# 面向机器翻译的蒙古语生成*

那顺乌日图(内蒙古大学)　刘群(中科院计算所)　巴达玛敖德斯尔(内蒙古大学)

**摘要：**蒙古语作为目标语言，在机器翻译中除了需要词语的转换以外，更重要的是生成出符合蒙古语语法的词语和语句。蒙古语属黏着性语言，其构词、构形都是通过在词干后缀接不同的词尾而实现，而且可以层层缀接，层层派生。其中蒙古语的构形词尾更是变化多端、错综复杂。譬如，蒙古语中的一个动词可以有 850 多种变化形式，在机器翻译中如何正确地选择和生成这些形式，是蒙古语机器翻译的关键所在。本文主要从蒙古语作为目标语的角度探讨了蒙古语语法形态的生成策略和技术。

**关键词：**汉蒙机器翻译 翻译系统 蒙古语 生成

# Mongolian generation in machine translation

Nasun-urtu (Ineer Mongolia University　Email:mgnasun@imu.edu.cn)
Liuqun (Institute of Computing Technology Chinese Academy of Sciences )
Badmaodsar (Ineer Mongolia University )

**ABSTRACT:** In machine translation, as the target language, Mongolian language generation is rather complicated. In Mongolian, word formation and grammatical formation are all achieved by attaching different suffixes to the stems. A certain Mongolian verb can have more than 850 forms. This article discusses the generation strategy and techniques of Mongolian grammatical morphology in the case that Mongolian is the target language.

**Keywords:** Machine Translation, Language Generation, Mongolian Processing

## 1、汉蒙机器翻译系统的基本构成

1998 年内蒙古大学蒙古学研究院蒙古语文研究所、中国科学院计算技术研究所、北京大学计算语言学研究所共同承担了国家 863 项目“面向政府文献的汉蒙机器辅助翻译系统”，旨在研究开发出一个面向政府文献（如，政府工作报告、各种决议等）的汉蒙机器辅

* 此项研究得到国家 863 计划资助，项目号：863-306-ZT04-05-3

助翻译系统。在机器翻译系统中，作为典型的黏着型语言的蒙古语的生成，比起众多的西方语言，更具特色。而且从孤立型语言的汉语到黏着型语言的蒙古语的机器翻译，更有其独特的一面。

该系统由汉语分析、蒙古语生成、翻译软件、用户接口等四个部分组成。

## 2、汉语分析部分

汉语分析部分包括对文本进行切词、标注、分析等一系列处理工作。本系统采用的语言模型主要来源于北京大学计算语言学研究所研制的《现代汉语语法信息词典》[①]（以下简称《词典》），并在该词典所采用的语言模型基础上修改扩充而成。

（1）汉语词语分类和属性。本系统采用的汉语词语分类和属性取自于《词典》，并作了少量的改动。《词典》中有大量的属性描述，我们根据机器翻译的需要对这些属性作一定的取舍，并增加少量新属性。

（2）汉语短语分类和属性。对汉语短语的分类，我们继承《词典》中对汉语词语分类时采用的“功能分类”思想，将短语（包括句子）分成 12 类。另外，我们还定义了内部结构、语气、被动、否定等短语属性。

（3）语义分类和属性。本系统是一个以语法分析为主，语义分析为辅的系统。虽然如此，为消解句法分析和转换时的歧义，语义分析还是起着重要的作用。

本系统采用的语义模型主要包括语义分类和配价分析两个方面。我们制定了一个比较详尽的语义分类体系，对每一个汉语实词都要填写其相应的语义分类，而对于名词、动词、形容词三类词语还要填写配价数以及相应配价成分的语义类。在规则的约束条件中，对某些短语的组合规定一定的配价关系，如果这种关系不能被满足，则合一失败。这样就排除了相当一部分由于搭配不当所造成的歧义。

## 3、蒙古语生成部分

蒙古语生成部分包括蒙古语语言模型、蒙古语生成规则集、从汉语到蒙古语的转换规则集和汉、蒙对照机器词典。

蒙古语作为目标语言，在机器翻译中需要词语的转换以外，更重要的是生成出符合蒙古语语法的词语和语句，其中所谓词语的生成就是蒙古语词的派生和各种语法形态的生成。由于在本系统中，蒙古语的生成属最具创新特色的部分，所以对蒙古语生成，做比较详细的阐述。

---

[①] 该词典是为计算机实现汉语句子的自动剖析与自动生成而研制的一部电子词典,收录了 51696 个词语（1995 年电子版）。有关这部电子词典的详细介绍请见《现代汉语语法信息词典详解》一书，清华大学出版社、广西科学技术出版社，1998 年。

（1）蒙古语语法形态的复杂性

众所周知，蒙古语属黏着性语言，其构词、构形都是通过在词干后缀接不同的词尾而实现，其中蒙古语的构形法更为复杂。我们拿 UILED[1]（做）来举例：

UILED（ 动词、做）接 BURI（名词构词附加成分），派生出 UILEDBURI（名词、工厂），再接 LEL（名词构词附加成分）派生新的名词 UILEDBURILEL（名词、生产），如果在 UILEBURI 后接动词构词附加成分，还会派生出很多新的动词词干。

动词 UILED 作为动词的第一词干，其后可以直接缀接动词的陈述式、祈使式、形动词、副动词等 33 个结尾形式后缀，我们不妨把这些称为第一层变化。

蒙古语的动词还有态、体的变化。如，UILED 本身是一个动词词干，在式范畴里它表示第二人称祈使式、在态范畴里它表示自动态，在体范畴里它表示一般体，并且它是动词各种词法变格的基础。在 UILED 后接态、体范畴后缀，还可以生成出 8 个新的词干（如果在其后接别的后缀，它们可以作为中间形式，如果不接别的后缀，他们则是结尾形式），我们可以称其为蒙古语动词的第二层变化。这 8 个词干上还可以接陈述式、祈使式、形动词、副动词等动词结尾形式，可谓第二层生成。我们可以计算出蒙古语动词的第二层生成的可能性是 8×33=297 个形式。蒙古语动词的又一个特点是，态范畴词缀之后还可以接体范畴词缀，表示某个动作的被动完成、被动进行等等语法意义。除此之外，某些态的形式还可以重叠，表示更为复杂的语法意义。如，UILED\U\GUL\JEGE、UILED\U\GDE\JEGE、UILED\U\LCE\JEGE(有些形式，在这个词上可能不会出现，但在其它词上就可以出现了)等等，通过这一层变化，又可派生出 24 个新的词干，而且其后面还要接陈述式、祈使式、形动词、副动词后缀，可谓第三层生成。这一层，如果按简单的数学方法计算，可以有 24×33=792 种变化。从中除去因语义因素而不可产生的一些形式至少可以产生 492 种变化形式，再加上第一、二层次所产生的形式 UILED 这个动词至少有 33+9+24+297+492=855 种变化形式。而且这还是对已确定的、具体的词而言的，如果是一个不确定的、概念上的“词”，那么每一个词尾还有阳性、阴性两种不同选择，这样，其实一个动词语法形态的变化至少可以有 1710 个选择项。在这样数目可观的选择项中正确地生成当前条件下的正确项，就必须编好生成规则和与之配套的属性字段和取值规范。更为重要的是应采取一些适合于蒙古语生成的恰当的策略。

（2）生成蒙古语的具体措施

在我们开发的汉蒙机器翻译系统中，是以一个语言模型、一个词法生成规则集、一部对译词典和一个结构转换规则集来实现蒙古语词语的转换和蒙古语词语、语句的生成的。蒙古语语法属性分静态属性和动态属性两类，静态属性是同一个词类内部词语之间的区别特征，如，有些名词是可数的、其后可以接复数附加成分表示多数，而有些词是不可数的，

[1] 由于现在所用的蒙古文系统与其他系统不兼容，在本文中采取了将蒙古文转写成拉丁文的方法。

其后面不能接复数附加成分；有些动词是及物的、其前面可以带直接宾语，而有些动词是不及物的，不可以带直接宾语；有些动词可以受副词修饰、而有些则不可以。这些是静态属性，通过在词典中设置语法属性字段来控制其生成。同一个词类所共同具有的语法特征属于动态属性，如大部分动词（除一些特殊动词）都有陈述式、祈使式、形动词、副动词等语法变化。这些语法形态的生成，是通过语言模型和生成规则的结合来实现的。

蒙古语词法生成规则总共 76 组，含 630 条规则，这些规则总共分成三个层次，第一层主要处理动词第一词干的结尾形式的生成，第二层规则主要处理动词的态、体等语法范畴的生成，这一层也是新的动词词干的派生环节。蒙古语动词态范畴之后还可以有体、态词缀，而且其后还要接终结形式，所以第三层是生成这些更为复杂的语法范畴及其形式的。

蒙古语词法生成的算法如下：

①将当前规则指针指向第一层的第一条规则；

②将当前词形设置为蒙古语单词的原形；

③执行以下循环，直到当前规则指针为空：

a) 将当前规则与当前词形匹配，如果匹配成功，那么：

i. 执行当前规则，得到新的当前词形；

ii. 跳过所有的同层次规则，将当前规则指针指向下一层的第一条规则；

b) 如果匹配失败，那么将当前规则指针指向下一条规则；

④返回当前词形。

下面是蒙古语陈述式现在时的生成规则：

```
@@*_A--V[TTT:TQGACI,CAG:CAGODO,GENDER:MALE] >>*AN_A
@@*_E--V[TTT:TQGACI,CAG:CAGODO,GENDER:FEMALE] >>*EN_E
@@*[N|L|B|G|R|S|D]--V[TTT:TQGACI,CAG:CAGODO, GENDER:MALE] >>*[]/V/N_A
@@*[N|L|B|G|R|S|D]--V[TTT:TQGACI,CAG:CAGODO,GENDER:FEMALE|NEUTRAL]   >>*[]/U/N_E
@@OG--V[TTT:TQGACI,CAG:CAGODO]>>OGGON_E
@@[HOB|JIB|CIB]--V[TTT:TQGACI,CAG:CAGODO] >>[]/BU/N_E
@@*--V[TTT:TQGACI,CAG:CAGODO,GENDER:MALE] >>*N_A
@@*--V[TTT:TQGACI,CAG:CAGODO,GENDER:FEMALE|NEUTRAL] >>*N_E
```

第一条规则表示，以分写的“_A”元音结尾的动词（规则中以*_A--V 表示），如果它是陈述式（TTT:TQGACI），它的时间属性为现在时（CAG:CAGODO）、并且它是一个阳性词（GENDER:MALE），那么，它会生成陈述式现在时形式*AN_A。这条规则实际上解决了两个问题，第一，它正确生成出了陈述式现在时词尾 N_A，第二，把原词干的分写形式转换成连写形式（蒙古语词干中除了最后一个音节可以有分写形式以外，其它音节不可以有分写形式），例如动词 TQGTAG_A（规定）通过这一规则就可以生成其陈述式形式 TQGTAGAN_A。第二条规则是分写的_E 元音结尾的词的生成，在这里除了阴阳性不同外其他部分是相同的。第三、第四条规则是辅音结尾词的生成规则，以辅音结尾的词，按蒙古语正字法规律不能直接接以辅音开头的词尾，中间必须加一个连接元音，本规则也满足了这一要求。第五、第六条规则是几个特殊动词的生成规则，上述六条规则都属于特殊规则，只有第七、第八条规则才是通用规则。如果我们把 TTT:这个属性字段的值改变成 TTT:TEMDEG，再通过 TEM:这一属性字段确定形动词的哪一个（如 TEM:TEMBAI 表示经常性形动词），并将生成的词尾及其正字法部分相应调整过来就会生成形动词词尾。这是生成第一层的例子，如果要是生

成第二层，那么还要限定其态或体的情况，如果我们加一条生成使动态的规则，就会生成出该动词使动态、陈述式、现在时的形式（如，UILED\U\GUL/U/N_E），即：

@@*[N|L|B|G|R|S|D]--V[HEB:HEBGVL,TTT:TQGACI,CAG:CAGODO,GENDER:FEMALE|NEUTRAL]

>>*[]GUL/U/N_E

第三层是在态范畴上再加体或态范畴，再加式和时范畴，就要写下列规则：

@@*[N|L|B|G|R|S|D]--V[HEB:HEBGVL,BAI:BAICIH,TTT:TQGACI,CAG:CAGODO,GENDER:FEMALE]

>>*[]UGULCIHE/N_E

这样就会生成出 UILED\U\GUL\CIHE/N_E 形式。

由于蒙古语里，态、体、式等范畴是按一定的顺序来缀接的，如果在一个词中这些范畴都有，那就按蒙古语的语法规律，以态、体、式的顺序来缀接。在生成规则库中虽然各有各的规则，但这些规则通过在程序中的合一运算，调整其前后顺序，生成出正确的形式。

那么，词语的这些生成信息是从和而来的呢？它们是由三个不同渠道来的，即语言模型，生成规则、转换规则和词典。

蒙古语语言模型(Monmodel)是由面向信息处理的蒙古语词语分类和标注规范以及词类、语法属性标记集构成的。该模型把蒙古语的词语分成 15 个基本类、6 个附加类。在基本类下面还有若干个子类，如，名词子类有专有名词、普通名词，普通名词还可分为可数名词和不可数名词。我们在语言模型中以

```
LexVal (NSUBC) : Hierar Limited
        {
            NPROP,          /*专有名词*/
            NNORM,          /*普通名词*/
            NNORM1=NCONT,   /*可数名词*/
            NNORM2=NUCNT    /*不可数名词*/
        } Default=NUCNT
```

来表示。对机器翻译词典中的每一个名词词条都有 NSUBC 和它的取值（NPROP, NCONT, NUCNT 中的一个）。语言模型不仅确定一个词类的标记、子类，还要确定这个词类有哪些语法属性，并规定其属性字段及其取值。如，名词的语法属性字段为：

```
LexAtt [MCAT:N] :
        {
            NSUBC,   /*名词子类*/
            NPLUR,   /*名词复数形式*/
            NREGU,   /*名词复数形式是否规则变化*/
            NQUAN,   /*名词能否做量词*/
            NHODL,   /*名词不定 N*/
            NHEMJ    /*构成数量词组时是否前接量词*/
        }
```

这些属性字段的取值为：

```
LexVal (NPLUR) : Symbol Unlimited
LexVal (NREGU) : Boolean Default=No
LexVal (NQUAN) : Boolean Default=No  /* Yes: 名词能作量词 */
```

LexVal (NHODL) : Boolean Default=No /* Yes: 名词有不定 n 形式 */

LexVal (NHEMJ) : Boolean Default=Yes /* No: 不要求量词*/

蒙古语语言模型包括 17 组 60 个属性值对。

因为转换模块和生成模块都从以语言模型中所规定的属性字段和属性值为基础而编制的转换规则集、及其词典和生成规则集中提取所需语言知识。语言模型的结构和整体设计必须符合机器翻译中语言生成的要求。

汉语到蒙古语的转换规则集（prsrbase）是汉、蒙两种语言的短语结构特征和对应规律的知识库系统。由于汉、蒙两种语言属不同语系，其短语、句子的构成等有着很大的差异，所以对汉、蒙两种语言短语结构的转换也是至关紧要的。如，汉语的介词结构，在蒙古语中可能对应于一些语法形态，也可能对应于一些虚词或短语结构，而且其位置和功能还要发生很大变化。譬如，

&& {pp8} pp->!p ap :: $.内部结构=介宾,%p.体谓=体,%ap.内部结构=的字,$.宾语=%ap,$!=%p @语气,$==%ap @语气,$.主题成分=%ap

|| IF %p.yx=对 TRUE

=> NP(NP/ap !G/p) %NP.CASE=OGH

在这一条规则里面，汉语的 ap 对应于蒙古语的 NP，且蒙古语的 NP 是以与格变化的，汉语的介词则对应于蒙古语的后置词，并处在中心词之后。在翻译过程中遇到这种结构，系统就会自动生成当前词的与格形式。转换规则库现有约 410 条这样的规则。

汉蒙机器翻译词典（dictn）是将汉语和蒙古语两种语言的词语进行对照的机器可读词典。它把汉语一个词所对应的蒙古语词语及其在当前所环境下的语法形态一同提供给系统。系统根据词典所提供的信息正确生成其语法形态。如，

$$ 报告

**{n} n $=[名词子类:na,数量名:YES,个体量词:份|个|篇,前名:可,前动:可,后名:可,名状语:NO,临时量词:NO,兼类:v,语义类:作品|信息]

=> N<ILEDHEL> $=[NSUBC:NCONT,GENDER:FEMALE,NPLUR:"-UD"]

=> N<MEDEGULULTE> $=[NSUBC:NCONT,GENDER:FEMALE,NPLUR:"-NUGUD"]

&&{npdzbg}np->np !np::$.内部结构=定中

=>NP(NP/np !NP//np)%NP.CASE=HAR

**{v} v $=[谓词性主语:NO,系词:NO,助动词:NO,趋向动词:NO,补助动词:NO,形式动词:NO,准谓宾:准,前名:可,后名:可,体谓准:体,双宾:YES,兼语句:NO,后动量词:动,后时量词:NO,动结:粘,动趋:趋,趋向补语:~过,不:YES,没:YES,很:NO,单作主语:可,单作谓语:可,单作补语:NO,兼类:n,语义类:对待,配价数:3]{主体:[语义类:人类],客体:[语义类:事情|信息],邻体:[语义类:人类]}

=> V<MEDEGUL> $=[VSUBC:VTVS,GENDER:FEMALE]

&&{djkzbg}dj->dj pp !vp w<",|", ">::$.内部结构=扩展,%vp.内部结构=述宾

=>DJ(VP/dj NP/pp !VP/vp W/w) %%VP.TTT=NOHEC,%%VP.NOH=NOHJER

在上面这个例子中，词典给出了汉语“报告”这一词的各种语法信息和与其对应的蒙古词语及其语法形态。汉语的语法属性字段提供汉语分析所需语法信息，蒙古语语法属性字段提供生成蒙古语所需语法信息。

## 4、翻译部分

这一部分包括一个汉蒙对照政府文献语料库，一部汉蒙对照机器词典和汉蒙机译软件三大部分。其中汉蒙机译软件是关键，它不仅包括汉蒙两种语言的分析、生成、转换，而且还涉及到输入、输出蒙古文的问题。翻译软件，我们利用由中国科学院计算技术研究所和北京大学计算语言学研究所联合开发的 863 成果-“通用机器翻译开发平台”，在该平台的基础上进行开发。该平台将机器翻译中常用的数据结构和算法以软件构件的形式提供出来，并已在该平台基础上开发成功了一个汉英机器翻译系统。使用该平台可以大大加快机器翻译系统的开发进度，节省大量人力、物力。该平台虽然以汉英机器翻译系统作为第一个开发对象，但其源语言的语言模型和计算模型实际上是通用型的，将其应用于汉蒙机器翻译的开发时，其描述语言和实现算法做一些改动，转换生成则需要作一些改变以外，主要的工作是研制蒙古语生成规则和相应程序。

我们采用给予转换的翻译方法，严格遵循独立分析、独立生成的设计原则。其中，汉语的词形分析阶段分为重叠词处理和切分两个步骤，汉语的切分采用双向最大匹配算法。出现切分歧义时，不做判断，保留到结构分析阶段进行处理。结构分析阶段采用 Chart Parsing 算法，转换阶段采用自顶向下自低向上相结合的局部子树变换算法。结构生成阶段采用自低向上的局部子树变换算法和自顶向下的全局子树易位算法。

## 5、用户接口

本系统开发了一个蒙古文显示输出系统，最终用户在 WINDOWS 环境下的蒙汉文同显窗户里进行编辑。虽然蒙古文是拼音文字，但与其它文字系统不同，蒙古文是在一个行里是由上而下连写，而每个行则由左而右易行。这样，蒙古文的显示输出不能只利用 WINDOWS 操作系统的造字功能实现蒙古文输入输出就可以解决，而必须开发一个竖向的蒙古文编辑器。由于蒙古文编码国际标准刚刚通过，还没有系统实现，我们在本系统中采用蒙古文的内部表示用罗马字的方法，通过一个转换规则，在屏幕上显示蒙古文，最终显示采用蒙古文显示窗口。

## 参考文献

[1] 冯志伟：“自然语言的计算机处理”，上海外语教育出版社，1996 年，上海。

[2] 冯志伟：“自然语言机器翻译新论”，语文出版社，1995 年，北京。

[3] 高文、钱跃良主编“’99 智能计算机接口与应用进展”，电子工业出版社，1999 年，北京。