蒙古语熟语资源库的初步构建

海银花[1]，那顺乌日图[2]，额尔敦朝鲁[3]

（1.内蒙古大学蒙古学学院，内蒙古呼和浩特市 010021；
2.内蒙古大学蒙古学学院，内蒙古呼和浩特市 010021；
3.内蒙古大学蒙古学学院，内蒙古呼和浩特市 010021；）

摘要：随着信息社会的迅猛发展，蒙古语熟语的语汇和应用面临着巨大挑战。构建“熟语资源库”是保护、开发和利用蒙古语熟语资源的最佳途径，也是机器翻译、语料库加工、文本校对等多个领域提供形式化知识从而能够解决蒙古文信息处理研究的燃眉之急。同时将其研究成果拓展到教学领域，提升蒙古语言文字的教学效率。目前，该资源库处于初步开发阶段。文中从资源库的规模与结构、属性字段及管理软件设计、应用前景分析等方面介绍该资源库的总概貌。

关键词：蒙古语；熟语资源库；初步构建

The Initial Development of the Resource Base of Mongolian Idiom

HaiYinhua[1], Nasun-urt[2],Eerdunchaolu[3]

(1. The Institute of Mongolian Studies, Inner Mongolia University, Hohhot 010021, China;
2. The Institute of Mongolian Studies, Inner Mongolia University, Hohhot 010021, China;
3. The Institute of Mongolian Studies, Inner Mongolia University, Hohhot 010021, China)

Abstract：With the rapid development of the information society, the vocabulary and application of Mongolian Idioms have facing a serious challenge. Building a "Resource Base of Idioms" is not only the best way of protecting, developing and utilizing Mongolian language resources, but also providing formal knowledge for machine translation, corpus processing, text proofreading, and other fields which can solve the immediate needs of Mongolian Information Processing. We can expand the research results to the field of teaching, improving teaching efficiency of Mongolian language. Currently, the resource base is in the preliminary stages of development. This paper will dissertate a total overview of the resource base from the perspective of the scale and structure, attribute field, and the design of management software, the analysis of application prospect and so on.

Key words：Mongolian, Resource Base of Idioms, Initial Development

1 引言

熟语作为蒙古语言资源的一个重要组成部分，源远流长而承载着蒙古族悠久的文化遗产，它能够形象地反映出蒙古族人民的生活习俗、价值取向以及思维方式，可为蒙古族文明多个领域研究提供宝贵的资源。但是，目前蒙古语熟语的发掘、开发和整理进展不尽人意，其数字化研究和形式化描述，亟待我们翔实而深入的研究，使之得到更好的保护和利用。

面向人的、传统的蒙古语熟语研究大多将熟语归入词汇学、句法学或文献学等领域，从其类

基金项目：国家社科基金项目“蒙古语熟语知识库的开发与研究”（12CYY062），国家自然科学基金重点项目“多民族文字识别及理解的理论与方法”（61032008）（与清华合作），国家自然科学基金项目“蒙古语语义信息词典”的设计与实现（项目号 60873084），内蒙古大学项目“蒙古语名词语义知识库的构建”（710067）。

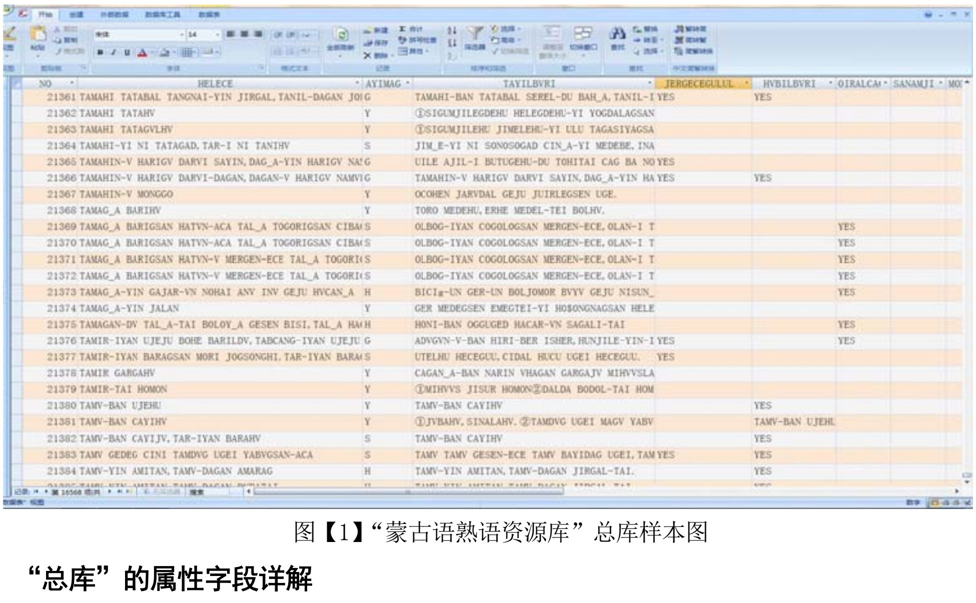
作者简介：海银花（1981-），女（蒙古族），博士，主要研究方向为蒙古文信息处理；那顺乌日图（1959-），男（蒙古族），教授，博士，博导，主要研究方向为蒙古文信息处理；额尔敦朝鲁（1976-），男（蒙古族），副教授，博士，硕导，主要研究方向为蒙古文信息处理。

型、文化含义、语言特征、表现形式或对照翻译等诸多视角进行过零散的研究。其中编撰蒙古语熟语辞书工作在19世纪末清朝时期已有木版，持有两百余年历史。如今《简明蒙古熟语解释词典》（2000）[1]、《蒙古语熟语大辞典》（2001）[2]等诸多工具书虽然很大程度上满足了人们多方面的需求。但是，从信息处理的角度来讲，随着服务对象的变换，印刷词典中的有些内容不能够直接应用到机器词典，其中面向人理解的分类、释义等信息对机器词典的非适应性更为突出。

自20世纪80年代至今，虽然蒙古文信息处理的基础研究和应用开发均有一定的成就，但是熟语的数字化研究相对滞后。熟语作为蒙古语多层级的、复杂系统的特殊单位，它横跨于词处理和句处理，往往被忽略。前人对于面向计算机理解的蒙古语固定短语语法属性的形式化方面有过深入的研究，其代表性成果有《现代蒙古语固定短语语法信息词典详解》[3]、《面向信息处理的蒙古语复合词研究》[4]等。形式上蒙古语熟语不但有短语形式，而且还有句子形式。类型上不仅仅只包括固定短语（）（它包括复合词、习用语、成语、固定词和名词术语等子类）还包括谚语（）、格言（）、讽刺语（）、忌讳语（）、典故（）、训词（）等许多不同子类。实质上固定短语与上述这些子类属于不同层面的分类。但是，这些句式熟语诸多子类的分类和澄清及其语义特征描述尤其面向机器理解的熟语语义特征研究至今仍处于空白阶段。这不仅直接影响信息检索、文本分类、机器翻译和语料库加工、树库构建等目前所进行的很多工作，而且对将来进行更深层次的研究和应用必将带来许多障碍。

2 资源库的规模与结构

构建熟语资源库的总体框架是基于词汇学和词典学理论，利用语料库和面向人理解的熟语辞书，抽取常用熟语作为基本词条构建熟语"总库"（样本库如图【1】所示）。目前总库已收录22000 条熟语。基于"总库"建立一些分库和辅助库表述熟语的单义或多义，相同或相近、相对或相反意义，以备计算机自动处理熟语时通过这些资源库实现歧义消解。



NO	HELECE	AYIMAG	TAYILBVRI	JERGECEGULUL	HVBILBVRI	OIRALCA	SANAMJI	MO
21361	TAMAHI TATABAL TANGNAI-YIN JIRGAL, TANIL-DAGAN JO	G	TAMAHI-BAN TATABAL SEREL-DU BAH_A, TANIL-I	YES	YES			
21362	TAMAHI TATAHV	Y	①SIGUMJILEGDEHU HELEGDEHU-YI YOGDALAGSAN					
21363	TAMAHI TATAGVLHV	Y	①SIGUMJILEHU JIMELEHU-YI ULU TAGASIYAGSA					
21364	TAMAHI-YI NI TATAGAD, TAR-I NI TANIHV	S	JIM_E-YI NI SONOSOGAD CIN_A-YI MEDEBE, INA					
21365	TAMAHIN-V HARIGV DARVI SAYIN, DAG_A-YIN HARIGV NA	G	UILE AJIL-I BUTUGEHU-DU TOHITAI CAG BA NO	YES				
21366	TAMAHIN-V HARIGV DARVI-DAGAN, DAGAN-V HARIGV NAMVI	G	TAMAHIN-V HARIGV DARVI SAYIN, DAG_A-YIN HA	YES	YES			
21367	TAMAHIN-V MONGGO	Y	OCOHEN JARVDAL GEJU JUIRLEGSEN UGE.					
21368	TAMAG_A BARIHV	Y	TORO MEDEHU, ERHE MEDEL-TEI BOLHV.					
21369	TAMAG_A BARIGSAN HATVN-ACA TAL_A TOGORIGSAN CIBA	S	OLBOG-IYAN COGOLOGSAN MERGEN-ECE, OLAN-I T			YES		
21370	TAMAG_A BARIGSAN HATVN-ACA TAL_A TOGORIGSAN CIBA	S	OLBOG-IYAN COGOLOGSAN MERGEN-ECE, OLAN-I T			YES		
21371	TAMAG_A BARIGSAN HATVN-V MERGEN-ECE TAL_A TOGORI	S	OLBOG-IYAN COGOLOGSAN MERGEN-ECE, OLAN-I T			YES		
21372	TAMAG_A BARIGSAN HATVN-V MERGEN-ECE TAL_A TOGORI	S	OLBOG-IYAN COGOLOGSAN MERGEN-ECE, OLAN-I T			YES		
21373	TAMAG_A-YIN GAJAR-VN NOHAI ANV INV GEJU HVCAN_A	H	BICIg-UN GER-UN BOLJOMOR BVYV GEJU NISUN_			YES		
21374	TAMAG_A-YIN JALAN	Y	GER MEDEGSEN EMEGTEI-YI HO$ONGNAGSAN HELE					
21375	TAMAGAN-DV TAL_A-TAI BOLOY_A GESEN BISI, TAL_A HA	H	HONI-BAN OGGUGED HACAR-VN SAGALI-TAI			YES		
21376	TAMIR-IYAN UJEJU BOHE BARILDV, TABCANG-IYAN UJEJU	G	ADVGVN-V-BAN HIRI-BER ISHER, HUNJILE-YIN-I	YES		YES		
21377	TAMIR-IYAN BARAGSAN MORI JOGSONGHI, TAR-IYAN BARA	S	UTELHU HECEGUU, CIDAL HUCU UGEI HECEGUU.	YES				
21378	TAMIR GARGAHV	Y	CAGAN_A-BAN NARIN VHAGAN GARGAJV MIHVVSLA					
21379	TAMIR-TAI HOMON	Y	①MIHVVS JISUR HOMON②DALDA BODOL-TAI HOM					
21380	TAMV-BAN UJEHU	Y	TAMV-BAN CAYIHV		YES			
21381	TAMV-BAN CAYIHV	Y	①JVBAHV, SINALAHV. ②TAMDVG UGEI MAGV YABV		TAMV-BAN UJEHL			
21382	TAMV-BAN CAYIJV, TAR-IYAN BARAHV	S	TAMV-BAN CAYIHV		YES			
21383	TAMV GEDEG CINI TAMDVG UGEI YABVGSAN-ACA	S	TAMV TAMV GESEN-ECE TAMV BAYIDAG UGEI, TAM	YES	YES			
21384	TAMV-YIN AMITAN, TAMV-DAGAN AMARAG	H	TAMV-YIN AMITAN, TAMV-DAGAN JIRGAL-TAI.		YES			

图【1】"蒙古语熟语资源库"总库样本图

2.1 "总库"的属性字段详解

（1）序号(NO)：自动生成的数字，表示资源库包含的熟语词条具体数目。

（2）词条(HELECE)：填置蒙古语熟语的拉丁转写形式。例如，""的对应字段中填置"NAMVR-VN EBESU HABVR-VN ALTA"。

我们收录常用熟语作为选词条原则，以下两种途径实现了词条的收录。

首先，以《蒙古语熟语大辞典》作为蓝本，从中人工选出常用熟语作为词条来源构建熟语总库。其次，从“100 万词现代蒙古语词性标注语料库”[①]中自动筛选出全部蒙古语熟语添加到“总库”。其中不能收录与之前收录的熟语重复的熟语。

（3）子类(BTOROL)：填置熟语语法分类的子类标记。例如，“ᠮᠣᠷᠢ ᠳᠤ ᠬᠥᠭᠵᠢᠮ ᠤᠨᠢᠭᠤᠯ ᠳᠤ ᠮᠣᠷᠢ” 的对应字段中填置“R”,表示“谚语”。

我们从信息处理的角度重新确定传统熟语研究中不明确的分类、界限、范围等问题，将会填补“信息技术 信息处理用蒙古文词语标记集”[②]中未能制定的熟语子类的标记，研制出合理而完整的“蒙古语熟语子类框架体系”及其标记集。目前已设定的子类标记如下表格所示。

蒙古语熟语										
固定短语					谚语	格言	讽刺语	忌讳语	训词	典故
复合词	习用语	成语	固定词	名词术语						
Y	X	K	J	NT	R	S	H	C	G	L

（4）主题(VTOROL)：填置熟语主题分类标记。例如，“ᠮᠣᠷᠢ ᠳᠤ ᠬᠥᠭᠵᠢᠮ ᠤᠨᠢᠭᠤᠯ ᠳᠤ ᠮᠣᠷᠢ” 的对应字段中填置“NuM”,表示“游牧”主题。

根据熟语的语义和语用属性,对其表示的主题意义进行分类,制定“熟语主题分类框架体系”及其相应的标记集。我们制定该分类框架体系时不但以语料分析结果为基础，而且以前人的熟语理论研究成果，譬如《蒙古语熟语研究》[5]、《蒙古语熟语中的文化积淀》[6]中的熟语文化含义分类建议作为参考初步制定熟语主题分类框架。

1. 物质（BOd）
  1.1 游牧(NuM)
    1.1.1 五种牲畜(TaM)
    1.1.1.1…其他家禽(BvM)
  1.2 狩猎(AA)
  1.3 饮食(IV)
    1.3.1 奶食品(CaI)
      1.3.1.1…
    1.3.2 肉食品（MaI）
  1.4 住宿(OR)
  1.5 服饰(HvJ)
2. 智慧(VV)
  2.1 道德(YOS)
    2.1.1 婚姻(HON)
    2.1.2 娱乐(NaN)
  2.2 家庭(GeB)
  2.3 信仰、祭祀($aT)
    2.3.1 宗教($aS)
    2.3.2 祭祀(TaH)
      2.3.2.1 祭天(TeT)
      2.3.2.2 祭火(GaT)
      2.3.2.2 祭地(GaJ)
    2.3.3 颜色(OnG)
    2.3.4 数字(TOG)
  2.4 忌讳(CeG)

[①] 该语料库的研发机构是内蒙古大学大学蒙古学院蒙古语文研究所；

[②] 这是 2008 年内蒙古大学蒙古学学院等几个机构共同制定的“中华人民共和国国家标准”（GB:ICS35.040 L71）；

…

（5）释义(TAYILBVRI)：填置每条熟语意义所对应的单词或短语。例如，“ ᠬᠦᠮᠦ ᠪ ᠬᠠᠪᠠᠷᠠᠯᠲᠠ ᠶᠢᠨ ᠬᠦᠮᠦᠨ ᠪ ᠲᠣᠳᠣᠷᠬᠠᠢᠯᠠᠭᠰᠠ”的对应字段中填写“ᠬᠠᠪᠠᠷᠠᠯᠲᠠ ᠶᠢᠨ ᠬᠤᠪᠢᠶᠠᠰᠤᠯᠠᠭᠤᠢ”（多余的、被排斥）。

该库中不管熟语的长度，使用与其意义相等或相近的词语诠释，这样使计算机更容易理解、处理熟语语义。我们参考《蒙古语熟语大辞典》的熟语释义，抽取出与当前熟语意义相互对应的单词或短语填置“释义”字段。

（6）实例(JI$IY_E)：填置语料库和纸质辞书中的实例。例如，“ᠬᠡᠨ ᠨᠢᠭᠡ ᠠᠮᠢᠲᠠᠨᠲᠤ᠂ ᠠᠪᠤᠷᠤᠭ ᠨᠢᠭᠡ ᠮᠣᠷᠢᠲᠤ᠂”的“实例”字段中填置“ᠲᠡᠷᠡᠭᠡ ᠴ — ᠬᠡᠨ ᠨᠢᠭᠡ ᠠᠮᠢᠲᠠᠨᠲᠤ᠂ ᠠᠪᠤᠷᠤᠭ ᠨᠢᠭᠡ ᠮᠣᠷᠢᠲᠤ᠂ ᠤᠨᠳᠤ ᠪᠠ/ ᠬᠡᠷᠡᠭᠯᠡᠭᠡᠴᠢ ᠪᠠᠶᠢᠭᠤᠯᠤᠯᠲᠠ ᠪᠠᠶᠢᠨ (ᠴᠢᠮᠡᠭᠲᠡ《 ᠵᠤᠪᠯᠠᠯ ᠪᠠᠶᠢᠴᠠᠭᠠᠯ ᠨᠤᠮᠤᠷ 》)”。

据观察，纸质辞书中的熟语例子较长。所以我们可用该库中的“实例”字段，独立构建 text 格式的“熟语语料库”，对未来研究提供基础语料。“熟语语料库”如图【2】所示。

图【2】“蒙古语熟语语料库”样本图

（7）是否短语式熟语（Phrase）：如果当前词条是短语式熟语应填置“Y”，如果不是“N”。

（8）是否句式熟语（Sentence）：如果当前词条是句式熟语应填置“Y”，如果不是“N”。

总体而言，目前已建立包括 22000 条熟语，具有“序号”、“词条”、“释义”、“子类”、“主题”、“实例”、“是否短语式熟语”、“是否句式熟语”等基本属性的蒙古语熟语资源库“总库”，如图【1】所示。

通过“总库”我们可以得知熟语的基本信息。譬如“总库”中的熟语“ᠤᠨᠳᠤᠰᠤ ᠤᠭᠯᠠᠭᠠᠨ᠂ ᠪᠠᠭᠠᠭᠠᠨ ᠪᠠᠢᠳᠠᠭᠠᠨ᠂”的基本属性为：1.【序号】：2213(用阿拉伯数字表示)；2.【词条】：DARVBACV DARDAYIN_A,BVLABACV BVLTAYIN_A；3.【释义】：ILE,TODORHAI(明显的)；4.【子类】：S(格言)；5.【主题】：常规事理；6.【实例】：OROS-VN ARAD TUMEN OLAN JIL-UN TVRSI OMOGTV BAGATVR ALDARTAI,TUMEN JIL-UN TEUHE-TEI,MINGGAN JIL-UN CVV-TAI,DARVBACV DARDAYIDAG BVLABACV BVLTAYIDAG,MOCIR-I HVGVLHV-DV UNDUSU NI ULEDEN_E(C.DAMDINSURUNG《JOBLELTU-YIN ARAD TUMEN-DU JORIGVLVGSAN IRUGEL》)；7.【是否短语式熟语】：N；8.【是否短语式熟语】：Y；

2.2 各分库和辅助库

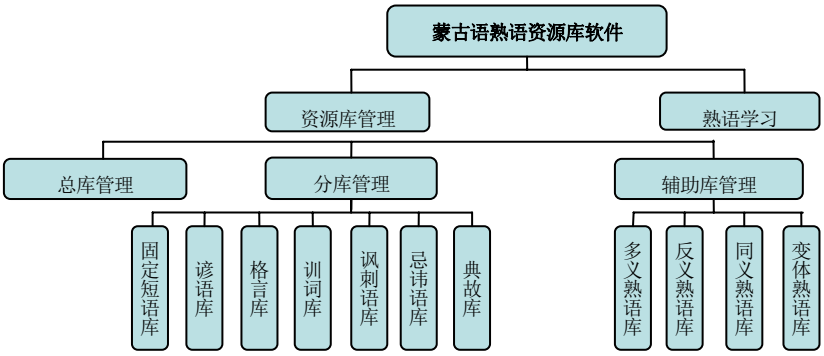
基于“总库”我们将构建“固定短语库”、“谚语库”、“格言库”、“讽刺语库”、“忌讳语库”、“典故库”和“训词库”等分库。各个分库继承“总库”的所有属性之外，还将要具备各自不同的特有属性。各个子类特征只能在其相应的分库中加以描述。同时我们还构建“多义熟语库”、“反义熟语库”、“同义熟语库”和“熟语变体库”等辅助库，描述熟语的多义、反义、同义和变体等特征。

多义现象在蒙古语熟语中普遍存在。例如，熟语“ᠨᠢᠭᠤᠴᠠ ᠪᠠᠷᠢᠯᠴᠠᠬᠤ”具备以下 3 种意义：【1】按照蒙古族习俗问候或见面；【2】摔跤；【3】打架。我们将多义熟语按照其不同的义项数设置不同的词条，分别描述其不同意义。熟语的同义和变体是有所区别的两种概念，我们通过属性描述将类似而异质的熟语单位作严格的比较，从比较中确定各自的特征和性质。

2.3 资源库相关软件设计

该软件由“资源库管理”和“熟语学习”等两大功能模块组成(图【3】所示)。前者通过对“总库”和各个分库、辅助库进行查询、添加、删除、浏览等诸多功能实现对数据库的管理和维护。后者是专设面向教育领域的模块，为用户提供掌握熟语的子类、主题、实例、释义…等基本信息之外，还有熟语的多义、反义、同义和变体…等多种语义知识的学习平台。

对于并列式熟语设“问答形式”功能，为学习者灵活掌握熟语提供便利工具。例如，熟语“ᠬᠠᠮᠤᠭ ᠶᠢᠨ ᠬᠦᠮᠦᠨ ᠪᠡᠷ ᠂ ᠬᠠᠮᠤᠭ ᠶᠢᠨ ᠠᠮᠢᠲᠠᠨ ᠪᠡᠷ”的上一句“ᠬᠠᠮᠤᠭ ᠶᠢᠨ ᠬᠦᠮᠦᠨ ᠪᠡᠷ”显示在学习界面上，下一句以（　　）填空形式让学习者猜测。当学习者不能填充下一句时可通过按“下一联”功能获取正确答案来掌握完整的格言。



图【3】熟语资源库管理和学习软件结构图

3　熟语资源库的应用前景分析

3.1　语言资源的保护

信息化、现代化新时代的浪潮不断推动着语言教育的发展，使书面语和口语渐趋接近，熟语收到越来越严重的冲击甚至有些地方已出现明显萎缩的势头，这是前所未有的巨大变化。由于当今蒙古语熟语语汇日益缩减,因此它所承载的蒙古族文化也将逐步消失,从而面临很大挑战。面向这一窘迫情况，搜集和记载蒙古族历代流传的文化遗产，以资源库形式保存正被遗忘或被遗漏的熟语是修缮整个蒙古语词汇、抢救濒临消失的蒙古语熟语的最佳途径。保护、开发和利用这一取之不尽的语言宝藏已成为我们研究者们刻不容缓的任务之一。“这样才能保证语言资源的健康、持续发展和长期利用”[7]。

3.2 蒙古文信息处理多个领域中的应用

（1）使计算机识别和理解蒙古语熟语知识，提高机器翻译的准确率。

熟语知识的形式化描述对于各种应用系统提供支持。由于蒙古语属于黏着性语言，其词汇体系的重要特征之一为一个“根词”不但能够派生单词，而且可以构成固定短语，甚至是句式熟语。例如，“ᠨᠢᠭᠤᠴᠠ -”（眼睛）这一“根词”通过缀接附加成分能够派生“ᠨᠢᠭᠤᠴᠠᠯᠠᠬᠤ”（注视、注目）、“ᠨᠢᠭᠤᠴᠠᠯᠠᠭᠳᠠᠬᠤ”（被注意）等单个动词，继而可以构成“ᠨᠢᠭᠤᠴᠠ ᠮᠥᠷᠥᠨ”（漫无边际）、“ᠨᠢᠭᠤᠴᠠ ᠪᠠᠷᠢᠯᠴᠠᠬᠤ”（歧视）等固定短语，甚至构成“ᠨᠢᠭᠤᠴᠠᠳ ᠪᠤᠰᠤ ᠬᠠᠮᠤᠭ ᠵᠢᠶᠠᠨ ᠪᠠᠷᠢᠵᠤ”（不放在眼里），“ᠨᠢᠭᠤᠴᠠᠳ ᠲᠤ ᠬᠠᠭᠤᠳᠠᠰ ᠬᠠᠮᠤᠭ ᠂ ᠵᠢᠷᠤᠭᠠᠳ ᠲᠤ ᠬᠠᠭᠤᠳᠠᠰ ᠬᠠᠪᠴᠠᠭ”（眼中钉）等句式熟语。无论是计算机还是人碰到一个熟语，不做语法结构分析的

前提下最大的难点是语义问题。熟语作为一个词汇单位，不管其长度有多长，最终表示一个词汇意义，因此诠释熟语时使用与其相等或相似的词语表示一个熟语语义单位是当今新型的“释义”方法。例如，熟语“ᠬᠠᠭᠤᠴᠢᠨ ᠪᠤᠤ ᠶᠢᠨ ᠬᠤᠭᠤᠷᠬᠠᠢ”（旧炮断子）的语义等于一个形容词“ᠬᠠᠲᠠᠭ”（熟练的），熟语“ᠣᠳᠣᠨ ᠤ ᠶᠡᠷᠤᠭᠠᠯ ᠬᠠᠷᠠᠬᠤ”（瞅天底）的语义等于一个动词“ᠪᠠᠳᠠᠷᠠᠬᠤ”（骄傲）等。这种方法使计算机更容易理解、处理蒙古语熟语语义，能够解决句法处理和机器翻译过程中熟语语义处理问题从而能够提高机器翻译的准确率，也为树库构建、语义网络甚至对隐喻、认知模式等高层研究提供知识支撑。

（2）对于语料库的词法标注和熟语标注提供基础资源，为构建蒙古语多级加工语料库奠定基础。

熟语是自然语言中普遍存在的现象。从语言分析的角度看，任何语料库的加工都很难避开这个问题。蒙古语熟语形式化研究方面，如同上述，德•青格乐图等开发的《面向信息处理的蒙古语固定短语语法信息词典》总库中囊括 7000 多条蒙古语常用固定短语，设置 17 项字段描述每个固定短语的语法属性。该库中定义的固定短语包括复合词（Y）、习用语（X）、成语（K）、固定词（J）、名词术语（NT）等 5 大类。复合词又可分为复合名词（Yn）、复合形容词（Ya）、复合代词（Yr）、复合时位词（Yo）、复合动词（Yv）和复合副词（Yd）[8]。我们可以采用人机互助方法，利用该词典的“词性”字段，在“100 万词现代蒙古语词法标注语料库”中进行固定短语标注。其中利用该固定短语数据库对“100 万词级词法分析数据库”进行匹配处理之后，通过人工校对构建固定短语标注语料库。由于该词典不能涵盖句式熟语，无法标注语料库中的句式熟语。如同下列例句 1 中带有下划线的语料是无法标注的熟语语料。

例句 1.

<</Wp1 <u>TANI/Ve1+LCA/Fe4+HV/Ft12-ACA/Fc41 EMUN_E/Oa TAL_A/Ne2-YIN/Fc11 GVRBA/Mu+N/Zx GOROGESU/Ne1 ,/Wp1 TANI/Ve1+LCA/Fe4+GSAN/Ft11-V/Fc12 H0YIN_A/On-BAN/Fx11 TANGNAI/Ne2-YIN/Fc11 GVRBA/Mu+N/Zx SVDASV/Ne1</u> >>/Wp1 GE/Vx+GCI/Ft33-BER/Fc51 []HESIGT0G TAHV/Nt1-YIN/Fc11 EG=MEG/Yd HI/Ve1+HU/Ft12 NI/Sf CU/Sq AYADA/Ve2+JV/Fn1 TEDE/Rb+N/Zx-U/Fc12 ASAGV/Ve1+GSAN/Ft11-I/Fc31 NI/Sf MEDE/Ve1+HU/Ft12-BER/Fc51-IYEN/Fx11 HELE/Ve1+JU/Fn1 ENE/Rj TEREGUR/Rj NI/Sf T0G0RI/Ve1+HV/Ft12 B0L/Vz2+JAI/Fs11 ./Wp1

我们可以利用“熟语资源库”对于上述句式熟语语料进行子类信息标注（这里用括号括上熟语，加上斜杠，之后添加熟语子类标记，“R”表示谚语）之后的语料如下例句 2 所示。

例句 2.

<</Wp1[T<u>ANI/Ve1+LCA/Fe4+HV/Ft12-ACA/Fc41 EMUN_E/Oa TAL_A/Ne2-YIN/Fc11 GVRBA/Mu+N/Zx GOROGESU/Ne1 ,/Wp1 TANI/Ve1+LCA/Fe4+GSAN/Ft11-V/Fc12 H0YIN_A/On-BAN/Fx11 TANGNAI/Ne2-YIN/Fc11 GVRBA/Mu+N/Zx SVDASV/Ne</u>1]/R >>/Wp1 GE/Vx+GCI/Ft33-BER/Fc51[]HESIGT0GTAHV/Nt1-YIN/Fc11 EG=MEG/Yd HI/Ve1+HU/Ft12 NI/Sf CU/Sq AYADA/Ve2+JV/Fn1 TEDE/Rb+N/Zx-U/Fc12 ASAGV/Ve1+GSAN/Ft11-I/Fc31 NI/Sf MEDE/Ve1+HU/Ft12-BER/Fc51-IYEN/Fx11 HELE/Ve1+JU/Fn1 ENE/Rj TEREGUR/Rj NI/Sf T0G0RI/Ve1+HV/Ft12 B0L/Vz2+JAI/Fs11 ./Wp1

课题组完成 80 万词级语料库的固定短语标注和人工校对之后发现“固定短语词典”尚未覆盖“100 万数据库”中的全部熟语。所以“熟语资源库”为语料库标注提供熟语子类信息、主题信息，为词法标注、熟语标注的多级加工语料库的构建提供资源。

3.3 语言教学中的价值

熟语资源库的构建及其学习软件的开发，它一方面对于蒙古族中小学生的学习母语、写作等

语言教学提供平台，培养和提高学生使用母语素养；另一方面，作为新型的教学辅助工具服务于对外蒙古语教学。熟语是具有修饰功能、表现力很强的一种固定形式，在写作和翻译中几乎不可或缺的元素。但是，近年来熟语的流失现象在蒙古族青少年尤其在城镇中小学生中表现得更为普遍。我们认为较长的句式熟语和并列式熟语是学习蒙古语熟语的难点。例如，当学生碰到如同“ᠬᠥᠳᠥᠯᠪᠦ ᠰᠠᠭᠠᠭᠠᠨ/ᠬᠥᠳᠥᠯ ᠬᠦᠴᠦᠨᠦ ᠲᠡᠷ ᠪᠠᠢᠬᠤ ᠪᠤᠶᠤ᠃”（只要努力会有精彩）等较长的句式熟语时，正确切分，指明其包含的褒贬色彩，从而能够正确理解、得体运用该熟语是个重要环节。对于“ ᠬᠦᠮᠦᠨ ᠬᠣᠭᠣᠷᠣᠨ ᠳᠤ ᠪᠡᠳᠦᠭᠡᠯ᠂ ᠬᠤᠷᠠ ᠬᠣᠭᠣᠷᠣᠨ ᠳᠤ ᠪᠡᠳᠦᠭᠡᠯ ᠪᠦᠬᠦ ᠪᠠᠢᠬᠤ ”（不能食言）等并列式熟语，正确连接上下联，避免导致掌握上出现偏误也是关键。除此之外，基于大规模真实文本取出来的熟语具有两方面的副产品可资利用：一是为词典提供真实的候选例句；二是为词典释义提供一定参考价值的释义信息。

4 小结

蒙古语词汇系统中的熟语词汇量大，类别多。传统研究中熟语子类之间的划分和界定、同义熟语和同一个熟语的不同变体之间的区分不明确等诸多欠缺或不一致性对于我们的研究带来了一定的难度；除此之外，根据熟语语义和语用进行主题分类，对短语式和句式熟语进行整体形式化描述和分析尚属第一次。因此，充分利用前人研究成果的同时借鉴英语、汉语的相关理论和方法致力构建有助于人理解熟语、有利于语言信息处理应用的资源库是此项研究的目标。

参考文献

[1] 格.阿凯姆编.简明蒙古熟语解释词典 [M] .内蒙古文化出版社，2000.

[2] 布林特古斯编.蒙古语熟语大辞典[M] .内蒙古教育出版社，2001.

[3] 德.青格乐图等.现代蒙古语固定短语语法信息词典详解[M].内蒙古教育出版社，2005.

[4] 德.青格乐图著.面向信息处理的蒙古语固定词组研究 [M] .内蒙古教育出版社，2001.

[5] 陈章太，论语言资源.中国语言资源论丛（一）.张普，王铁琨主编.商务印书馆，2009：13.

[6] 德 •青格乐图等.现代蒙古语固定短语语法信息词典详解[M].内蒙古教育出版社，2005：4.

[7]王海风.蒙古语熟语研究 [D] .内蒙古大大学博士学位论文，2010（6）：113-128.

[8]陈玉荣.蒙古语熟语中的文化积淀[D].内蒙古师范大学硕士学位论文，2010（5）：18-61.