpus Of Mongolian Systrm)的建设是蒙古语研究的不可缺少的一部分。随着对蒙古语研究的逐渐深入，比如机器翻译、语音合成、语音识别、文字识别、智能检索等的研究，蒙古语各个特性的量化描述需求越来越大。正是蒙古语料库为蒙古语研究提供了确凿的数据。国内外各种语言的语料库研究有很大的进展，并逐渐成熟。所以，作为蒙古文信息处理平台(MIPP)的一个重要的子课题，蒙古语语料库 COMS 的建设与研究，有它的理论价值和使用价值。

一、语料库的基本特征和类型

(一) 语料库的基本特征

语料库来自拉丁词 corpus,原意为“汇总”、“文集”等,复数形式为 corpora 或 corpuses。语料库是按照明确的设计标准,为某一具体目的而集成的大型文本库。Renouf 认为,语料库是“由大量收集的书面语或口头语构成,并通过计算机储存和处理,用于语言学研究的文本库”[2]。

语料库具有以下基本特征：

(1) 具有明确而具体的研究目标。语料库是按照明确的设计标准,为某一具体目的而集成的大型文本库。语料库的设计和建设是在系统的理论语言学原则指导下进行的,语料库的开发具有明确而具体的研究目标。

(2) 具有一定的容量。随着语料库的发展,语料库变得越来越大,由各种商业和学术组织负责收集的大型电子数据、电子文档也纷纷建成。这些大型电子文档广泛收集了各种类型的数量可观的电子文本,以满足各种用户的不同需求。

(3) 具有人机交互特性。计算机语料库实际上提供了一种人机交互,这种交互方式随

着语料库工具的发展而逐步加强其自动化特性。

(4) 自动化特性。如今人们所谈的语料库，实际上就是指计算机语料库。语料库语料以电子文本形式储存，并且是通过计算机自动处理的。计算机技术的发展和相关软件的开发，使语料库语言学得以蓬勃发展。

(5) 资源丰富。可获得的语料资源丰富，获得渠道方便。如今大量的在线语料资源，光盘资料，因特网资源，包括新闻、邮件列表、电子邮件等，使得语料库的建设和扩充变得非常快捷方便。

(6) 多样化发展。语料库类型和品种数量的高速增殖。一方面，同一语言的语料库在类型上出现分化，针对不同文类和领域建成了各种专用语料库。另一方面，语料库在品种、数量上呈现高速增长的趋势。

(7) 量化研究。语料库既是一种研究方法，又代表着一种新的研究思维，并以当代先进的计算机技术为技术手段。

基于语料库的语言分析与研究已深入人心。任何对语言进行描述和分析的研究如果想得到可靠的发现，就必须以大量的、客观的语言运用量化数据为依据。语料库是语言学研究可靠的数据源，基于语料库的研究方法是语言学研究的基本方法之一。通过语料库可以得到语言的定量数据。

基于语料库的研究以量化研究为基石，以概率统计为基本手段，以“数据驱动”为基本理念。其基本方法是通过对实际语言运用的抽样，确定其对语言整体的代表性，通过对样本特征的描述概括整体特征。

(二) 语料库的类型

语料库的类型可以从以下几个层面来划分：

- (1) 语言种类：可分为单语语料库、双语语料库以及多语语料库。
- (2) 语料信道：可分为书面语语料库和口语语料库。
- (3) 载体媒介：可分为印刷文本、电子文本、口语转写、影像等。
- (4) 语料状态：可分为共时语料库和历时语料库。
- (5) 应用层面：可分为通用语料库和专用语料库。
- (6) 语料处理：生语语料库和标注语料库。

根据编码标注的复杂程度，又可细分为以下几种情况：

- (1) 不加任何处理的纯文本语料库。
- (2) 经过格式属性标注的语料库，如对段落、字体、字号进行标注。
- (3) 对识别信息进行标注，如作者、体裁、语域等；以及词性标注。

(4) 殊标注，如错误附码等。经过标注和附码的语料库，使得语料库数据分析更加系统精确，也便于对特殊数据信息的提取和处理。但是，不经任何人工介入的生语语料库同样具有独特的价值。在语料库建设中，一般是保持一个干净的生语语料库，而把经过标注和句法分析的语料另存为一个子语料库或者独立的版本。

(5) 语料库有“静态的”，还有“动态的”；即该语料库是有限的，又有无限的。有限的语料库只收集某一固定时期的共时语言材料，语料库建成后，就不再扩充。

二、蒙古语语料库 COMS 的框架结构

（一）蒙古文的特点

蒙古文是拼音文字，是属于阿尔泰语系的蒙古族文字。蒙古文有 31 个字母，其中有元音 7 个，辅音 24 个。蒙文的字典排序严格按照字母排序进行。蒙文的拼写规则是以词为单位竖写，词与词之间用空格分开。一个词内各个字母连写。总的书写规则是，蒙古文采取从上到下的书序，从左到右的行序。蒙古文中的词又分阳性、阴性和中性。每个字母在字首、字中、字尾有不同的变体，有的变体有多种形式。所以，31 个字母有一百多种左右书写变体形式。这些变体字母的书写形式的物理形状的长短各异，存在着一字多形、多字同形的现象。在蒙古文当中有一套非常复杂而严密的正字规则和语法规则。这些特点给蒙文信息处理带来许多困难。

对语料库进行标注和统计，将用于解决随研究过程而出现的一些问题。如蒙语语音合成、蒙古语文字识别后处理、蒙古语语音识别、蒙汉电子词典等。

（二）COMS 基本框架

COMS 建设具有具体的目标。为实现目标，可以采用适合于自己的结构框架。COMS 是为了解决蒙古语言文字研究中的问题而建设的，并且它是一个开放的、动态的、易用的。它可以不断地修改、更新、扩充。

COMS 框架结构见图 1：

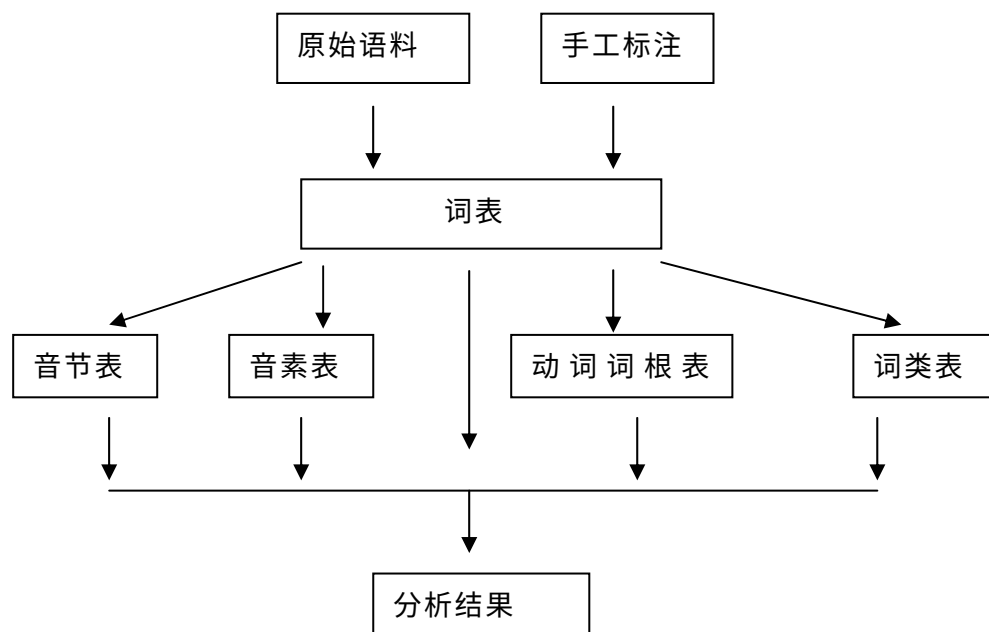


图 1：语料库的框架

COMS 是加工原始语料而形成的。原始语料是未经加工的原始文本。蒙古语的输入方法有几种，如直接输入法、蒙文拼音输入法和蒙文拉丁转写输入法等。为简单起见，COMS 的原始材料，没有考虑蒙古文的书写形式，直接采用蒙古文拉丁转写输入法建立。

从原始语料中读入蒙古文，对它进行预处理。即从文本中只读取蒙文词，其它标点符号、数字等非词对象被略去，把词加到语料库当中，形成词表。

词表是标注的基础。所谓的标注就是用一些符号代表一定的意思，对语料做标记。我们 COMS 中

做了分音素、分音节、词类、词的汉语意思等标注。在标注过程中,自动和手工结合。自动标注音素、音节等比较简单的标注、手工标注词义等。

从已做标注的词表中得到音素表、音节表、词类、动词词根表等统计表,也可以根据需要生成更多的表,且对各种统计表进行分析。

(三) 语料库的表结构

本语料库表结构是用 Microsoft Access 2000 数据库实现的。Access 提供一种建立表与表之间“关系”的方法。词表、音素表、音节表、词类表、动词词根表的数据之间都存在这种“关系”,这种关系将数据库里各张表中的每条数据记录都和数据库中唯一的主题相联系。只需要通过一个主题就可以调出来使用,非常方便。

每一个表中设有相关的字段,例如,词表中有词类、词的音素、词的音节、词的汉语意思、词频和音标;音素表中有“蒙文读音输入法”对应的 35 个音素及其统计数据;音节表中有可能出现的所有音节及其统计数据。下面详细论述词表。

1. 词表数据的组成

词表数据以词条为单位。如果把词典数据看成是一个文件的话,那么一个词条就是词表文件的一条记录。一个词条由若干个词表字段项组成,即词条记录中的若干字段。一个词条一般由以下几个字段项组成:

- (1) 词目部分词目:一个词条所处理的词。
- (2) 音标:某些特殊的词,如缩略语、词缀、构词成分等,可能没有音标。
- (3) 曲折变化形式:列出单词不规则的变化形式。
- (4) 词性:标注每一个词在语言环境中所具有的词类。
- (5) 音素:一条词目由哪些音素组成,在这里用空格隔开每一个音素。
- (6) 音节:一个词的所有音节,两个“/”中间表示一个音节。
- (7) 中文释义号:词条有多种释义时,用 1 2……等数字表示。

(8) 中文释义:大体相同的释义列在同一个释义号下,词义相近及较远的释义用不同的符号表示。

2. 词表数据的特点

由于词表数据的特殊性,词表数据字段多,关系复杂,有以下特点:

- (1) 词表规模不同,词表收录的词表字段项也有所差别。
- (2) 同一词表中各词条中出现的字段个数、位置、次数完全相同。
- (3) 复合字段项多。一些字段是复合字段项。如参见字段中包含多种参见信息。
- (4) 位置关系复杂。多数字段项出现位置不定,而且处于不同的位置其含义亦有所不同。如说明信息可以出现在词性中,表明对词性使用方法的说明;出现在所有释义的前面,表明对所有释义的说明;出现在释义项中,表示对该释义项的说明。
- (5) 例外情况多。词表编纂前先确定词表体例结构(即词表数据的表示形式,相互之间的位置关系等),但实际词表数据中有很多例外情况。如一般规定一个词条中只说明一个词目,但实际上有很多例外。所有这些都给词表数据的组织带来了很大的不便。

3. 结构模型

(1) 文本模型：以文本文件的方式组织词表数据。用标识语言标识每一个词表字段项。用此方式组织词表数据时,可以很好地区分每一个词表字段项(只需要在字段项的前后加上标识符号即可)。基于文件方式管理大量数据,给词表数据的维护带来了极大的方便。因而它仅适合固定的词表数据。

另外,还需要专门设计一个与该词表标识语言配套的检索软件,以保证从词表中查找(定位)到相关的信息。

此种模型适合于光盘版、磁带版或单机运行的词表数据。

(2) 数据库模型：当词表数据处于不定型状态,需要经常修改时,文本模型显然不合适了。此时必须采用数据库管理系统进行管理。目前关系型数据库(RDBMS)是数据库的主流,新推出以及流行的数据库管理系统都很好地支持关系模型。

由于词表结构本身的特点,用 RDBMS 描述词表数据结构并不是非常合适。首先,根据 RDBMS 的范式,字段应是最小的不可拆分的元素。因此,每个词表字段项必须用一个单独字段项描述,这样就必须考虑许多独立的字段项以及它们之间的关系。其次,要描述词表字段项之间复杂关系,必须有若干个专门的表。这样就会使得词表数据被分散在每个表中,也就意味着在处理时将跨越多个表,从而加剧了处理复杂程度,效率明显下降。第三,词表数据中出现了许多例外情况,关系数据库不能穷尽所有的处理方法。

(3) FDB 模型：网络环境下,词表数据的文本模型不能够很好的管理数据,也不利于词表数据的各种处理。在 C/S 体系结构中,对词表数据的管理应采用大型数据库管理系统。用纯关系模型管理词表数据是不合适的。我们需要在纯关系数据库和纯文件管理之间找一个平衡点,既有利于数据管理和处理,又能提高系统的效率。

最后,采用文本+数据库(FDB)的数据组织方式是比较好的。整个词表数据结构是基于数据库管理系统(DBMS),但是数据库中的字段不是原子数据,每个数据库字段项不对应于词表数据中的一个词表字段项。数据库字段项是一个复合项,包含多个词表字段项,每个词表字段项之间用词表标识符号标识。采用这种结构后,由于一部分词表的字段项之间的关系被隐藏在数据库的字段项中,因而词典数据库中对词典字段项之间的关系描述相对少一点,从而大大减少了词表数据库中的表的数目,这样就在运算时减少了数据库之间的关联,可以提高效率。当然,由于数据库中字段项不是基本字段项,所以在处理时,对字段项中的内容还需筛选和过滤。这样才能得到实际的结果,因而需要有一组对字段内容的处理方法。

MFC 提供了访问 ODBC 的接口,不用考虑数据库的交互细节。并对 Microsoft Access 数据库提供了更好的支持。Visual C++6.0 有相当强的数据库编程功能。用 Visual C++6.0 通过 ODBC 访问 ACCESS 数据库可以提高建语料库的效率,对大量的统计工作带来很大的方便。

用 VC 的数据库编程和 Access 数据库管理功能来管理语料库可以兼顾自动和手工的工作。让自动和手工工作有机地结合起来,使语料库的管理变得简单、易用、开放。

三、蒙古语语料库 COMS 的应用

计算机能处理大量的数据,从而使 COMS 的建设与应用变得方便。语料库数据来源覆盖面大,相对全,不象人工处理的数据那样数据少而语料库的应用受到限制。通过计算机处理,大量的自然语言数据可为研究工作提供可靠的依据。语料库不仅可以用于机器翻译,还可以做语音分析,指音段分析(主要用于语音识别和语音合成),词法分析(指词性、词的组成的分析)。

(一) 统计分析

在语料库语言学中,基于统计的处理技术是从语料库中获取各种所需要的知识的主要手段。它的基本思想是：

使用语料库作为唯一的信息源,所有的知识(除了统计模型的构造方法)都是从语料库中获得的。使用统计方法获取知识。知识在统计意义上被解释,所有参量都是通过统计处理从语料库中自动习得的。

蒙古语是一种拼音文字,它的书写是按词连写的。虽然蒙古语也有分写形式,但在拉丁转写输入法中,与词对应的拉丁转写码不会有词的分写形式。词与词之间用空格隔开。所以对蒙古文章进行分词只以空格分界就可以了,从而便于做词频统计。

蒙古语的音素与“蒙文读音输入法”的35个码一一对应。所以对词进行分音素,只按输入码分开即可。因此,音素统计数据反映在输入码的统计数据上。

蒙古语音节有“辅音+元音”、“元音+辅音”、“辅音+元音+辅音”、“辅音+元音+辅音+辅音”等情况。考虑到这几种情况,我们可以分音节,并做音节统计。

（二）语音合成与语音识别

语音合成是计算机智能接口的一个重要分支,而语料是进行语音识别、合成的起点。“语料是指包括有可能利用来发现和研究语言学和语音学知识的一切素材”[3]。在语音合成中,对语料库的管理工作大致可归纳为两部分:一是对原始语料文本资料的管理;二是对大存储容量录音数据的管理。建立起语音数据与文字描述之间的联系,是一件繁琐细致的工作。通过语料库管理系统,将语音的文字描述(如字、音、调)与语音数据、语音波形显示有机的结合起来,展现在用户面前是一个文字、声音、图形综合管理的语音数据库。通过语料库,完成对各类语音合成基元的分析、统计、检索等研究,也力图为语音合成的评测提供素材。

蒙古语语料库的语音分析是 COMS 建设的主要目的之一。蒙古语在口语中的发音与书面语中的发音有很大的差别。利用语料库可以知道各音素在不同位置上的发音变化,例如音的同化、异化、颠倒等现象的统计数据;音节之间的“连读”,也就是读词时各音节之内的元音被省略,从而发生一词在口语中的音节数与书面语中音节数不同,得到某词在口语中音节数变少的相关数据。在语料库中,标注词的读音(口语读音),得出双音子、三音子出现的统计数据,从而得到一些出现频率高的音子;在合成中,直接用这些音子的音,而不是从音素或音节的拼接得到这些音子的音,进而达到基于 PSOLA 算法的语音合成的最佳效果。

（三）形码到音码的转换

音形转换是蒙古语自动转换问题中很关键的问题。它的解决对于人机自然语言交互通讯、蒙古语的键盘输入和汉语语音识别及合成都有重要意义。然而蒙古语的音字不是一一对应,即一音多字、一字多音的现象。这给音形转换的解决带来了极大的困难。语料库语言学的发展,为研究者提供了一种新思路。

蒙古语语料库的词法分析也是目的之一。蒙古语中有“同形不同音”现象,我们通过以下方法来区分这些词:一,我们通过词的阴阳性来解决一部分词;二,还可以通过音节的构成规律来区分;三,可以在语料库当中查找要区分的词,找出相匹配的对象。

还有一些词有两种读法,用上述几种方法,如它的阴阳性、音节构成或词一级的语料库中查找的方法等,都无法知道它的读法,只能从词所在的文章或说话环境(上下文)来决定其读音。

（四）文字识别后处理

在蒙文文字识别过程中,我们不太可能保证百分之百的正确识别,总有一些词会识别错误。这时可以用语料库的统计数据作处理。在语料库中,用可能识别错的词去串匹配,匹配程度最高的词就是与被识别词对应的正确词。

在 COMS 建设中, 已经从原始语料中提取出了词汇, 形成了词典, 作了音素、音节标注, 并得到了词频、音节、音素的出现频率的统计数据。

COMS 是动态语料库, 它可以在使用过程中不断地更新。其它的统计数据, 如: 双音子、三音子、一个音节与音节同时出现的可能性等, 都可以对语料库稍微进行加工而得到。

COMS 的建成, 对于各种基于统计的蒙古语处理技术的发展, 无疑会起巨大的推动作用。

参考文献

- [1]清格尔泰. 蒙古语语法[M]. 呼和浩特: 内蒙古人民出版社, 1991. 251-P328.
- [2]<http://ps.henannu.edu.cn/fl/project/features.htm>
- [3]周强. 基于语料库和面向统计学的自然语言处理技术介绍[J]. 计算机科学, 1995, (4). 3-40.
- [4]赵铁军. 机器翻译原理[M]. 哈尔滨: 哈尔滨工业大学出版社, 2000.
- [5]赵世霞, 蔡莲红, 常晓磊. 汉语语音合成语料库管理系统的建立[J]. 小型微型计算机系统, 2000, (3).
- [6]李灿伟, 杨震. 基于 HMM 的说话人确认系统的研究[J]. 南京邮电学院学报(自然科学版), 2001, (6).

Design of mongolian corpus COMS

MA Li¹, SONG Jian-bin², Ochir³

(College Of Computer Science Inner Mongolia University, Inner Mongolia, Hohhot, 010021)

Abstract: In this paper I intend for introduction of characteristics and sorts of Corpus. And this paper analyses the necessity of COMS construction, introduces structure design of COMS(Corpus Of Mongolian System). Word-table structure and structural model of COMS can be found in this paper, including text model、database model and FDB model. Then, this paper introduces the application of COMS in statistical analysis、voice operation demonstrator and phonetic analysis. Finally, this paper gives a conclusion about what we have done on this field.

Key words : Mongolian; corpus; structure; application

收稿日期: 2004-5-13;

基金项目: 教育部人文社会科学研究重大项目“蒙古文信息处理平台(MIPP)的研究”(02JAZJD850003); 国家自然科学基金项目“基于WEB的蒙文图书信息管理系统”(60163003)

作者简介: 1, 马丽(1978-), 女, 汉族, 内蒙古鄂尔多斯市人. 内蒙古大学计算机学院研究生, 主要研究信息处理和数据库技术;

2, 宋建斌(1977-), 男, 汉族, 内蒙古呼和浩特市人. 内蒙古大学计算机学院研究生, 主要研究蒙文信息

处理和计算语言学；

3，敖其尔（1941-），男，蒙古族，内蒙古兴安盟人。内蒙古大学计算机学院教授，内蒙古大学蒙古学研究中心兼职教授，硕士生导师，主要研究蒙文信息处理和计算语言学。