

博士论文丛书

RESEARCH ON MT-ORIENTED TRANSFORMATION RULES OF
CHINESE-MONGOLIAN PHRASES

面向机器翻译的汉蒙 短语转换规则研究

巴达玛敖德斯尔 著

内蒙古教育出版社

博士论文丛书

RESEARCH ON MT-ORIENTED TRANSFORMATION RULES OF
CHINESE-MONGOLIAN PHRASES

面向机器翻译的汉蒙 短语转换规则研究

巴达玛敖德斯尔 著

内蒙古教育出版社

图书在版编目(CIP)数据

面向机器翻译的汉蒙短语转换规则研究/巴达玛敖德斯尔著. —呼和浩特:内蒙古教育出版社, 2005. 12
ISBN 7—5311—6208—3

I. 面... II. 巴... III. 蒙古语(中国少数民族语言)—短语—机器翻译—研究 IV. H212.59

中国版本图书馆 CIP 数据核字(2005)第 129358 号

博士论文丛书

面向机器翻译的汉蒙短语转换规则研究

巴达玛敖德斯尔 著

出版·发行/内蒙古教育出版社

经销/新华书店

印刷/内蒙古邮电印刷厂

开本/890×1240 毫米 1/32 印张/4.5 字数/125,000

版本/2005 年 12 月第 1 版 2006 年 3 月第 1 次印刷

印数/1—1 100 册

社址/呼和浩特市新城区新华东街维力斯大厦 9 层

电话/(0471)6608179、6608165 邮编/010010

出版声明/版权所有,侵权必究

书号:ISBN7—5311—6208—3/G·5710

定价:17.00 元

如发现印、装质量问题,影响阅读,请与内蒙古教育出版社联系调换。

序

巴达玛教德斯尔博士的学位论文将要出版了。这对他是个值得庆贺的事情,同时对蒙古语文信息处理,尤其是对汉蒙机器翻译来说,也是一个可喜的事情。他让我给写一篇“序”。这对我是个不大不小的“难题”。因为我虽然某种程度上接触了一点汉蒙机器翻译这个信息处理的“最终难点积蓄地”,但是基本上停留于学习、观看他人的研究成果这样一个“皮毛”之处,自己没有做过专门研究。因而很难说出像样的“子丑寅卯”。不写吧,他是我的最后一位博士生,其他几位出书时都给写了个“序”,就他一个不写也实在说不过去,只好硬着头皮写。

巴达玛教德斯尔博士在这篇学位论文中主要探讨了汉蒙机器翻译中的汉蒙短语转换规则问题。论文可分为蒙古语短语的研究和短语转换规则两个部分。在蒙古语短语研究部分,作者提出短语里包括“实词与虚词的组合”;在短语的分类部分,提出了数量词短语、时间词短语、时位词短语和后置词短语等较新的分类,也提出了与此相应的“短语标记集”;还对10种短语的结构类型进行了分类研究。这些研究成果不仅对汉蒙机器翻译有参考价值,对传统蒙古语文研究也会起良好的启迪作用。这个论文的主要“亮点”在于它的各类“转



换规则”。论文在对比分析研究汉蒙短语的基础上提出了有关 np、vp、ap 和 pp 短语的 100 余条全局性转换规则和 100 余条局部性转换规则。据测试,翻译准确率达到 90% 左右。这里必须强调指出的是,这个“准确率”只是在本短语规则的“运行”中得出的,这个数字不能代表整个翻译系统的“准确率”。我们知道,运行某几种特定“局部性规则”的结果和整个翻译系统的运行结果不是一回事。往往几个特定局部性规则的运行和全局性规则的完整运行会产生一些“冲突”,影响整个结果。

巴达玛教德斯尔博士的这篇论文是基于“规则”的汉蒙机器翻译研究的部分—阶段性成果。之所以称它为部分—阶段性成果,是基于如下情况:首先,这个论文从汉语的那么多种短语里只选了 np、vp、ap 和 pp 等 4 种短语,对其余短语的转换规则还没有涉及;其次,构成 pp 短语的汉语介词有 80 多个,论文涉及的只是其中的 29 个,其余的 50 多个还没有研究;再次,机器翻译不研究解决“词义”难题,也不可能取得突破性进展,而这个论文还没有涉足于此。因而它只能是部分—阶段性成果。我认为,这种定位,决没有任何贬低的意思。准确定位是进一步研究的前提和动力。现在社会上流行的那种“吹嘘和自我吹嘘”的极其有害的“风气”我们应当拒之门外。饭要一口一口地吃,问题要一部分一部分地解决。我们不能指望一个论文或者一个人就把汉蒙机器翻译的问题都解决了。只能靠好些个人、好些个相互配合的团队的阶段性成果的一步一步地长期积累来接近最终结果。

据有关专家的分析,机器翻译的研究经历了几个阶段的探索,大致可以分为基于“规则”方法的阶段和基于“语料库”方法的阶段。基于“规则”的机器翻译的研究还远远不能满足实际需要。在这个方向上必须做的工作还有很多很多。在认清这个问题的同时,我们必须看到问题的多样性或复杂性。

以近年来的研究结果看,只依靠“规则”的机器翻译受到语言研究、翻译研究的很大限制,有些疑难问题,至少在可以预见的近期内几乎无法圆满解决。因而出现某种从以“规则”为主的机器翻译研究转到以“语料库”为主的机器翻译研究或者是两者并用的研究的趋



势。以我的初步判断,这两种方法的结合很可能会较好地解决问题。但这两种方法应该各占多大比例?哪些问题用“规则”解决好?哪些问题应该用“对齐语料库”解决?没有现成的答案。我们需要结合汉蒙翻译的实际很好地充分地学习、利用他人的研究经验和成果,研究出自己的“答案”。为此必须付出很艰苦、很“枯燥”的劳动。这是问题的另一方面。

巴达玛教德斯尔博士,在汉蒙机器翻译方面开始做了很好的工作,并取得了一定的成就。这些成就具有一定的“开创性”,这很不容易。这个成就应该是继续研究的起点,而决不是、决不应该是某种“终点”或“转折点”。

机器翻译,是个庞大的系统工程,只能是依靠“兵团作战”,只能是依靠“持久战”。我们内蒙古大学蒙古语文研究所现在做的只是个“开头”,远远没有达到“目的”。我想,没有那么多一批高水平研究人员紧密协作、长期坚持,就不会有汉蒙机器翻译的具有较高实用价值的“产品”。我希望,巴达玛教德斯尔博士在此基础上继续做一系列后续研究,取得更好的一系列成果。

布扎精

于呼和浩特市学府花园

2005年11月





摘 要

在汉蒙两种语言的短语结构之间实现转换,是基于转换的汉蒙机器翻译系统的核心技术之一。本书在借鉴传统蒙古语语法研究成果的基础上,面向汉蒙机器翻译的实际需要,尝试以形式化的方式对汉语短语的蒙古语转换规则进行描述,探讨有关短语转换方面的语法和翻译问题,并提出了自己对有关问题的解决方法。全书共四章。

第一章简明扼要地介绍了以往学者们对本书所关注问题的研究概况、课题研究所要达到的目标以及具体工作步骤等。

第二章提出了一个面向信息处理的蒙古语词语和短语分类体系及标记集,介绍了我们研制的汉蒙机器翻译系统的基本情况和结构。这两项内容,是本课题的理论背景和实践基础。

第三章对汉语的 np、vp、ap 和 pp 四类主要短语的蒙古语转换规则进行了具体而系统的形式化描述。这部分工作可以概括为,将以往面向传统翻译所做的有关汉语这四类短语的蒙译研究的成果,加上本人的研究和实践,组织成一部可以直接服务于汉蒙机器翻译的短语转换规则库。本章介绍了有关上述四类短语的转换规则 100 余条、局部性转换规则 100 余条。

第四章展示转换规则的调试结果,并从汉语分析和蒙古语转换

两个方面对短语转换规则研究中存在的难点问题进行了分析。

汉蒙短语转换规则研究是跨现代蒙古语语法和汉蒙机器翻译两个领域进行的。一方面,从蒙古文信息处理的角度关注蒙古语词语分类和短语研究的问题,可以为该领域的研究工作提供一个借以参考的理论框架;另一方面,短语转换规则研究的具体成果对汉—蒙乃至外—蒙机器翻译的发展具有一定的应用和参考价值。



摘

要



Abstract



To implement the transformation between Chinese and Mongolian phrases is one of the key technologies for the rule-based Chinese-Mongolian machine translation system. On the basis of the research of traditional Mongolian grammar and for the practical requirements of Chinese-Mongolian machine translation, the author tries to describe and formalize the Mongolian transformation rules of Chinese phrases. He discusses the grammar and translation problems that relate the phrase transformation, and presents his solutions for them. The book includes four chapters.

In the first chapter, the author briefly introduces the research concerning the problems discussed in this book. The purpose of the research and the concrete approaches are explained here too.

In the second chapter, the author puts forward an information processing-oriented classification system of Mongolian words and phrases and its token sets. He also introduces the basic state and structure of the Chinese-Mongolian MT system that has been developed. The two items that mentioned above offer a theoretical back-



ground and practical basis for the subject.

The Mongolian transformation rules for the four main types of Chinese phrases of np, vp, ap, and pp, are formalized and described systematically and detailedly in the third chapter. This part of work may be generalized as below: to build up a rule base, which can directly serve Chinese-Mongolian MT system. The rule base is founded on the accomplishments acquired in the traditional translation research, which regards the Mongolian translation of these four types of Chinese phrases, and the results from the author's personal research and practice. In the same chapter, the author introduces 100 transformation rules of the four types of phrases that mentioned above, and another 100 local transformation rules too.

In chapter four, the debugging result of the transformation rules is showed. From the aspects of Chinese analysis and Mongolian transformation, the author analyzes the difficult problems that exist in research of phrase transformation rules.

The research on Chinese-Mongolian transformation rule goes across two fields of contemporary Mongolian grammar and Chinese Mongolian machine translation. On the one hand, to focus on the classification and research of Mongolian phrases from the aspect of Mongolian information processing, can offer a theoretical framework for the field as a reference. On the other hand, the material achievements of the research on the phrase transformation rules would have some application and reference value to the development of Chinese-Mongolian MT system and even to the Foreign—Mongolian MT system too.

关于本书所用符号、代码的说明

1. 汉语词语(语素、词、标点)分类、短语分类标记

代码	含义	代码	含义	代码	含义	代码	含义
a	形容词	p	介词	np	名词短语	dp	副词短语
b	区别词	q	量词	ap	形容词短语	dj	主谓短语
c	连词	r	代词	vp	动词短语	fj	复句短语
d	副词	s	处所词	mcp	数词短语	zj	整句短语
f	方位词	t	时间词	mp	数量短语	npbar	名词性准短语
g	语素	u	助词	tp	时间词短语	vpbar	动词性准短语
m	数词	w	标点符号	sp	处所词短语	apbar	形容词性准短语
n	名词	z	状态词	pp	介词短语		

2. 蒙古语词语分类、短语分类标记

代码	名称	标记的记忆提示
A	形容词	取英语 ADJECTIVE(形容词)的第1个字母。
B	惯用语	取蒙古语 HEBSIMEL (UGE)的第3个字母,因其第1、2个字母已用于情态词、字符。
C	连接词	取英语 CONJUNCTION(连接词)的第1个字母。
D	副词	取英语 ADVERB(副词)的第2个字母,其第1个字母已用于形容词。
E	字符	取蒙古语 TEMDEGTU 的第2个字母,其第1个字母已用于时间词。
F	后缀	取英语 SUFFIX(后缀)的第3个字母,其第1、2个字母已用于语气词、摹拟词。





G	后置词	取蒙古语 DAGABVRI (UGE)的第 3 个字母,其第 1、2 个字母已用于副词、形容词。					
H	情态词	取蒙古语 HANDVLG_A (UGE)的第 1 个字母。					
I	感叹词	英文的其他字母都已经有了归宿,把这个字母用来标记蒙古语最具特色的一个词类。					
L	缩略语	取蒙古语 T0BCILAGSAN (UGE) 的第 6 个字母,前几个字母已用于时间词、方位词、惯用语、连接词、感叹词。					
M	数词	取英语 NUMERAL(数词)的第 3 个字母,N、U 已有他用。					
N	名词	取英语 NOUN(名词)的第 1 个字母。					
O	时位词	取蒙古语 ORON(- V NER_E)的第 1 个字母。					
Q	量词	取英语 QUANTIFIER(量词)的第 1 个字母。					
R	代词	取英语 PRONOUN(代词)的第 2 个字母。					
S	语气词	取蒙古语 SVLA (UGE)的第 1 个字母。					
T	时间词	取英语 TIME(时间)的第 1 个字母。					
U	摹拟词	取蒙古语 DURSULEHU (UGE)的第 2 个字母,第 1 个字母已用于副词。					
V	动词	取英语 VERB(动词)的第 1 个字母。					
W	标点符号						
Y	复合词	取蒙古语 H0RSIY_A (UGE)的第 6 个字母,其前几个字母已由他用。					
AP	形容词短语	GP	后置词短语	MP	数量词短语	MCP	数词短语
NP	名词短语	OP	时位词短语	RP	代词短语	TP	时间词短语
VP	动词短语	DJ	主谓短语				

目 录



面向机器翻译的汉蒙短语转换规则研究

摘要	1
Abstract	1
关于本书所用符号、代码的说明	1
第一章 导论	1
一、课题的提出	1
二、本课题研究的基础	2
三、课题研究的内容和工作步骤	4
第二章 课题研究的理论背景和工程实践	5
一、理论背景	5
二、工程实践	26
第三章 汉语到蒙古语的短语转换规则研究	40
一、关于短语规则形式的几点说明	41
二、np 的转换规则	47
三、vp 的转换规则	61
四、ap 的转换规则	73
五、pp 的转换规则	77



第四章 结论和余论—转换规则实验结果示例及难点分析	100
一、转换结果基本准确的短语的实例	101
二、短语转换不出正确结果的原因分析	109
参考文献	115
攻读学位时发表的论文	125
后记	127



目

录



第一章 导 论

一、课题的提出

本课题的研究工作在跨现代蒙古语语法和汉蒙机器翻译两个领域进行。

在迄今国内外发表的、面向人的、传统的蒙古语语法著作和教材中都或多或少地涉及到了蒙古语短语研究方面的一些问题。

国内的蒙古文信息处理工作,是从 20 世纪 70 年代起步的。以内蒙古大学为例,从 20 世纪 80 年代以来,在蒙古文信息处理研究方面做了很多基础性的工作,还开发了一些应用系统。20 世纪 90 年代由蒙古语文研究所建立的《500 万词级现代蒙古语文数据库》,就是其中的一项基础工程。该所研究人员也承担完成了国家自然科学基金项目《对蒙古语语料库的词类标注与统计》。进入 21 世纪以后,也开始了以短语为对象的切分、标注研究。

在传统蒙古语语法研究中,短语研究是薄弱环节。但是在蒙古文信息处理的各种应用系统中,短语结构规则却起着非常重要的作用。这种“供需矛盾”,使蒙古文信息处理工作的进展受到一定的制约。目前国内外蒙古语语法的研究,无论是面向人的传统语法,还是

以计算机为对象的形式语法,都迫切要求对蒙古语短语的结构规则进行全面、系统地研究。本课题立足汉蒙机器翻译,面向多种领域。通过本课题的研究,我们研制、开发出一个面向汉蒙机器翻译的、从汉语到蒙古语的短语转换规则库,并对在研制过程中遇到的理论和实践方面的有关问题进行探讨。

二、本课题研究的基础

开始本课题研究工作之前,有必要了解以往学者们对本课题所关注的短语结构规则的研究情况。同时,由于我们的研究首先在汉语短语结构分析规则(我们假定这个分析结果是正确的)的基础上,写出对应的蒙古语短语结构转换规则,然后用计算机翻译一定数量的短语或简单句,再根据所得出的结果对转换规则做进一步的改进,所以,也有必要对开展本课题研究所依托的机器翻译系统环境以及选用的语料作一些说明。

一、前人的研究

在过去出版的有关蒙古语语法的论著和高等学校“现代蒙古语”课程的教材中,都有对现代蒙古语词组(与“短语”有些区别)的结构、功能进行研究的部分。总的来说,对词组进行宏观分类的多,发现结构规则的少,对词组结构进行形式化描述的更少。在国内,很少几篇论文研究了某一类短语结构规则的形式化描述问题^①。当然,以前人们也一直在用形式化的方法描述蒙古语短语和句子的结构,只是不太自觉而已。短语(词组)研究的有些理论问题上还存在意见分歧。比如,什么是词组?实词与实词的结合是词组,这方面意见一致,但对实词与虚词的结合算不算是词组,就有不同的意见。还有,在蒙古语词组的结构、功能分类方面的意见也不尽一致。如蒙古语词组中有没有“主谓结构词组”?等。

^①特图克,“现代蒙古语名词词组的基本规则”,《内蒙古民族师范学院学报》,1991,(3);特图克,“蒙古语出发格动词词组”,《蒙古语文》,1992,(12)。在撰写本章的时候,作者也有幸读到了蒙古语文研究所研究生达胡白乙拉的硕士学位论文《现代蒙古语名词短语结构规则研究》。

二、我们的基本想法和有关界定

如上所述,在蒙古语短语研究领域确实有很多值得研究的理论和实践问题。但是,我们在本课题中并不对短语的所有问题都进行研究,而是从汉蒙机器翻译的需要出发,研究汉语名词短语(np)、形容词短语(ap)、动词短语(vp)和介词短语(pp)在汉蒙机器翻译中的转换规则。

为此,我们在这里作以下几个方面的界定:

一、我们所说的短语是从翻译角度说的,被翻译成一个整体或结合体(或结构体)的单位,与过去所说的词组、句子既有相同的一面,又有不同的一面;

二、研究的是汉语到蒙古语的转换规则,所以首先假定,在汉蒙机器翻译系统中的汉语短语分析规则基本上是正确的,对其我们打算介入,对蒙古语短语的分析规则也不进行更多的研究。简单地说,就是研究把一个汉语的短语翻译成蒙古语的时候怎么转换其结构的问题;

三、我们的短语转换规则是面向机器翻译的,所以对它进行的描述是形式化的。翻译系统中的汉语分析规则是汉语语法专家人工总结,并用形式化方法描述的。蒙古语转换规则也应与此对应。

短语转换规则不仅关系到具体技术,更关系到语言学问题,所以在编写转换规则的同时需要解决蒙古语言学的相关理论问题。为解决所遇到的语言学问题,我们在蒙古语言学某些领域进行了前期的探索性研究,在蒙古语的词法和句法的形式化方面做了一些工作。这方面的研究工作集中体现在蒙古语词语分类研究的系列成果《面向信息处理的蒙古语词语分类研究》^①中。

①该系列论文曾在1999年于包头召开的中国蒙古语文学学会第八届年会上宣读,它们是:《关于面向信息处理的蒙古语词语分类及制定其标记集问题》(那顺乌日图);《关于时间词》(白音门德);《关于在信息处理中如何处理蒙古语固定词组的问题》(德·青格乐图);《关于在信息处理中如何处理蒙古语摹拟动词的问题》(扎·义兰);《语文词典中无词类标注词的处理研究》(巴达玛敖德斯尔)等,见:《中国蒙古语文学学会学术论文集》(4),pp. 114~139,呼和浩特:内蒙古人民出版社,2001。