论蒙古语词素切分的实现*

通拉嘎[1,2]，赵小兵[3]
[1]泉州师范学院 图书馆，福建 泉州 362000
[2]中央民族大学 少数民族语言文学学院，北京 100081
[3]中央民族大学 信息工程学院，北京 100081
E-mail: bolor@163.com

摘　要：词素切分即视词根与附加成分为基本词素，自动识别词根的词性及意义、附加成分类型信息。基于词素的切分能有效处理歧义和未登录词，有效缓解数据稀疏问题，促进语言信息处理深层次发展。目前蒙古文信息处理还尚未进入词素切分层面。论文探讨了实现词素切分的理论和实践基础、面临的难题，认为：实现蒙古语的词素切分需要充分利用定性与定量，规则加统计的方法，先借助语言学的定性研究成果，建立信息处理用词根及附加成分词典，制订《信息处理用现代蒙古语切分规范》，然后以规范为指导，以电子词典为基础，建立词素切分理念的语料库，修改与验证规范及词典，进一步解决词根与附加成分的量化与切分问题，实现词素切分。

关键词：词素；蒙古文信息处理；词根；附加成分

A Study on Mongolian Morpheme Segmentation

Tong Laga[1,2], Zhao Xiaobing[3]
[1] Quanzhou Normal University, Quanzhou 362000
[2] Department of Minority Language and Literature, Minzu University of China, Beijing 100081
[3] College of Information Engineering, Minzu University of China, Beijing 100081
E-mail: bolor@163.com

Abstract: The morpheme segmentation take the root and suffix as basic morpheme, automatically identifying POS and meaning of root, and the type of suffix. Morpheme segmentation can effectively resolve problems of ambiguity and unknown words, effectively alleviate sparse data problem, improve development of Mongolian language information processing.This paper aims at solving the problems of Mongolian morpheme by ways,of rule and statistics. with the help of the achievement of linguistic qualitative research. And will be constituted. dictionary of root and suffix ,will be set up a 《Contemporary Mongolian Language Segmentation Specification for Information Processing》,was being perfected while in use, will be set up morpheme segmentation corpus. and improvement of Specification and dictionary, to solve morpheme segmentation.

Keywords: morpheme segmentation; Mongolian information processing; root; suffix

1　引言

词根、词干、附加成分是蒙古语的语言单位。词根是无法再切分的基本义，是构成新词的基础；词干是可切分的具新意的词，由词根缀接各种附加成分而成；构词附加成分续接在词干后，表示新的词汇义和语法义；构形附加成分接续在词干后，表示语法义。迄今为止，国内的少数民族信息处理都只推进到词干与构形附加成分层，尚未推进到词根与构词附加成分层，而词根与附加成分切分（即切分词素）是语言信息处理深化的前提。蒙古文信息处理目前大致也是以词干与构形附加成分的切分理念为基础，基于内蒙古大学语料库的国内相关研究领域，如蒙古语切分与词性标注、机器翻译也以词干切分理念进行相关研究。论文尝试探讨词素切分对蒙古文信息处理及语言学发展的意义，并探索实验的意义、可行性及难点，尝试提出解决方法。词素是音义结合

* 本文承国家科技支撑计划“藏语 / 维吾尔语语言资源监测关键技术研究与示范应用”（2009BAH41B04）资助。

的最小语言单位，词素切分即视词根、各类附加成分为基本词素，切分与标注出词根的词性及意义、附加成分类型信息。

2 词素切分的研究意义

作为粘着语，词法构词法，即词干后接续构词或构形附加成分构成新词是蒙古语最主要的构词法，因而蒙古语词素切分理念对蒙古文信息处理和语言学研究都有重要意义。

2.1 对蒙古文信息处理及其他少数民族信息处理的意义

日语、韩语、蒙古语、维吾尔语、哈萨克语是粘着语，有非常丰富的形态变化，附加成分一般只表达一种意思或仅具有一种语法功能。英语、俄语是屈折语，也具有十分丰富的形态变化，与粘着语的显著区别是一个附加成分可以表达多个意思。冯志伟早在1996年就已探讨了芬兰语等粘着语、英语、法语、德语等屈折语的词素切分问题。[1]刘颖认为英语的词法分析只有分析到词根层，才能解决歧义、未登录词的问题。[2]日语至少存在3种小于句子的语言单位：词、词素和句节，以什么作分词单位是日语分词遇到的首要问题，日本现有的分词软件绝大多数都是以形态素（词素）为单位进行切分。[3]词素切分是粘着语和屈折语词法分析深层次发展必须面临的问题，但限于理论和技术基础，很多语言尚未推进到这一层面，这其中也包括蒙古文及国内的其他少数民族语言信息处理，不过尝试性研究已逐渐进行，维吾尔语已有研究构词、构形附加成分的详细的词类标记集，也已建立了有2万多条数据的语料库用词根词典。而随着非监督式形态切分方法及词素理念的发展，蒙古文信息处理已有将词根、附加成分视为词素，进行统计分析的论述。

通过词素切分可以掌握词根的词性及基本信息，附加成分的词汇或语法义信息，还可了解词根与附加成分的搭配信息。由于粘着语附加成分一般只有一种意义，其接续有规律可循，因而蒙古文等粘着语也可从附加成分推断词根的词性、词义信息，判断与词根的搭配信息，这也十分有益于了解语义及搭配。词素切分方法与词干加构形附加成分的切分方法有明显区别，用词素切分理念切分“BICILGE”,需切分成“BICI+LGE”，“BICI”是词根，缀接“LGE”构词附加成分后动词变成了名词，“写”意变为“写法”意，词素切分方法可显示意义及词性的变化，而词干与附加成分的切分理念无法切分“BICILGE”，自然也无法通过切分来显示词性和词义的变化。

维、哈、柯、朝是典型的粘着语，词素切分将带动蒙古文信息处理更深层次的发展，也必将给维、哈、柯、朝等国内其他少数民族的语言信息处理带来积极、深远的影响。蒙古语词素切分理念能有效处理歧义和未登录词，还可缓解数据稀疏问题，从而对信息检索、词典编撰、词频统计、词法分析、句法分析、语义分析、机器翻译的研究有很大促进。

歧义和未登录词是自然语言处理遇到的两大问题。

不论规则方法还是统计方法，蒙古语都需借助词根的词性及词义信息、附加成分的类型特征，上下文语境信息来解决歧义及未登录词问题，如：TAHIY_A,词根为“TAHI”，既有名词“鸡”的意思，也有及物动词“祭祀”的意思；JABSARLAG_A,词根为“JABSAR”，既有名词“隔阂、间休”意，也有及物动词“隔开”意，[4]FL0RTV=NAtRI（氟化钠）、DORBELJI=CILAGV（方解石）是未登录词，词根分别为“FL0R=NAtRI”，“DORBE=CILAGV”，词素切分方法可以切分出词根的词性信息、附加成分类型信息，并通过统计方法了解词语搭配、语境信息，十分有助于理解并切分歧义和未登录词。词干与构形附加成分的切分理念颗粒度较大，无法深入到基本词素，因而难以解决歧义、未登录词、语义问题。

对机器翻译来说，词法分析是句法分析及后续研究的基础，只有更完善详尽的词法分析，才能更好为句法分析及语义分析提供服务。数据稀疏是统计机器翻译面临的大问题，尤其在少数民族语言信息处理领域，熟语料的缺乏及从而导致的数据稀疏是基本问题之一。词素切分理念能更

充分利用语言自身的形态信息，使词根和附加成分的出现次数得到显著提高，使训练更加充分，进而缓解数据稀疏问题。

2.2 对语言学研究的意义

词素切分理念先基于蒙古语言学研究成果，提出定性假设，然后通过语料库的定量研究方法验证假设，促进蒙古语言的教学与研究。因而蒙古文信息处理的词素切分理念能为词根及附加成分、词典学、语义学、词法、句法研究提供量化内容，促进语言学发展。

3 词素切分的理论与实践基础

词素切分理念需要成熟的语言学及信息处理理论、技术路径、大量的数据资源作理论与实践基础。这里将词素切分基础分为理论基础、数据资源、技术上的实现路径进行分析。

3.1 理论基础

3.1.1 信息处理基础

自然语言信息处理发展多年，汉、英、日、藏、维及其他粘着语的词法分析与词性标注、句法理论与句法分析、语义理论和语义分析、机器翻译及语料库语言学研究、规范和标记集研究成果都是我们的理论基础。蒙古文信息处理发展较快，已有 5 种以上的切分与词性标注系统，短语分析、句法分析、语义分析、机器翻译也获得较大突破，内蒙古大学已建立了《蒙古语语法信息词典》，并发布了完善的《面向信息处理的蒙古语标记集》。附加成分的研究可见斯琴著《现代蒙古书面语构词附加成分研究》、淑琴著《〈蒙古语语法信息词典构形附加成分分库〉的设计与实现》，语义及同形词研究可见额尔敦朝鲁著《面向信息处理的蒙古语动词语义研究》、德·萨日娜，那顺乌日图著《〈蒙古语语义信息词典〉的初步构建》、淑琴著《蒙古文同形词知识库的构建》、哈斯格日乐著《面向信息处理的蒙文同类同音同形词自动识别研究》、齐红莲著《〈蒙古语多义词数据库〉建设》…等成果。

3.1.2 语言学理论基础

词素切分理念还需以语言学作定性研究基础。蒙古语已有清格尔泰著《蒙古语语法》、确精扎布著的《蒙古文编码》、沙.罗布桑旺丹著《蒙古语法》等十分成熟的理论著作，还有、陈乃雄著《蒙文同形词》、达.巴特尔编著《蒙古语派生词倒序词典》、内蒙古大学蒙古语文研究所编《蒙汉词典》等词典做支撑，词根与附加成分的研究可见斯琴朝克图编《蒙古语词根词典》、宝玉著《现代蒙古语常用词缀详解教程》、诺尔金著《蒙古语构词后缀汇总》、特格希都楞著《蒙古语构词法研究》等专著或词典，这些为词素切分奠定了雄厚的语言学基础。

3.2 数据资源

蒙古语语料库的建设从 20 世纪 80 年代开始，现已有中世纪蒙古语语料库、100 万词级的详细切分与标注的现代蒙古语语料库、1000 万词级的未标注语料库等单语语料库，还有汉蒙平行语料库、英蒙双语语料库及英汉蒙三语语料库等多语语料库，不过这些数据资源都基于词干与构形附加成分的切分理念，还尚无以词素切分理念为基础的语料库。

3.3 技术上的实现路径

蒙古语是拼音文字，粘着语，所以英语和中文信息处理的先进经验只能借鉴，无法移植，技术上的实现路径需要从语言特点出发进行探索。蒙古文信息处理现已有：多种文字处理系统及输

入法、基于规则和基于统计的切分与词性标注系统、达尔罕电子词典、《蒙古语语法信息词典》及管理平台、语料库加工工具及管理平台、语义标注技术、多义词、同形词自动识别、汉蒙词语对齐系统、汉蒙机器翻译系统、蒙英机器翻译系统等技术成果，这些技术和经验是蒙古语词素切分理念的技术基础。

4 词素切分难点及解决

词素切分虽然有重大研究意义，但目前理论和技术基础薄弱，实践起来相当困难，现剖析这些难题，并尝试提出解决方法。

4.1 《信息处理用现代蒙古语切分规范》的建立

切分规范用来指导信息处理发展，促进语语料库的兼容与共享。目前《信息处理用现代蒙古语切分规范》尚未建立，这也制约了词素切分的实现。蒙古文信息处理需要突破语言学研究存在的难点及盲点，根据现代蒙古语的特点及规律，建立以信息处理用为目的，制订完备、实用的现代蒙古语切分规范。该规范应根据《面向信息处理的蒙古语标记集》，对蒙古语基本词类及比词大的语言单位（如惯用语、成语、固定词等）、比词小的语言单位（如附加成分、字母、数字、标点符号等）、同于词的复合词等语言单位制订明确的切分规范，并随语言信息处理与语言学的发展不断修正。

4.2 语料库构建

蒙古语现有的语料库都是以词干与构形附加成分理念为基础，还尚未有基于词素（词根与附加成分）切分理念的语料库，因而实现词素切分必须建立相应语料库。该语料库的建立首先需要量化词根及附加成分，以语言学的定性研究形成假设，然后以定量研究方法加工语料库，在语料库上验证切分理论，进一步修改语料库。

4.3 词根与附加成分的量化与切分还原问题

蒙古语的词根可分为死词根（MUHUGSEN=IJAGVR）、独立词根、非独立词根等3类，独立词根较好处理，死词根及非独立词根需借助附加成分才明确意义，在句中无法独立使用，所以其切分与还原是词素切分面临的难题。附加成分可分为构词、构形、构词-构形等3类，构词附加成分添加词汇义，构形附加成分添加语法义，构词-构形附加成分不仅添加词汇义，还添加语法义。蒙古语的粘着特性导致附加成分层层缀接，所以附加成分的边界问题很难解决，量化自然也成问题。我们需要突破语言学的定性研究思维，以各种语法论著为基础，穷尽列举备选词根及附加成分，然后以是否保持基本意义为确认词根的标准，以不同的添加意义为确认附加成分的标准，辅之量化指标，建立面向信息处理的词根词典及附加成分词典，并在语料库上实践切分。对词根和附加成分应采取从易到难的量化顺序，先解决独立词根，再解决死词根和非独立词根；先解决构形附加成分，再解决构词-构形附加成分，最后解决构词附加成分。

词根与附加成分的切分还原涉及增音、减音、同化、异化等语言现象，需要逐一识别并进行切分标注，如“BERIYED”,词根为“BERI”,缀接构形附加成分“D”时，需增加连接字母“E”,而“BARIGVL”，词根为“BARI”,缀接附加成分“GVR”时，词根的“R”遇到附加成分的“R”，其附加成分被异化成“GVL”。词根与附加成分的切分还原比较复杂，特殊规则较多，需要不断积累与总结。

4.4 词根与附加成分的歧义

不论哪种语言信息处理，歧义始终是影响精度的两大难点之一,如“0NDVRVLG_A”,词根为

“0NDVR”，既有名词“生计”义，也有名词“喷涌”义，如何让机器了解并正确切分？附加成分歧义指附加成分有两个或两个以上的类型信息，如“GCI”可以构成形动词、形容词、名词、数词，“GSAN”可以构成形动词、名词，那么在切分当中如何辨识附加成分歧义的类型？这就需要用规则加统计的方法，首先对词根及附加成分进行量化，收集词根的词性、词语搭配、上下文信息，收集附加成分的类型信息、与词根的搭配信息，之后统计词根与附加成分的搭配频率、互信息及Z分值，结合定性及定量研究方法，解决歧义。

4.5 复合词根词切分

由两个或两个以上独立的词根结合而成的词即是复合词根词，语言每时每刻都在变化，语言学上认定的复合词根词有些早已演变成普通的单词根词，而有些词从蒙古文信息处理角度来看，是根本没有切分必要的，所以确定复合词根词也是很难解决的问题，如 D00GARHV,应切分为“DAGV=GAR+HV”还是“D00=GAR+HV” ？“ABCIRAHV”，应切分为“AB+CV=IR_E+HU”，还是“AB+CIRA+HV”?这些词是否是复合词根词?

我们需要对所有复合词根词进行统计，然后从蒙古文信息处理出发，将人名（如：TEMURBAGAN_A）、地名（如：HOHENAGVR)、机构名（如：VLAGAN=MOCIR）、名词术语（如：VSVTURUGCI）、外来词语（如：KIL0MetR）、有明确的附加成分形式的词（如：0D0HI）剔除，经过上述筛选后，对备选复合词根词再结合定性与定量的方法进行研究，得到最终的蒙古语复合词根词，并提出切分方案。

4.6 复合附加成分切分

复合附加成分指两个或两个以上的附加成分接续在一起，共同表明一种意义，切分后，还分别表示不同的意义，如“CILA(E)”,不切分时是动词附加成分，切分后，“CI”是名词附加成分，“LA(E)”是动词附加成分。此类问题会导致遇到部分附加成分时不知何时断开，何时接续，语言使用环境在此较为重要。作者认为，解决该问题的最有效的方法是基于语料库，对复合附加成分加入定量信息，确定具体环境下的切分。

综上所述，词素切分虽然对少数民族语言信息处理及语言学发展有重大深远的意义，但目前相关研究基础较为薄弱，完成难度很大。我们需借助语言学的定性研究成果，建立信息处理用词根及附加成分词典，制订《信息处理用现代蒙古语切分规范》，然后以规范为指导，以电子词典为基础，建立词素切分理念的语料库，修改与验证规范及词典，解决词根与附加成分的量化与切分问题，充分利用定性加定量，规则加统计的方法来解决词素切分的难题。

本文符号说明

等号（=）为连接符号，表示被连接的是一个切分单位，如：“HEDUI=CINEGEN ”

下划线（_）表示词根或词干自身的分写附加成分，如：“JIRGVG_A”。

加号（+）表示是与词根或词干连写的附加成分，如：“ERGI+GUR”。

参 考 文 献

[1] 冯志伟. 自然语言的计算机处理[M]. 上海：上海外语教育出版社，1996，64-80.

[2] 刘颖. 计算语言学[M]. 北京：清华大学出版社，2002.

[3] 隋福民. 面向机器翻译的日语形态素解析[D]. 大连理工大学硕士学位论文，2004.

[4] 陈乃雄. 蒙古语同形词词典[M]. 呼和浩特：内蒙古教育出版社，1982.

[5] 清格尔泰. 蒙古语语法[M]. 呼和浩特：内蒙古人民出版社，1992.

[6] 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.

[7] 那顺乌日图. 蒙古语语法信息词典框架设计[D]. 内蒙古大学博士学位论文，2000.
[8] 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，2007(6): 115-122.
[9] 李文，张建，李淼. 一种带权值参数的非监督式形态切分方法[A]. 第三届全国少数民族青年自然语言处理学术研讨会论文集[C]. 乌鲁木齐：新疆大学, 2010，30-33.
[10] 阿里甫·库尔班，吾买尔江·库尔班，吐尔根·伊布拉音. 信息处理维吾尔语词语分类体系及标记研究(I)[J]. 新疆大学学报(自然科学版)，2009(4)：476-481.
[11] 阿里甫·库尔班，吾买尔江·库尔班，吐尔根·伊布拉音. 信息处理维吾尔语词语分类体系及标记研究(II)[J]. 新疆大学学报(自然科学版)，2010(1)：106-112.
[12] 百顺. 基于派生文法的日-蒙动词短语机器翻译研究[J]. 中文信息学报，2008(2)：47-54.
[13] 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]. 内蒙古大学硕士学位论文，2005.
[14] 斯琴. 现代蒙古语书面语构词附加成分研究[M]. 呼和浩特：内蒙古教育出版社，2004.
[15]（英）米特科夫. 牛津计算语言学手册[M]. 北京：外语教育与研究出版社，2009.
[16] 张晨，祁坤钰. 基于互信息的词汇搭配研究方法[J]. 西北民族大学学报，2009(3)：57-59.