《蒙古语语义信息词典》的初步构建*

德·萨日娜[1]　那顺乌日图[2]

[1]内蒙古社会科学院蒙古语言文字研究所 呼和浩特 010010 [2]内蒙古大学蒙古学学院 呼和浩特 010021

E-mail: saran352@yahoo.com.cn　qingjirvm@126.com

摘要:《蒙古语语义信息词典》是面向信息处理的语义知识库，词典用现代蒙古语的“词语”为基本单位来组织记录它们在语言运用中的多种语义信息。初步构建的词典内容由语义属性库和语义分类库组成，语义属性库又包括总库和事物类分库、运动类分库和性状类分库等若干分库。各库根据其主要功能对所收录的词目或词汇从不同的角度进行语义描述，所记录的语义信息要满足句法和句义处理中所需的语义知识。

关键词: 现代蒙古语，语义信息，词典

The Preliminary Construction of "Mongolian Semantic Information Dictionary"

De•Sarina[1]　Nasun-urtu[2]

1 Institute of Mongolian Language of Inner Mongolia Academy, Hohhot, 010010

2 The Mongolian College of Inner Mongolian University, Hohhot, 010021

E-mail:saran352@yahoo.com.cn　qingjirvm@126.com

Abstract: "Semantic Information Dictionary" is the semantic knowledge base for information processing. Taking the modern Mongolian "words" as basic unit, this dictionary organizes and records their varieties of semantic information in language use. The contents of the preliminary construction of the dictionary consist of semantic properties and semantic classification base, the semantic properties base also includes general base and a number of sub-bases such as object-type sub-base, sports-type sub-base and character-type sub-base. According to the primary function, each base describes the semantic information of recorded words and vocabulary from a different point of view. The recorded semantics information should meet the requirement of semantic knowledge needed in the syntax and sentence meaning processing.

Key words: Modern Mongolian　semantic information　dictionary

1 引言

近年来蒙古文信息处理研究取得进一步的发展，随之该领域对语义知识的需求越来越紧迫。这种需求涉及到语言单位的义素、语素义、词义及句义等多个层面。分析实际需求，结合蒙古语语义研究的进展情况，建立一个综合性的蒙古语语义知识库是解决语义问题的根本途径。《蒙古语语义信息词典》作为这种语义知识库用现代蒙古语的“词语”为基本单位来组织记录它们在语

* 该研究受国家自然科学基金项目（项目号：60873084）的资助。

言运用中的多种语义信息。建立该词典的目的一要满足蒙古文信息处理中句法分析、句义分析、相似度计算等所需求的语义知识；二要为日常语言工作和语言学习提供词义知识；三要为语言理论研究提供多方面的、具有实际意义的语义依据。

该词典将收录《蒙古语语法信息词典》的全部常用单词（共 38000 多个）和附加成分，在此基础上还要追加部分常用复合词来构成词典的主要词目或词汇，所谓“词汇”指的是纳入在语义分类系统中的词汇。词典内容由若干库组成，各库所收录的词目、词汇保持一致性。各库根据其主要功能对所收录的词目或词汇从不同的角度进行语义描述，所描述的语义知识包括词语的义项类、语义类、配价类、角色搭配类、近义类、反义类及释义等多方面。

2 词典建立的基础

蒙古文信息处理研究自上世纪 80 年代开创以来，在应用系统研究和基础理论研究方面取得了较好的成果。如，“方正电子出版系统蒙文版”、“华光轻印刷系统蒙古文版”等排版系统、“蒙古文 WPS Office”办公软件的诞生和应用，以及蒙古文机器翻译、蒙古文自动校对系统的研究等成为该领域研究良好的开端。在基础建设方面，蒙古语语料库建设与加工、词概率统计、《蒙古语语法信息词典》的研制、术语数据库建设等工作也取得一些成果并积累了经验。近几年在应用研究和基础建设的带动下面向信息处理的蒙古文理论研究受到研究人员的重视，研究内容涉及到蒙古语词法属性、句法、语义、文本识别等方面。这些研究从各个方面反映出建立蒙古语语义知识库的必要性。同时，这些研究成果也正成为语义知识库建设的重要基础和理论依据。

《蒙古语语义信息词典》的建立旨在将词语的语义分类、词义内部结构分析及其在句法、句义中的角色功能等各知识点有条不紊地组织在若干库中，在研究中有效地管理和利用。从词典建立的理论方面来讲，词语的语义知识来源有多方面，其中包括蒙古语词典学中词汇、词义分析研究成果；词汇学、语义学的相关研究成果；蒙古语计算语义研究中的语义分类、总语义场的设计、词义搭配研究、配价理论及格框架研究等诸多成果。

从蒙古文信息处理的现状来看，研究工作面临着句处理，在某些领域尝试着句结构分析生成、短语分析识别、句切分、语义识别等研究。蒙古语是黏着型语言，它的句结构相对自由多样，词法变化极为丰富。蒙古语句处理究竟立足于哪一种理论框架，具体采取什么方法比较可行，这是我们必须面对并解决的问题。现在蒙古语句处理研究中的多半遵循短语结构语法理论，主要凭借短语结构形式分析来完成句分析或生成。很显然在这类研究中必然产生一些词法、词义层面的歧义现象，如果没有相关语义知识的支持，歧义难以得到解决。此外，我们在语料库研究中也尝试着引用格语法理论进行句义分析识别的研究，由此认为在句处理研究中根据蒙古语本体特征，充分利用其词语形态变化等属性进行句义分析也有望得到较好的效应。

3 词典的主要内容

从上述研究中所需的语义知识点来看，《蒙古语语义信息词典》中涵盖的语义知识既要满足句法处理中所需的词语义类、配价量、配价质等知识，又要满足句义处理中所需的角色搭配类、义项分析等知识。当然有些语义知识有可能在句法、句义处理中同样有价值，甚至在其他领域的语言研究、语言运用中也有所价值。如，义项、释义、近义词、反义词等等。

《蒙古语语义信息词典》初步建立在两大语义分库之上，即语义属性库和语义分类库。

3.1 语义属性库

《蒙古语语义信息词典》中的语义属性库包括 1 个总库和 6 个分库。这些库所建立的目的和功能有所不同。总库主要根据文本处理研究中句法分析生成所需的语义知识而建立，而各个分库主要根据句义分析生成所需的语义知识而建立。

总库收录上述常用词语为词目，词目以义项为单位，即单义词语的一个词语为一个词目，而多义词语的一个义项为一个词目。总库的词目信息中要设置“*序号*、*词目*、*读音*、*词类*、*同形序列*、*义项序列*、*义项范围*、*义类*、*主义素*、*配价义类*、*配价量*、*配价质*、*释义（例）*、*汉译*、*代价*(相当于频率)”等 15 个字段。

这些字段中的“词目、读音、词类、同形序列”等信息要与《蒙古语语法信息词典》中的同类信息基本相符。而“义类、主义素、配价义类、配价量、配价质”等信息的来源主要参考或根据面向信息处理的蒙古语语义研究相关成果。而“义项序列、义项范围、主义素、释义（例）”等信息的依据有待于进一步详细研究，这些问题基本解决之后词典的所有字段信息才能够得以充实。例如：在总库中以常用动词“BARI”的第一个义项为词目，那么它的主要语义信息内容为：“义项序列→1；义项范围→物；义类→ Vaj4；主义素→ ADHV；配价义类→ Nbw2；配价量→ 2；配价质→ Ud,Ht；释义→ YAGVM_A-YI GAR-TAGAN ADHVJV ABHV；汉译→抓、拿；代价→”。

语义属性库中的 6 个分库分别是（1）事物类分库；（2）运动类分库；（3）性状类分库；（4）句型成分分库；（5）附加成分分库；（6）词缀分库。下面简单阐述这 6 个分库各自的选取词目、描述内容及功能。

（1）**事物类分库**：词目由总库中的所有名词和代词（除了形容词性的、数词性的之外）词目构成。根据蒙古语事物类词语格框架中多充当“主体、客体、邻体”语义格，而很少充当谓语中心词和谓语的修饰成分等特点，对事物类分库中的词目设置了“*序号*、*同形序列*、*词目*、*读音*、*义项序列*、*词类*、*词连接成分*、*主体格 1*、*主体格标 1*、*主体格 2*、*主体格标 2*、…、*客体格 1*、*客体格标 1*、*客体格 2*、*客体格标 2*…、*邻体格 1*、*邻体格标 1*、*邻体格 2*、*邻体格标 2*、…、*近义词*、*反义词*、*复合词例*、*俗语例*、*惯用型例*”等字段。

这些字段中的 “词目、读音、同形序列、义项序列、词类” 信息要与语义属性总库中的同类信息相符。“词连接成分”指词语在格结构中与其他词语或短语搭配时所要的连接形式，即附加成分或虚词。“主体格、客体格、邻体格”分别指格框架中不同的语义成分，其后的“1,2,3…”序列号分别指大语义格下层的小语义格，如，“主体格”为大格，而“实施、当事、主题”为其下层小格。这一属性反映词目在句中能否充当该“格”中心词的功能，如果能则填其相应代码，否则就不填。“格标”属性指词目充当相应“格”时可能要的“格标志”种类。另外字段中的“近义词”、“反义词”的概念主要参考蒙古语词汇学、词义学研究中合理解释来定义。例如：在事物类分库中用常用名词“CAI（茶）”的第三个义项作为词目，那么它的主要语义信息内容为：“词连接成分→ -YIN|-TAI|BA| BOLON|HIGED|；主体格 1→；主体格标 1→；主体格 2→{uij}；主体格标 2→ ∅ |NI|MINI|CINI；主体格 3→{sed}；主体格标 3→ ∅ |BOL|GE/DEG BOL|GE/GCI|·|NI| MINI |CINI；客体格 1→{hur}；客体格标 1→ -YI|-BAN| ∅；客体格 2→{θrt}；…；客体格标 2→ ∅；…；近义词→；反义词→；复合词例→ HAR_A CAI；HOHE CAI …；俗语例→ CAI CV UGEI，CIRAI CV UGEI…；惯用型例→”。

（2）**运动类分库**：词目由总库中的所有动词词目构成。蒙古语的动词在格框架中主要充当谓语中心词，它能够支配句中各语义格的数量和种类。动词在单句中所支配的语义格的数量是有限的，它一般不超过个数。各语义格在句中所扮演的角色有主次之分，并且这种主次之分的规律很大程度上取决于充当谓语中心词词语的义类。我们在格框架中根据语义格角色的这种主次之分又把它们区分为“必选格”和“可选格”两种。我们将“必选格”和“可选格”的概念作为动词句义结构中的主要特征纳入到运动类分库中。在运动类分库中设置的字段有“*序号、同形序列、词目、读音、义项序列、词类、助动词、状动词、必选格 1、必选格标 1、必选格 2、必选格标 2、必选格 3、必选格标 3…、可选格 1、可选格标 1、可选格 2、可选格标 2、可选格 3、可选格标 3…、近义词、反义词、复合词例、俗语例、惯用型例*”等。

该分库字段信息的填置方法与事物类分库的大致相同。但是在“必选格、可选格”字段中填置的内容应当是词目词语句中能够支配的“格”种类，“必选格标、可选格标”中填置的内容也是所支配相应“格”可能要的“格标”种类。例如：在运动类库中以常用动词“AB（要、索要、索取）”的第一个义项作为词目，那么它的主要语义信息内容为：“义项序列→1；助动词→ ‘Y’；状动词→ ‘Y’；必选格 1→{uid}；必选格标 1→∅|·|NI|MINI|CINI| BOL|-DV -TV|；必选格 2→{hur}；必选格标 2→-YI -I|-BAN -IYAN|∅；…；可选格 1→{oro}；可选格标 1→-ECE；可选格 2→{toh}；可选格标 2→∅|VDAG_A；…；近义词→ OL, ABCIRA；反义词→ OG；复合词例→ AB/HV TANVG UGEI…；俗语例→AB/HV HOMON GEDEYIDEG, OG/HU HOMON BUHUYIDEG…；惯用型例→ AB/V/N DOOR_A-BAN…”。

（3）**性状类分库**：词目由总库词目中的形容词、副词、数词、量词、代词（形容词性的、数词性的）、摹拟词、感情词词目构成。蒙古语的这些词语在格框架中的主要功能是充当谓语的修饰成分，我们在格框架研究中把它也作为一个“语义格”来看待，统称为“修饰体”。修饰体格的小类有“性状、数量、样态、频度”等几种，它们无论什么条件下都不能充当必选格。但其中的“形容词、数词”当名词用时也可能充当“主体”或“客体”等必选格。该分库中的有些词类，如“形容词、数词、模拟词”等少数情况下能够充当谓语中心词。综合考虑这些情况，我们将性状类分库的属性字段设置为“*序号、同形序列、词目、读音、义项序列、词类、修饰名词、修饰动词、修饰形容词、充当谓语中心词、充当语义格、可选格 1、可选格标 1、可选格 2、可选格标 2…、必选格 1、必选格标 1、必选格 2、必选格标 2…、近义词、反义词、复合词例、俗语例、惯用型例*”等。例如：在性状类分库中以常用形容词“VLAGAN（红）”的第一个义项为词目，那么在它的语义信息内容为：“义项序列→1；修饰名词→ ‘Y’；修饰动词→ ‘Y’；修饰形容词→ ‘Y’；充当谓语中心词→ ‘Y’；充当语义格→ ‘Y’；可选格 1→{cin}；可选格标 1→-IYAR|∅；可选格 2→；可选格标 2→；必选格 1→{sed}；必选格标 1→∅|BOL|NI|MINI|CINI|GE/DEG|GE/GCI|-DV|-ACA|；必选格 2→{uid}；必选格标 2→|-IYAR|-DV|；必选格 3→{ hur}；必选格标 3→-YI|-IYAN；近义词→VLABVR, VLABTVR…；反义词→ NOGOGAN；复合词例→ VLAGAN ALAG, VLAGAN ARIHI…；俗语例→ VLAGAN-I UJE/BEL VRBA/JV, HOHE-YI UJE/BEL HURBE…；惯用型例→”。

（4）**句型成分分库**：词目由总库词目中的连词（在整句、分句之间起连接作用的）、连接形式、语气词、助动词、情态词词目及情态复合成分等构成。这些词语或成分在格框架中的作用是连接整句和整句、分句和分句或表示句子的情态义等。该分库的属性字段设置为“*序号、同形序列、词目、读音、词类、谓语动词后、谓语动词前、谓语体词后、谓语体词前、义类*”等。在属性字段“谓语动词后、谓语动词前、谓语体词后、谓语体词前”中填置的内容为词目在这些位

置上出现的可否属性。

（5）**附加成分分库**：词目由部分连词（只在词或短语之间起连接作用的）、后置词、语法附加成分、领属词、领属附加成分和数附加成分构成。这类词语和附加成分在格框架中要么连接词或短语来构成语义格，要么表示语义格的领属义、数义等。尤其它们的多数在句义中能够充当各语义格的标志，这一特征为语义格和格框架的分析起到重要作用。需要说明的是附加成分与格标志之间的关系不是一对一的关系，而是多对多的关系。该分库的属性字段设置为“*序号*、*词目*、*读音*、*词类*、*附加成分类*、*义类*、*词或短语连接*、*主体格标 1*、*主体格标 2*…、*客体格标 1*、*客体格标 2*…、*邻体格标 1*、*邻体格标 2*、…*修饰体格标 1*、*修饰体格标 2*、…”等。在属性字段“词或短语连接”中填置的内容为词目充当的可否属性，而“主体格标 1、主体格标 2……”中添置的内容为词目能够充当其标志的语义格种类。

（6）**词缀分库**：词目由动词的诸多构形词缀构成，其中包括语法书中列举的动词时式、陈述式、态、体、形动词、副动词的各词缀及这些词缀的同词干后叠加使用形式、两个同词干后并列使用形式、两个不同词干后并列使用形式等。这些词缀在句义中表示主要和非主要动作行为的发生时间、陈述语气、主体作用、进行状态等。从这个意义上来讲它们能够构成句义的重要组成部分。另外，有些词缀的单式或叠加使用与格框架中的主体、客体的类别及形式有着密切的关联。该分库的属性字段设置为“*序号*、*同形序列*、*词目*、*读音*、*词缀类*、*义类*、*后现体词*、*后现动词*、*后现助动词*、*后现语气词*”等。在属性字段“后现体词、后现动词…”中填置的内容为在词目后面这些成分出现的可否属性。

3.2 语义分类库

建立蒙古语语义分类库的目的一是为文本处理句法研究中歧义消解提供语义知识，二是为配价研究、格框架研究中的句法、句义结构分析提供相关的义类知识。另外该库中的语义分类系统还为上述语义属性库中的“义类”属性字段提供可靠信息。

建立蒙古语语义分类库要以前期研究成果为基础，利用按类别层次进行标注的语义体系框架，将所选取的词汇逐一填进去。被选的词汇要以义项作为单位。对体词（名词、代词等）主要根据其本身意义分类。例如：“NABCI（叶子）”这一名词属于名词（N）→事物（Nhb）→物（Nbw）→生物（Nbw1）→构件（Nbw13）→植物的（Nbw133），那么它的语义分类代码为“Nbw133”。对述语词（动词、形容词等）主要根据其本身义的同时，还要考虑其与体言词搭配的质量特征来进行分类。例如：“YABV（走、去）”这一动词属于动词（V）→运动动词（Vho）→移动（Vho4），那么它的语义分类代码为“Vho4”（上述二例的分类请见下面的语义分类系统部分例中的斜体字行）。分类要体现词语的纵向继承、横向对比的语义关系，又要适当斟酌语义颗粒度的均衡。

N 名词

Nhb 事物

Nhe 事

Nhe1 政治

Nhe2 经济……

Nbw 物

Nbw1 生物

Nbw11 动物

Nbw111 人类

Nbw112 牲蓄

Nbw1121 兽类

Nbw1122 禽类

Nbw1123 微生物

Nbw12 植物

Nbw13 构件

Nbw131 人的

Nbw132 牲畜的

Nbw133 植物的……

V 动词

Vho 运动动词

Vho1 感知

Vho2 表露

Vho3 行为

Vho4 移动……

4 结语

建立面向信息处理的蒙古语语义知识库能够满足当前蒙古文信息化研究多方面的需要。这项研究无论在应用研究还是理论研究方面都有一定的基础。尽管如此，我们所构建的语义信息词典中的有些主要问题到目前还未能得到很好的解决。如，在词典中词语的主义素和义项范围怎么制定，多义词的义项怎么划分更合理，词语的释义和例句采取什么样的模式才能够在研究中更为实用等问题都亟待解决。并且有些问题也需要进一步的研究和验证，如，词语的语义分类、配价属性及格框架属性等问题。因此，该论文中探讨的只是《蒙古语语义信息词典》初步的构架问题，从整体上来讲，这一构架需要更细致的研究和完善。

参 考 文 献

[1] 那顺乌日图. 蒙古文信息处理. 赤峰：内蒙古科技出版社, 1998：201~264.

[2] 王惠, 詹卫东, 刘群. 现代汉语语义词典的设计与概要//黄昌宁. 中文信息处理国际会议论文集. 北京：清华大学出版社, 1998：361~367.

[3] 吴尉天. 汉语计算语义学. 北京：电子工业出版社, 1999：141~153.

[4] 那顺乌日图.『蒙古语语法信息词典』框架设计. 呼和浩特：博士学位论文, 2000.

[5] 王惠, 李康年. 大型词典编纂的计算机辅助开发与管理系统. 上海《辞书研究》, 2004 , 2：73~81.

[6] 何莲喜. 蒙古文词的多义研究. 呼和浩特：内蒙古人民出版社, 2002 年.

[7] 陈群秀. 一个现代汉语语义知识库的初步实现. http:// www.china-language.gov.cn

[8] 额尔敦朝鲁. 面向信息处理的蒙古语动词语义研究. 呼和浩特：博士学位论文, 2005.

[9] 德·萨日娜. 蒙古语格框架研究. 呼和浩特：博士学位论文, 2006.