汉、蒙、藏、维分词与词性标注技术发展现状研究

通拉嘎[①]

中央民族大学少数民族语言文学学院 北京 100081　　泉州师范学院 福建 362000

国家语言资源监测与研究中心少数民族语言分中心 北京 100081

E-mail:bolor@163.com

摘 要：中文信息处理用各种统计方法实现精度的提高，而规则的运用更多是在歧义和未登录词的研究上。蒙古语的统计方法发轫于近几年，但更多的是一种理念的进步，仍是以词干加附加成分的规则方法为主，现有 AYIMAG 和 DARHAN 切分和词性标注系统。藏文较多运用规则加统计的方法，也有直接运用 viterbi 算法进行分词的成果，其基于规则的方法是用格助词和词典库进行分词；现有北大计算语言学研究所和国家语言资源监测与研究中心少数民族语言分中心的藏文自动分词系统，不过还未见藏文词性标注系统的实现成果。维吾尔语有基于隐马尔科夫模型的 viterbi 算法与基于 n-gram 模型的尝试，其规则方法仍是词干加附加成分的切分理念，现还未见可用的切分与标注系统。

关键词：分词、词性标注、发展现状

Research on the development of Chinese, Mongolian, Tibetan, Uighur segmentation and POS tagging technology

Tonglaga

Department of Minority Language and Literature Minzu University of China Beijing 100081

Quanzhou Normal University Fujian 362000

National Language Resource Monitoring and Research Center minority languages Branch

E-mail:bolor@163.com

Abstract: Chinese information processing achieved its improvement of precision through various statistical methods, and the use of rules applied mainly in the research of ambiguity and unknown words; Mongolian statistical methods began in the last few years, its approach remains the "stem plus an additional component" method, and the existing achievements including AYIMAG and DARHAN Segmentation and Part of Speech Tagging System; Tibetan adopted the "rules plus statistics" methods, its segmentation also uses VITERBI algorithm directly, whose rule-based method is to use auxiliary word and the dictionary word library. Existing achievements include Tibetan Word Automatic Segmentation System from Peking University Institute of Computational Linguistics and National Language Resource Monitoring and Research Center minority languages Branch. However, no Tibetan speech tagging system has been achieved so far; Uighur makes attempts to utilize VITERBI algorithm based on Hidden Markov Model and N-GRAM based model. Nevertheless, its segmentation approach remains the "stem plus an additional component" concept and no available segmentation and labeling system can be seen so far.

Keywords: Segmentation; POS tagging; Development Status

[①] [作者简介] 通拉嘎（1976-），女（蒙古族），内蒙古科尔沁右翼中旗人，中央民族大学少数民族语言文学学院在读博士、泉州师范学院图书馆馆员，研究方向为计算语言学与应用语言学。

本文获国家自然科学基金项目“基于动态语料库的汉语基本词汇特征曲线跟踪与自动提取研究”的资助，项目号 60663008。

1 前言

词即是能够独立运用的，有意义的，最小的语法单位。只有在词的层面上做到了准确、详尽的分析，才能在语义、语用、句法等层面上达到更高的发展。在信息处理上，分词与词性标注是句法分析、语义分析、语用分析、机器词典等工作的基础。

汉、蒙、藏、维的分词与词性标注技术有许多相通的地方，也有很多不同，本文旨在分析各语言的分词与词性标注技术，为面向信息处理的多语种研究提供帮助。

分词是按照特定的规范，对语言按分词单位进行切分的过程。[1]主要有基于词典的、基于规则的、基于统计的、基于规则和基于统计相结合的等分词方法。

词性标注是给定句子中判定每个词的语法范畴，确定其词性并加以标注的过程。[2]词性标注的前提是确定相关语言的标记集，较之传统语言学的词类，面向信息处理的词类更加具体与详尽，力求覆盖语言中所出现的各种现象。因而确定词性标记集是进行词性标注的基础。主要有基于规则、基于统计的、规则和统计结合的、基于转换的错误驱动的等词性标注方法。

2 中文自动分词和词性标注技术

汉语属于汉藏语系汉语支。汉语是按句连写的，汉字之间无形态变化与变格信息可以表示词的间隙与类别，语序灵活，因而中文信息处理首先需要解决的是中文自动分词问题。

汉语分词与词性标注现已有很多成果。

CDWS 分词系统是我国第一个实用的自动分词系统，目前较典型的分词系统有：中科院计算所 ICTCLAS 分词与词性标注系统、北大计算语言所分词与词性标注系统、清华大学 SEG 分词系统、复旦大学分词与词性标注系统、山西大学的 ABWS 分词系统、哈尔滨工业大学分词系统。

中文信息处理关键问题在于歧义和未登录词的处理上。目前对歧义和未登录词，有两种处理方式：1. 与分词分开处理的方法，专门对某一类歧义或未登录词进行研究，将歧义分为交集型、多义性及混合型歧义字段，将未登录词分为人名、地名、组织机构名、日期、时间、金额、百分比等，分别找寻解决方案。如，交集型歧义识别（孙茂松 1997）[3]、组合型歧义识别（侯敏、孙建军 1996）[4]、中国人名的识别（郑家恒、刘开瑛 1994）[5]、中国地名的识别（谭红叶 2002）[6]、组织机构名的识别（张小衡、王玲玲 1997）[7]。这样处理的结果是虽然在歧义和未登录词的处理上达到一定的切分精度，但整个系统缺乏统一算法，也缺乏统一的评估体系，也无法充分利用上下文信息。2. 将分词与歧义、未登录词的处理一并解决。如白栓虎（1995）的分词与词性标注一体化算法[8]、何燕的未登录词一揽子解决方案（2000）[9]、俞鸿魁将中文分词和命名实体识别有机地结合（2004）[10]，这些方法都有不同程度的效果，但尚未达到应用水平的识别率。也都有不足。

汉语歧义的处理方法有：基于词典的，有联想-回溯算法（刘开瑛 2000）[2]231-246；基于统计的，基于马尔可夫模型的 viterbi 算法、决策树[11]、专家系统[12]、神经网络等[13]；规则与统计相结合的方式，目前很多统计方法都或多或少也应用了一些规则，只是规则与统计的侧重点不同。如：规则与统计结合的新词识别方法[14]。

汉语未登录词的处理方法有：基于规则的方法，如郑家恒、张辉（2002）[15]总结机构名称的组织规则；基于统计的方法，如孙茂松，左正平（1998）[16]的基于词的二元模型互信息与 T-

测试差排歧算法；统计与规则相结合的方法，如：宋柔，朱宏（1993）的基于语料库和规则库的人名识别法[17]。

2003 年,863 中文与接口技术评测组在中科院计算技术研究所对中文信息处理的多个项目进行了评测,杨尔弘，方莹（2006）等对评测结果进行了介绍，分词精确率最高为93.44%,词性标注精确率最高为87.47%,命名实体整体的（人名、地名、机构名）精确率最高为76.45%，[18]歧义和未登录词始终是中文信息处理的必须要闯过的关。

1998 年，俞士汶等建立了《现代汉语语法信息词典详解》[19]，中文信息处理也有《信息处理用汉语分词规范》、《信息处理用现代汉语分词词表》和《信息处理用现代汉语词类标记集规范》等以供参考，这为新时期中文信息处理的繁荣发展做了非常良好的铺垫。中文信息处理将在更加细化的歧义和未登录词识别方面开展更加深入研究。

3 蒙古语词语切分和词性标注技术

蒙古语属阿尔泰语系蒙古语族。蒙古语名词有数、格、领属范畴，形容词有比较级的范畴，数词、代词有格、领属范畴，动词有体、态、式、时、人称以及形动词、副动词等语法范畴。语序是主宾谓，限定词在中心词之前，有较严谨的元音和谐律。书写方式是从上到下，从左至右移行。蒙古语是拼音文字，属于粘着语，具有非常丰富的形态变化，词与词之间可以自然分割，因而不用像汉语进行分词。但蒙古语需要进行词语自身的切分工作。

内蒙古大学与内蒙古师范大学长期致力于蒙古语信息处理。

蒙古语的词语切分，即是将蒙古语词干的词性信息和附加成分的类型信息标注出来。蒙古语的词语切分与词性标注较多采用了规则方法：那顺乌日图(1997)[20]在"蒙古文词根词干词尾的自动切分系统"中第一次提出建立蒙古语形态分析的方案，将蒙古语的近两百种词尾分成三大类自动切分系统，主要采用了编制词典、设计各种语法条件、设计生成规则等方法；淑琴(2005)[21]分析和归纳蒙古语构形附加成分的各种语法属性，建立了构形附加成分的切分与还原规则，设计出易于机器处理的各种属性字段及取值规格；王斯日古楞（2007）[22]通过对蒙古语单词的构成规则进行研究,建立各类单词的词根库和词缀库,让计算机自动识别文本中每个单词的词性;那日松、淑琴（2009）[23]在内建的资料库基础上，进行了词干还原系统设计，并做了形态知识的实验，该测试在歧义与未登录词上仍有很多问题需要解决；胡冠龙，张建，李淼（2007）[24]以基于转换的错误驱动方法对拉丁蒙古文进行词性标注，并针对速度过慢的问题，减小了搜索空间，该篇论文是错误驱动法在蒙古文上的初次尝试。

统计方法在蒙古语中虽然应用较晚，但也有一些进展。图格木勒（2007）[25]设计出基于附加成分和词干的混合切分方法，但蒙古语词类切分系统无法解决歧义问题，因而图格木勒(2007)以基于统计的方法改进了以前的歧义标注的不足，不过未对统计方法展开具体叙述；赵斯琴（2006）[26]介绍了蒙古语词性标注系统的设计思想，采用基于规则和统计相结合的方法,对蒙古语句子进行分类,并对已分类的句子进行词性自动标注。

蒙古语歧义和未登录词的研究性成果仅见少数几篇，那日松，敖其尔（2004）[27]以最大概率和同现概率方法进行兼类词自动标注，准确率分别为75%和81%。富涛、包志红（2008）[28]以记录好的未登录词词干与拟状动词二维表进行比较,认为不断分析和研究蒙古语每个词类的构词特征,归纳出规则,将会识别更多的未登录词。该方法还是与规则法为主，辅以匹配法。阿红(2008)

[28]215-220以互信息和MI值的计算为蒙古文3个典型动词寻找到搭配词，认为利用词类的兼类词搭配的统计和分析方法，可以抽出全部兼类词的显著搭配词，并在此基础上，分析出兼类词消解歧义的规则。该方法虽然引进了概率统计方法，但仍以规则思维为主。

目前,蒙古语大多以基于规则的方法去处理切分与标注,正确率很难达到应用水平。蒙古语仅有少数几篇专门论述歧义与未登录词的文章，没有全面性的研究与突破性进展。面向信息处理的蒙古语各类规范和标记集还未见国家标准，但已有几篇研究成果，那顺乌日图（2007）[29]认为研制出具有较强通用性的蒙古语词语分类及标注规范是当务之急；巴达玛敖德斯尔（2004）[30]提出的词语分类体系包括21个类，并且对分类体系中的15个词类进行了描述；内蒙古大学建立的“面向信息处理的蒙古语词语分类及标记集” 词干标记为77个，附加成分标记为154个。蒙古语将切分与词性标注一体化处理，已有AYIMAG和DARHAN等词语切分和词性标注的系统，那日松、淑琴（2009）也设计了蒙古语词干还原系统，不过正确率也是70%-80之间。蒙古语信息处理应该转变传统方式与思维,更多引入其他方法，胡冠龙，张建，李淼（2007）用基于转换的错误驱动方法进行拉丁蒙古文标注，即是不错的尝试。

4 藏文分词与词性标注技术

藏文属于汉藏语系藏缅语族。词与词之间没有明显的间隙，以格助词来表明各种语法、语义关系。格助词粘着在实词后，不单独使用。词法上藏语的名词、动词、形容词有词形变化，动词还具有时、态、式等屈折变化形态，句法上藏文属于SOV型语言，即谓语动词后置型语言，这是藏语不同于现代汉语的一个明显的特征。藏文书写方式是从左至右，是逻辑格语法体系的拼音文字，属于屈折型语言。

藏文信息处理技术的先行者有北京大学、中国社会科学院、国家语言资源监测与研究中心少数民族语言分中心、西北民族大学、青海师范大学等。

在藏文分词与词性标注中，较多运用了统计与规则相结合的方法：陈玉忠（2003）[31]结合藏文各类形态特征，在分析比较了两种基本分词方法— 最大匹配法和格助词分词法的基础上，提出了一种基于格助词和接续特征的书面藏文自动分词方案；孙媛、罗桑强巴、杨锐、赵小兵（2009）[23]210-218提出利用格分析法将句子分块，在此基础上，再用词典匹配和统计相结合的方法认词，对歧义，采用双向扫描和统计的方法进行识别与切分，对未登录词，采用建立词库和加构词规则的方法及建立统计模型的方法去识别,该方法是串频统计和词形匹配相结合的切分方法；苏俊峰，祁坤钰，本太（2009）[32]分析藏语的构形特征，得到词性标注集，从人工标注的语料中统计词和词性频率以及训练得到二元语法的HMM模型参数，运用Viterbi算法完成词性标注。

基于规则进行分词的研究成果有：祁坤钰（2006）[33]建议三级切分方法：一级是格切分,在单句中查找格助词将单句分割成若干个词群；二级切分，切分对象主要是短语，三级切分的核心任务是词汇验证和再切分。才智杰，索南仁欠（2007）[34]在分析用格助词进行句子分块的基础上，提出了基于格助词和词典库的分词算法。

藏文信息处理技术萌芽于20世纪末,目前较通行的分词技术是以格助词为基础,先将原文切块,再以统计方法分词，是规则与统计相结合的方法。研究分词规范的论著较少，扎西加、卓杰（2009）[35]在借鉴汉语的分词规范的基础上，将藏文词类划分为26个基本词类、9个特殊词类，并阐述了制订分词规范的思路与细则。藏文歧义和未登录词的研究成果较少，仅见孙媛，罗桑强

巴，杨锐，赵小兵（2009）[23]219藏语交集型歧义字段切分方法研究。在藏文词类划分方面，有陈玉忠(2005)[36]的 26 类藏语词语分类体系，扎洛（2006）[37] 的27个大类词性体系等。现已见北京大学计算语言学研究所（正确率96%）和国家语言资源监测与研究中心少数民族语言分中心（正确率93%）的藏文自动分词系统，不过还未见藏文词性标注系统的实现成果。

5 维吾尔语的词语切分与词性标注技术

维吾尔语属于阿尔泰语系突厥语族。维吾尔语的名词有数、格、从属人称等语法范畴，形容词有比较级的变化。代词一般都有格的变化。数词在作名词使用时，有数、人称、格的变化。动词有体、态、时、人称、数以及形动词、副动词等语法范畴。语序是主宾谓，限定词在中心词之前。构词和构形附加成分很丰富，一般是词根在前，构词词尾在中，构形词尾在后，有严格的次序顺序，有较严谨的元音和谐律。维吾尔语属于拼音文字，从右向左书写，是粘着型语言。

新疆大学与新疆师范大学在维吾尔文信息处理方面走在前列。

在维吾尔语词语切分与标注中，统计方法运用较广。毕丽克孜(2003)[38] 以 1 篇维吾尔文中篇小说为语料，探索计算机自动识别和处理维吾尔文语料的途径，阐述与维吾尔语词频统计技术相关的具体步骤与方法；陈鹏（2006）[39]采用了双向匹配和全切分相结合的方法来实现维吾尔语词干提取，首次采用概率统计的方法研究了维吾尔语词性标注问题，从而证明，基于概率统计的一阶隐马尔可夫模型以及 Viterbi 算法能有效的解决维吾尔语词性标注的问题；买合买提. 买卖提（2008）[28]206-209认为基于规则的词性标注成效不太理想，所以用 N-GRAM 模型对维吾尔语进行标注；艾则孜·吐尔逊，买合木提·买卖提（2009）[40]以基于隐马尔可夫模型的 Viterbi 算法对维吾尔文词类包括兼类词进行自动标注。

而基于规则的一些研究成果有：古丽拉. 阿东别克（2004）[41]提出以“词=词根+附加成分”结构进行维语词语切分的一些规律和实现方法；赛麦提·麦麦提明（2006）[42]在介绍词性自动标注系统原理的基础上，初步探讨了兼类和同形而引起的歧义的三种方法，即词语结构分析法、搭配词统计法和分布特点规则法等，是规则法和统计法的结合。

维吾尔语切分与标注技术的研究始于 20 世纪末，标注标记集的研究已有玉素甫·艾白都拉 Version 1.0[43]和 2. 0[44]的报告及新疆大学吐尔根. 依不拉音（2006）关于词性标记集的探索可供参考。维吾尔语词语切分、词性标注技术都处于起步阶段，现有成果有基于统计的解决方式，如陈鹏（2006）的一阶隐马尔可夫模型以及 Viterbi 算法，也有基于规则处理问题的，如古丽拉. 阿东别克（2004）的基于“词=词根+附加成分”结构进行维语词语切分的方法，也有统计与规则相结合的方法，如赛麦提·麦麦提明（2006）的搭配词统计法和分布特点规则法。但目前为止，还未见可用的切分与标注系统，仅见吐尔根·依不拉音等的（2006）[45]（2009）[28]210-214的基于词典和基于规则的词性标注系统设计，但尚未实现。维吾尔语切分与词性标注正从基于规则向基于统计方法的转变。

6 讨论

分词与词性标注技术是语言信息处理技术的基础工程。蒙、藏、维分词与词性标注技术相比中文信息处理起步较晚，收获成果较慢，存在着许多理论和技术上的难题。汉语成功吸收了英语

的先进经验，运用各种统计方法实现精度的提高，而规则的运用更多是在歧义和未登录词的研究上。蒙古语的统计方法发轫于最近几年，但更多的是一种理念的进步，而不是具体哪种算法的实现，还是以词干加附加成分的规则方法实现标注。藏文较多运用的是规则加统计的方法，即先用格助词切块，再用统计方法分词，也有直接运用 viterbi 算法进行分词的成果；基于规则的方法是以格助词切分为基础，用格助词和词典库进行分词。维吾尔语有基于隐马尔科夫模型的 viterbi 算法与基于 n-gram 模型的尝试，其规则的方法也是词干加附加成分的切分理念。少数民族语言信息处理技术应该立足于自身语言，借鉴其他语言的先进成果，并相互借鉴，发展符合本民族语言的信息处理技术。在完整的、穷尽式规则建立耗费巨大人力、物力及时间的情况下，少数民族信息处理技术可以借助大规模语料库，采用统计与规则相结合的方法，实现技术突破。各少数民族语言信息处理要从以往依赖语言学理论转向在大规模文本上的统计分析，从以往依赖人工的长期总结转向机器自动学习.

分词与词性标注需要遵循特定的规范，除了蒙古语有内大自建的《蒙古语语法信息词典详解》外，其他少数民族语言还没有类似成果，更没有面向各少数民族语言的信息处理各类规范与标记集，这导致了有限资源的重复建设与互不兼容，是技术及研究人员的极大浪费，直接制约了少数民族语言信息处理技术的纵深发展。蒙、藏、维语言信息处理技术应该尽快建立完备的面向信息处理的各类规范与标记集，争取确定为国家标准，更加推进少数民族语言的信息化和自动化。

参 考 文 献

[1]吴立德著. 大规模中文文本处理[M]. 上海：复旦大学出版社，1997，16

[2]刘开瑛著. 中文文本自动分词和标注[M]. 北京：商务印书馆，2000,162

[3]孙茂松. 利用汉字二元语法关系解决汉语自动分词中的交集型歧义[J]. 计算机研究与发展,1997(5)332-339

[4]侯敏,孙建军. 汉语自动分词中的歧义问题[J]. 语言文字应用,1996(1)69-72

[5]郑家恒，刘开瑛 汉语姓名自动辨识初探[J]. 语言文字应用，1994(2)65-68

[6] 谭红叶,郑家恒,刘开瑛. 中国地名自动识别系统的设计与实现[J]. 计算机工程，2002(8)128-129

[7]张小衡,王玲玲. 中文机构名称的识别与分析[J]. 中文信息学报,1997(4)22-31

[8]白栓虎. 汉语词切分及词性自动标注一体化方法[J]. 计算语言学进展与应用清华大学出版社，1995

[9]何燕. 任意类型的未登录词的识别研究[J]. 北京语言文化大学硕士学位论文，2000

[10]俞鸿魁. 基于层次隐马尔可夫模型的汉语词法分析和命名实体识别技术[D]. 北京化工大学硕士学位论文，2004

[11]孟迎. 基于统计的机器学习的中文命名实体识别[D]. 昆明理工大学硕士学位论文，2004

[12]何克抗. 徐辉. 孙波. 书面汉语自动分词专家系统设计原理[J]. 中文信息学报. 1991（2）1-14

[13]徐秉铮，詹剑. 基于神经网络的分词方法[J]. 中文信息学报,1993（2）36-44

[14]聂颂,何丕廉. 孙越恒. 统计与规则结合的一种新词识别方法[J]. 微型机与应用,2003（10）

[15]郑家恒 张辉. 基于 HMM 的中国组织机构名自动识别,计算机应用,,2002（11）1-2

[16] 孙茂松,左正平. 汉语真实文本中的交集型切分歧义[C]. 汉语计量与计算研究. 香港：香港城市大学出版社,1998

[17]宋柔，朱宏. 基于语料库和规则库的人名识别法[C]. 计算语言研究与应用. 北京语言学院出版社，1993

[18]杨尔弘，方莹，刘东明，乔羽. 汉语自动分词与词性标注评测[J]. 中文信息学报，2006（1）44-49

[19]俞士汶. 现代汉语语法信息词典详解[M]. 北京：清华大学出版社，1998
[20]那顺乌日图. 蒙古文词根词干词尾自动切分系统[J]. 内蒙古大学学报(人文社会科学版), 1997(2)53-57
[21]淑琴. 《蒙古语语法信息词典构形附加成分分库》的设计与实现 [D]. 呼和浩特:内蒙古大学硕士论文，2005
[22]王斯日古楞. 蒙古语单词词性自动识别研究[J]. 内蒙古师范大学学报自然科学版，2007（3），319-321
[23]那日松. 淑琴. 蒙古文词干还原系统设计和研究中的若干问题[C]. 第十二届全国少数民族语言文字处理学术研讨会论文集, 2009，149-154
[24]胡冠龙，张建，李淼. 改进的基于转换方法的拉丁蒙文词性标注[J]计算机应用, 2007（4）
[25]图格木勒. 蒙古文资源库建设相关技术研究[D]. 内蒙古大学硕士论文，2007
[26]赵斯琴. 蒙古语词性标注系统的设计[J]. 内蒙古师范大学学报(自然科学汉文版)，2006（2），186-188
[27]那日松，敖其尔. 蒙古语兼类词词性标注的处理[J]. 蒙古学集刊, 2004（3）
[28]富涛，包志红. 蒙古语未登录模拟动词识别方法[C]. 第二届全国少数民族自然语言处理学生研讨会, 2008, 191-194
[29]那顺乌日图，淑琴. 面向信息处理的蒙古语规范化研究[J]. 中央民族大学学报(哲学社会科学版)，2007年第6期
[30]巴达玛敖德斯尔. 面向信息处理的蒙古语词语分类体系研究[J]. 中央民族大学学报(哲学社会科学版)，2004(3)93-99
[31]陈玉忠. 基于格助词和接续特征的藏文自动分词方案[J]. 语言文字应用，2003(2)75-82
[32]苏俊峰，祁坤钰，本太. 基于HMM的藏语语料库词性自动标注研究[J]. 西北民族大学学报（自然科学版），2009(2)42-45
[33]祁坤钰. 信息处理用藏文自动分词研究[J]. 西北民族大学学报，2006(4)92-97
[34]才智杰, 索南仁欠. 藏文分词算法研究[C]. 第十一届全国民族语言文字信息学术研讨会论文集，2007
[35]扎西加、卓杰. 面向信息处理的藏文分词规范研究[J]. 中文信息学报，2009(4)113-117
[36]陈玉忠. 信息处理用现代藏语词语的分类方案[C]. 第十届全国少数民族语言文字处理学术研讨会论文集， 2005
[37]扎洛. 语言信息处理的现代藏语词性分类方法研究[J]. 青海师范大学学报, 2006(1)38-41
[38]毕丽克孜. 现代维吾尔语语料库词频统计实验性研究[D]. 新疆大学, 2003 年
[39]陈鹏. 基于语料库的维吾尔语词干提取和词性标注[D]. 新疆大学硕士学位论文, 2006 年
[40]艾则孜·吐尔逊，买合木提·买卖提. 基于隐马尔科夫模型的维吾尔语词性自动标注系统的设计与实现[J]. 和田师范专科学院学报，2009(5)217-218
[41]古丽拉·阿东别克. 维吾尔语词切分方法初探[J] . 中文信息学报, 2004(6)61-65
[42]赛麦提·麦麦提明. 现代维吾尔语同形词词性自动标注探析[J]. 语言与翻译(汉文)，2006(3)35-38
[43]玉素甫·艾白都拉. 现代维语语料库的词类标注研究[J]. 民族语文, 2005(4)
[44]玉素甫·艾白都拉, 阿不都热依木·沙力, 阿拉帕提古丽. 信息处理用维语词汇标注标记集的确定[J]. 计算机应用，2009(7) 2006-2008
[45]吐尔根·依不拉音，阿里甫·库尔班. 基于词典的现代维吾尔语词性自动标注系统的研究[A]. 中文信息处理前沿进展——中国中文信息学会二十五周年学术会议论文集，2006