《蒙古语语法信息词典量词分库》的建设*

海银花 那顺乌日图

内蒙古大学蒙古学学院 呼和浩特 010021

E-mail: haiyinhua@imnu.edu.cn

提 要：《蒙古语语法信息词典》是为实现蒙古语语句的自动分析与自动生成而研制的、囊括词法形态、句法功能、搭配特征以及正字法规则等信息的一部机器词典。它由《蒙古语语法信息词典总库》（以下简称《总库》）和各个分库等不同层次构成，《蒙古语语法信息词典量词分库》（以下简称《量词分库》）是其第二层的有机组成部分。文中详细介绍了有关《量词分库》的建设情况，包括理论基础，属性字段以及属性值等，分析了《量词分库》建设中的一些难点，并且提出了相应的处理方法。

关键词：《蒙古语语法信息词典》，《量词分库》，属性字段，难点，处理方法

The Development of "The Classifiers Bank Of The Mongolian Grammatical information Dictionary"

Hai.Yinhua Nasun-urt

Academy of Mongolian studies, Inner Mongolia University, Huhhot 010021

E-mail: haiyinhua@imnu.edu.cn

Abstract: The "Mongolian Grammatical Information Dictionary" is a machine dictionary, which is developed for automatic analysis and automatic generation of the Mongolian language. The dictionary includes the information about morphological forms, syntactic functions, collocation features and orthographical rules. It is consisting of "Mongolian Grammatical Information Dictionary general bank"(for short "general bank") and the sub-banks of various word classes. Every sub-bank is organic part of "Mongolian Grammatical Information Dictionary", and "the Classifiers Bank of the Mongolian Grammatical Information Dictionary"（for short "Classifiers Bank"）also be included. The paper will particularly introduces about "Classifiers Bank" and mainly discusses some difficulties.

Key words: Mongolian Grammatical Information Dictionary, Classifiers bank, attribute field, difficulty, measure

一、《蒙古语语法信息词典》概要介绍

蒙古文信息处理工作历时 20 余年所取得的研究成果之一为《蒙古语语法信息词典》。它的研制工作始于 2000 年，并且本项研究先后得到教育部、国家语委、国家自然科学基金项目的支持，目前《蒙古语语法信息词典》基本成形。

《蒙古语语法信息词典》是实现自动分析与自动生成蒙古语语句而研制的一部机器词典。根据收词一贯原则，《蒙古语语法信息词典》收录了约 4 万个词语；根据蒙古语词类划分原则，制定了面向信息处理的蒙古语标记集，完成了这 4 万个词语的归类；采用关系数据库文件格式描述

* 此项研究得到国家自然科学基金（项目号：60363005），教育部、国家语委民族语言文字规范标准建设与信息化项目（项目号 MZ115-005）资助。

作者简介：海银花（1981—），女，内蒙古大学蒙古学学院博士研究生，研究方向为蒙古文信息处理；那顺乌日图（1959—），男，内蒙古大学蒙古学学院教授，博士生导师，主要研究方向为蒙古文信息处理。

每个词语及其语法属性的二维关系。词典中共有 20 个数据库文件，其中包含全部词条的总库 1 个，各类词库 19 个。总库设 19 个共同的属性字段，各类词库又分设若干属性字段，如名词分库设 36 个属性字段，动词分库设 33 个属性字段等分别描述其更翔实的语法属性。词典所有的库均可以根据主关键字段“编号”（NO）[1]、“词语”（MONGOL）、“标音”（GALIG）、“词类”（UGSAIMAG）等相互连接。我们如果定义词典每个库所包含的词条数同该库的属性字段数的乘积为该库的信息量，那么《总库》的信息量约达 76 万，各个库的信息量之和约达到 200 万。并且我们正在研制具有添加、修改、删除、查询、统计和浏览等功能的词典管理工具《蒙古语语法信息词典管理平台》。目前该词典已被应用到语料库标注、语料库对齐、文本转换、文字识别、机器翻译等领域，例如《面向政府文献的汉蒙机器辅助翻译系统（达日罕系统）》、《多字体印刷蒙古文（混排汉英）文档识别系统》等系统中，作为蒙古语语言基础资源，该词典发挥了重要作用，而且在国内、国外已有一定的用户。

如同上述，词典的框架由不同层次构成。其第一层是包括该词典所有词条的《总库》，第二层是各类词的分库，《量词分库》是其第二层的有机组成部分。在自动分析、自动生成、机器翻译、自动标注、自动校对等信息处理工作中该词典所起的作用是通过语法属性字段及其取值所含信息得以实现。因此我们建设《量词分库》时充分利用《总库》、《100 万词级现代蒙古语语料库》和各种蒙古语词典以及蒙古语语法著作等知识资源，设置了易于计算机处理的属性字段及其取值。

二、理论基础

丰富的量词是蒙古语词汇的重要组成部分，是蒙古语词汇的一大特点。量词在蒙古语名词的数量表达式里是必不可少的，表达名词的数量时，必须选用与之相配的量词。例如，“NIGE JUSUM HVRVD”[2] 一条奶豆腐 、“HEDUN JAH_A HVBCASV”（几件衣服）、“JIRGVGAN TOHOI BUS”（六尺布）分别用“JUSUM”、“JAH_A”、“TOHOI”来修饰“HVRVD”、“HVBCASV”、“BUS”等名词而具有极强的修辞功能。但是蒙古语量词的研究和规范还存在一些问题，包括对量词的名称、量词范围的确定、量词的再分类等。这些传统蒙古语法学界中研究不够透切或未进行研究的空白点对于《蒙古语语法信息词典》的研制会带来理论障碍甚至实践难点。研发语法信息词典都要涉足到语言基础理论和应用技术等两项研究工作。例如，中文信息处理要开发《现代汉语语法信息词典》时将朱德熙先生创立的“词组本位语法体系”作为理论指导[3]，保证了词典具有丰富的语法信息和简洁的表述方式，从而保证了词典的质量。从蒙古语的基础理论来讲，我们研制和开发此词典时直接利用或借鉴的现代蒙古语计算语法学体系尚未形成的前提下，通过全面分析或考察传统语法学著作和前人研究成果，从中吸取营养是我们目前采取的最佳方法。

对于量词的定义及其词法、句法等语法特征学术界向来保持一致。首先，对于量词的定义上历来学者们的观点基本保持一致，将其定义为“表示事物和行为的计量单位的词叫做量词”，其次，对于量词的（1）词法方面具有不完全的名词格变化，有领属范畴变化；（2）句法方面以数量形式与数词共同充当句子成分等语法特征上学者们无分歧。

---

[1] 括号里的拉丁转写蒙古文表示《蒙古语语法信息词典》的属性字段标记。

2 文中全部蒙古文以拉丁转写蒙古文形式提供。

3 详见张：“面向自然语言处理的现代汉语词组本位语法体系”，《语言文字应用》1997 年第 4 期。

对于量词的名称、词类划分和再分类问题上语法学界保持不一致。

1．量词名称方面，不同的语言学家对之有不同的称呼，运用不同的名词术语。譬如，内蒙古大学蒙古语文研究所编的《现代蒙古语》（1964 年）称谓“HEMJIGUR-UN NER_E”；布和吉日嘎拉编的《蒙语语法》（1977 年）称谓“HEMJIGUR-UN UGE”；清格尔泰著《现代蒙古语语法》（修订版）（1999 年）中称谓“HEMJIYEN-U NER_E”；甚至有些论著中与数词统称为“TOG_A HEMJIGUR-UN UGE”等，这些名称虽然存在一些细微的差别，但都是从其功能方面命名的名称。对此我们根据与蒙古语其他词类例如“JINGHINI NER_E”（名词）、“TEMDEG NER_E”（形容词）、“CAG ORON-V NER_E”（时位词）等名称的一致性以及离它最近的蒙古语“TOGAN-V NER_E”（数词）的名称统一起来最终采用“HEMJIGUR-UN NER_E”这一术语比较科学。

2．蒙古语量词的词类划分方面存在三种状况：

（1）将量词归类为数词：大多数蒙古语语法著作将量词归类为数词，阐述数词的语法特征的同时随其概述量词的有关内容，这种归类法在蒙古语语法学界占据主导地位；

（2）将量词归类为数量词：量词并列与数词归属于数量词，例如清格尔泰著《蒙古语语法》（1991 年）把蒙古语词类分为静词类、动词类和无变化词类三大类，继而把静词类分为名词、形容词、数量词、时位词和代词，其中又将数量词分为数词和量词两类，并且分别说明了其语法特征和子类划分等内容。布和吉日嘎拉编的《蒙语语法》（1977 年）的量词词类划分也属此类；

（3）未有专用量词的观点：纳·格日勒图著《蒙古书面语语法研究》（1998 年）中阐述了蒙古语没有专用量词，以计量工具或可计量的事物、现象名称作为计量单位的观点。但文中之前已说明数词能够支配量词以短语形式共同充当句子成分的语法特征，这是量词词类划分问题上具有独特见解的唯一著作。

虽说蒙古语量词的词类划分方面学者们多有分歧，但是面向信息处理的蒙古语研究中将其视为独立的词类并制定了其相关标记。

3．语法学界从不同的角度，按不同的标准对于量词进行了子类划分。例如（下面以列表形式提供各种蒙古语语法著作中的量词的再分类情况）：

著作名称	量词的再分类
内蒙古大学蒙古语语文研究所编的《现代蒙古语》（1964 年）	物量词和动量词
松日布编的《蒙古语语法知识》（1976 年）	名量词和动量词
布和吉日嘎拉、恩和编的《蒙语语法》（1977 年）	表示容量、重量和时间的量词，表示种类、钱单位的量词，计量事物的量词等 5 种
那森柏、哈斯额尔顿等编的《现代蒙古语》（1982 年）	长度、容量、重量等的专有词语，以名词为替代的物量词，动量词，时量词等 4 种
达瓦编的《现代蒙古语基础知识》（1982 年）	人、物量词，动作状态量词，时量词等 3 种
涛高、援朝等人编的《现代蒙古语》（1993 年）	长度、容量、重量的量词，以名词为替代的物量词，动量词，时量词等 4 种
哈斯额尔顿、贡其格苏荣编的《现代蒙古语》（1996 年）	长度、容量、重量词，以名词为替代的物量词，动量词，时量词等 4 种
清格尔泰著《蒙古语语法》（1991 年）	物量词，动量词，时间量词等 3 种
清格尔泰著《现代蒙古语语法》（修订版）（1999 年）	度量词、动量词、时量词
嘎日迪等人编的全国高等学校教材“现代蒙古语”（2001 年）	距离、长度量词，重量词，容量词，个体集合量词，动量词，钱单位量词等 6 种。

对于量词的这些内部子类划分有的比较粗浅，有的相对详细，但其共同点在于均属按照量词

的运用状况和语义作出的分类结果。除此之外，布和吉日嘎拉、恩和编的《蒙语语法》(1977年)和达瓦编的《现代蒙古语基础知识》(1982)把量词分为专用量词和借用量词两大类，这是从蒙古语量词的来源角度对其进行的分类法。

4. 量词范围涉足到量词来源问题。蒙古语量词与名词关系密切，量词一部分是由名词转化而来，如，“SURUG、JIL、GALBA”等，也有的从动词转化而来如“ADHV、CIMHI、OGOCI”等，而且现实生活中大量存在着名词临时借用为量词的现象，并且这些量词大部分都以单词形式存在，这是扩充蒙古语量词范围的直接原因。另一方面随着蒙古族悠久历史文化发展和漫长游牧生产生活的需求而产生了一些复合量词，它们均属由两个或两个以上的词而组成的词组结构，例如“MVHVR SOGOM、SOGOM GAJAR、EREHEI ORIYAM_A、HOOS ALDA、JVLVG SIGUREM_E、UJUGUR TOHOI、BILAGV SIDAM_A、HUJUGUU CINEGE”等与某一个事物进行比较估量的复合量词。这些复合量词虽然由两个或两个以上的词而组成，但在人的思维里其表达的意义是估量距离或时间的固定的一种概念，因而我们将它们视为蒙古语词汇的组成部分。蒙古族的这种丈量的习俗以及从而产生的量词对于蒙古族语言思维的丰富发展有着深远影响，例如，“UHER-UN CINEGE”、“HONIN CINEGE”等词语分别比喻计量巨大实物和植物成长过程，随之产生“UHER CILAGV”、“HONIN TARIY_A”等专用固定名词词组[4]。

三、建设《量词分库》中的难点及相应的处理方法

1.《量词分库》收词问题

（1）《总库》是各分库词条的主要来源。《总库》里总共收录了144个量词，我们建设《量词分库》时将其作为基本词条来收录。另一方面，根据笔者对于蒙古语量词的收集和统计，传统蒙古语语法著作中“ALHVM、ALCAM、BADAG、JVRBVS、SIRHEG、JAH_A、EDUR、GAJAR、SAGVRI、TOLOGAI、BULUG、HESEG、TOGORIG、MONGGO、SVMV、JUIL、TASVLG_A、MOR、DEBTER、OBOG_A、HERCIM、GARAG、DVSVL、HAGVDASV、HAYIRCAG”等词被看作量词。但是，这些词作为名词收录在《总库》，因而未被收录在《量词分库》中。这些词作为兼类词，名词兼类量词是其语法特征之一，从而将它们视为名词收录在《名词分库》，并且《名词分库》里专门设定“Q”属性字段来描述能否充当量词这一语法信息。

（2）《蒙古语语法信息词典》主要以单词为基本单位来收录词语，因此《量词分库》中未能涵盖蒙古语复合量词。那些“HOOS ALDA、EREHEI ORIYAM_A”等富有民族特色的复合量词也是蒙古语词汇的重要组成部分，我们将它们收录在《蒙古语语法信息词典》的组成部分—《蒙古语固定短语语法信息词典》的《习用语分库》(X)中。

2. 量词再分类问题

如同上述，在传统蒙古语研究中对于量词子类划分问题上众说纷纭，从其应用意义、来源等不同的角度，按照不同的标准提出了各种分类法。

（1）我们本着“表示事物和行为的计量单位的词叫做量词”这一描述量词的语义性定义，并兼顾量词自身的词法特征、句法功能，在制定“面向信息处理的蒙古语标记集”时将量词(Q)分为名量词(Qn)、时量词(Qc)、动量词(Qv)等三种。在《量词分库》的“UGSAIMAG”属性字

[4] 俞士汶：“面向自然语言处理的现代汉语词组本位语法体系”，《语言文字应用》1997年第4期

段的属性值为这三种量词的相关标记。其中名量词指的是事物的计量单位，该分库中名量词占绝对多数，将近达占到90.3%，包含了表示重量、容量、度量、个体、集合、过程、种类等量词；时量词和动量词分别指时间和动作的计量单位，分别占8.3%和1.4%，这里仅收录了“VDAG_A”和“DAHIN”两个动量词。面向信息处理的蒙古语研究中这种综合性的分类法比较粗略，未能满足自然语言计算机处理的更高需求，针对这种要求，我们在《量词分库》中专门设置了“VDH_A”属性字段，填写每一个量词能够计量的意义。例如，“KILOgRAM”、“ALDA”、“BITEGUU”的“VDH_A”字段里分别填写“HUNDU”、“VRTV”、“BAGTAGAMJI”等，各自表示重量词、度量词、容量词等信息。

（2）蒙古语量词也和其他民族语言一样，有固定量词和临时量词之分。根据《“蒙古语语法信息词典”框架设计》的基本思路，像“ALDA、DELIM、SOGOM、TOGE、IMAHV”等一部分蒙古语固有词和“JING、LANG、KILOgRAM、LItR”等汉语借用词以及国际通用量词属于固定量词，“ADHV、OGOCI、JAGVN、HONOG”等兼类量词则属于临时量词。根据这一特征，我们在该分库里设置了“VVGAL”属性字段，表示是否属于固定量词。这里我们除了蒙古语固有的量词之外、从名词、动词转换而来的和一些外来语一律归属了临时量词。

3．量词与数词搭配问题

从语法意义上讲，蒙古语量词表示计量的单位或等级、编号单位。它本身并不包含数量的意义，只有与数词结合后数量词组整体才能表示数量。因此量词受数词和数词词组的修饰作定语是其最主要的功能。可是蒙古语量词并不与所有数词能够搭配，与量词构成数量词词组的主要是基数词、 概数词、 分配数词、分数词和序数词。其中蒙古语全部量词与基数词、概数词、分配数词搭配组成数量结构，通过分库中的属性字段未能显现量词的这种共有的语法特征，从而未能区分量词相互间的语法、语义差异；与分数词和序数词搭配时却具有选择性，例如，“DAHIN”一词可以与“GVRBAN-V NIGE”等分数词搭配，不可以与“HOYADVGAR”等序数词搭配，但“VDAG_A”却可以与“HOYADVGAR”搭配。针对这一语法特征，我们在此分库里设置了“前面能否与分数词搭配”（属性字段标记为“Fm5+Q”）和“前面能否与序数词搭配”（属性字段标记为“Fm2+Q”）两个字段，分别表述了此项语法属性。

四、《量词分库》属性字段介绍

下面以列表形式显示《蒙古语语法信息词典量词分库》中设置的15个属性字段及其取值。

属性字段	属性值及其说明
NO	每个词条的词典中的编号
“MONGOL”	填写每个词条的传统蒙古文
“GALIG”	填写拉丁转写蒙古文，例如，“VDAG_A”中填写“VDAG_A”
“UGSAIMAG”	表示词类，填写词语的所属词类标记，例如，“DAHIN”中填写“Qv”
“VVGAL”	是否属于专用量词，属性值为逻辑值，是填写“Yes”，不是填写“No”，例如，“DELIM”中填写“Yes”，“BITEGUU”中填写“No”
“Q-G”	能否与后置词或具有后置词功能的词搭配，属性值为逻辑值，能，填写“Yes”，不能，填写“No”，例如，“ADHV”中填写“Yes”，“JUIREI”中填写“No”。

“Q+N”	后面能否与名词直接搭配，属性值为逻辑值，能，填写“Yes”，不能，填写“No”，例如，“LANG”中填写“Yes”，“BER_E”中填写“No”。一个量词后往往可以与多个不同的名词直接搭配，例如，“NIGE ADHV AMV”、“DOLOGAN NIVtON HUCU”等，蒙古语名量词具备这一特点区别于动量词和时间量词。而动量词、时间量词与名词搭配时一般通过一些蒙古语语法形式变化，例如，“NIGE JIL-UN TOLOBLEGE”、“HEDU VDAGAN_V YABVDAL”，并且动量词、时间量词与动词搭配的较多，例如，“NAYIMAN SAR_E AJILLAHV”、“HOYAR VDAG_A VNGSIL_A”等等。
“Fm5+Q”	前面能否与序数词搭配，属性值为逻辑值，能填写“Yes”，不能，填写“No”，例如，“GALBA”中填写“Yes”，“DAHIN”中填写“No”。蒙古语时间量词和表示团、堆、集合的量词具备了此语法特征
“Fm2+Q”	前面能否与分配数词搭配：其属性值为逻辑值，能，填写“Yes”，不能，填写“No”，例如，“TON”中填写“Yes”，“JAGVN”中填写“No”。
“Q-OYIRCG”	有无同义量词，其属性值为填写其同义量词，有，填写相关同义量词，无，不填。例如，“BADVN”中填写“dAN”，“BAGCA”中不填。该分库中具有此语法特征的量词不占多数，有“EBHEGE(EBHEGESU)、PUU(PUUD)、JIRAN(RABJVNG)、UNc(UNS)、UR_E (MU)、BADVN(dAN)、BARIM_A (BARIM)”词语等
“Q-ONDOO”	有无词义相同、词形不同的量词，属性值为填写相关量词，有填写相关量词，没有，不填，例如，“MILI”中填写“MILLI ”，“JING”中不填。该分库中“MILIgRAM”(MILLIgRAM)、“MILILItR”(MILLILItR)、MILIMetR (MILLIMetR)、SeKUInd(SeKUNd)等量词具备了此项语法特征；
“QQ”	能否重叠使用，其属性值为逻辑值，能，填写“Yes”，不能，填写“No”，例如，“OGOCI”中填写“Yes”，“KALVRI”中填写“No”。有些专用量词以重叠形式作定语或状语，例如，“OGOCI OGOCI VSV”
“Q-HORSY”	有无搭配使用的量词，其属性值为填写其搭配的词语，有，填写相关搭配量词，无，不填。例如，“ALDA”中填写“DALIM”，“LItR”中不填。
“Q-VDH_A”	填写所表示的语义，属性值为填写所表示的意义的拉丁转写蒙古文，例如“MIL”中填写“VRTV”，表示长度量词；“MILIGRAM”中填写“HUNDU ”，表示重量词
“Q-TOD”	能否独立充当定语，其属性值为逻辑值，能填写“Yes”，不能，填写“No”，例如，“ADHV”中填写“Yes”，“DAHIN”中填写“No”。在句子中量词一般以数量短语形式充当定语，但有些量词却独立充当定语有别于其他量词。例如“ADHV BVDG_A CV UGEI”

小结

本文旨在对蒙古语量词相关研究进行查询、整理并归纳其语法特征的基础上初步制定了《量词分库》属性字段标记及其属性值。为了保证词典的质量我们试着制定合理、规范的属性字段标记，填写属性字段时虽然依据语法著作进行查询，并兼顾真实文本中的实际使用情况，但受目前对蒙古语量词研究基础的薄弱，以及作者水平所限，难免有遗漏，在今后的研究工作中遇到实际问题时，我们将及时对“量词分库”的内容做相应的调整，不断地完善该分库是我们长远任务之一。

参 考 文 献

[1] 《现代蒙古语》，内蒙古大学蒙古学学院蒙古语文研究所编，内蒙古人民出版社 1964 年；

[2] 《蒙古语语法知识》，松日布编，黑龙江人民出版社 1976 年；

[3] 《蒙语语法》，布和吉日嘎拉、恩和编，内蒙古人民出版社 1977 年；

[4] 《现代蒙古语》，那森柏、哈斯额尔顿等编，内蒙古教育出版社 1982 年；

[5] 《现代蒙古语基础知识》，达瓦编，内蒙古少年儿童出版社 1982 年；

[6] 《蒙古语语法研究》，确精扎布著，内蒙古大学出版社 1989 年；

[7] 《现代蒙古语》，涛高、援朝等编，内蒙古少年儿童出版社 1993 年；

[8] 《现代蒙古语》，哈斯额尔顿、贡其格苏荣编，内蒙古教育出版社 1996 年；

[9] 《蒙古语语法》，清格尔泰著，内蒙古人民出版社 1991 年；

[10] 《蒙古书面语语法研究》，纳·格日勒图著，内蒙古教育出版社 1998 年

[11] 《现代蒙古语语法》，清格尔泰著（修订版），内蒙古人民出版社 1999 年；

[12] 全国高等学校教材《现代蒙古语》，嘎日迪主编，内蒙古教育出版社 2001 年；

[13] “蒙古语量词—蒙古人的丈量习俗”，那顺乌日图《蒙古语言文学》1991 年第 5 期；

[14] “蒙古语语法信息词典”框架设计，那顺乌日图，内蒙古大学，博士学位论文，2000 年；

[15] 《蒙古语语法信息词典》管理平台的设计与实现，王斯日古楞，《应用语言学学术研讨会》，内蒙古大学，2007 年；

[16] “蒙古文识别文本后处理相关技术研究”，包艳花，内蒙古大学，硕士学位论文，2007年；

[17] 《现代汉语语法信息词典详解》（第2版），俞士汶等著，清华大学出版社，2002年；

[18] “综合语言知识库的建设与利用”，俞士汶等，《中文信息学报》2004年第5期；

[19] 《蒙古语辞典》，《蒙古语辞典》编纂组，内蒙古人民出版社 1997 年；

[20] 《蒙汉词典》（修订版），内蒙古大学蒙古学学院蒙古语文研究所编，内蒙古大学出版社 1999 年；