《现代维吾尔语语法信息词典》数据库建设的研究*

加米拉·吾守尔[1,2], 瓦依提·阿布力孜[1,2], 吐尔根·依布拉音[1,2]

[1]新疆大学 信息科学与工程学院，新疆 乌鲁木齐 830046

[2]新疆多语种信息技术重点实验室，新疆 乌鲁木齐 830046

E-mail: jamila@sina.cn

摘 要：《现代维吾尔语语法信息词典》是为实现维吾尔语的自动分析与自动生成而研制的一部机器词典。是维吾尔文信息处理的支柱工程。在自动分析、自动生成、机器翻译、自动标注、自动校对等工作中语法信息词典所起的作用都是通过语法属性字段及其取值所含信息得以实现。本文从计算语言学的角度着重讨论《现代维吾尔语语法信息词典》数据库设计与实现中遇到的若干问题及解决得的基本方法。

关键词：现代维吾尔语；语法信息；电子词典；数据库

Research on Database Construction of Modern Uyghur Grammar Information Dictionary

Jiamila Hoxur[1,2], Wayit Abliz[1,2], Turgun Ibrahim[1,2]

[1] College of Information Science & Engineering, Xinjiang University, Urumqi 830046

[2] Xinjiang Laboratory of Multi-language Information Technology, Urumqi 830046

E-mail: jamila@sina.cn

Abstract: "Grammar of Modern Uyghur Information Dictionary" is a machine dictionary developed for realize the automatic analysis of Uyghur language, is a backbone tools of Uyghur information processing projects. In automatic analysis, automatic generation, machine translation, automatic tagging, automatic correction, and other work implementing Its role by various syntax properties and contained information .This article from the perspective of computational linguistics focus on discussion of the "Grammar of Modern Uyghur Information Dictionary" database in the design and implementation of basic methods of problems encountered and resolved.

Keywords: Modern Uyghur; grammatical information; electronic dictionary; database construction;

1 引言

作为维吾尔语基本语言知识库的《现代维吾尔语语法信息词典》为语言信息处理提供基础资源。它包括维吾尔语词法形态、句法功能、搭配特征以及正字法等方面的知识，是维吾尔文信息处理的重要基础。自动分析、自动生成、机器翻译、自动标注、自动校对等应用系统都可以从中提取所需要的维吾尔语语法知识。《现代维吾尔语语法信息词典》是面向维吾尔语信息处理的基本语言知识库，其开发《现代汉语语法信息词典》的初衷是为语言信息处理提供基础资源。词典的构建对维吾尔语文本检索、校对、翻译、摘要，乃至让计算机"理解"语言，以及语言知识的获取、表示与运用，对维吾尔语从计算语言学的角度深入研究有深远的理论意义和实用价值。

本文采用计算语言学、语料库语言学以及自然语言处理的方法，在现代维吾尔语词法、句法进行分析的基础上，深入研究维吾尔语的名词、动词、形容词等词类语法特性，建立现代维吾尔语词语分类体系。以此分类体系为指导，采用关系数据库技术设计《现代维吾尔语语法信息词典》结构，根据语法功能和义项相结合的原则，从现代维吾尔语平衡语料库中遴选词语导入《现代维吾尔语语法信息词典》中，并对每个词语赋予其语法属性，最终构建具有实用价值的《现代维吾尔语语法信息词典》数据库。

* 基金资助：国家社会科学基金重点项目[10AYY006]，国家自然科学基金项目[60663006]，国家工信部电子发展基金项目[工信部财（2009）453]的资助

2 语料整理

目前维吾尔语语料库的主要数据来源有非统一编码的各种维吾尔文文字处理系统的排版文件（如：方正，潍坊，Ilikyurt，Alkatip 等）、网站或各种非标准电子资料(2006 之前基本上非标准和扩展区)。这些数据要转换成标准 Unicode 代码后才能使用。主要解决的问题之一怎样将不同编码的语料转换成统一的编码？这要对编码特征进行分析：编码转换时，系统先要识别出什么编码生成的文件，根据识别出的编码和编码特征进行标准 Unicode 编码的转换，例如：方正排版系统中每个字符前有 FA 标记，潍坊排版系统中每个字符前有 8081 标记。根据这些特征我们将它们转换成标准的 Unicode 编码，实现标准化。代码转换的质量直接关系到后续工作的数据质量。目前我们涉及到的系统能转换的编码类型有：方正、潍坊、标准 Unicode 扩展区代码（Alkatip, Ilikyurt)。编码转换后又要解决排版命令的清理、数字序列的还原(如：把 0102 年 01 月 72 日 0012 还原成 2010 年 10 月 27 日 2100)、特殊符号和乱码的清理。为此开发了代码转换引擎。代码转换操作自动完成后，要用新疆多语种信息技术重点实验室设计开发的维吾尔文校对引擎自动进行拼写校对，不能自动转换部分人工完成。

词典基础语料库的收集、整理和统一的质量要求高，因此，首先必须要有高质量的文本语料库，我们选择的语料文件主要有：政府文献、新闻、法律类，文学作品类和医学类词语数量极少，需要不断地充实语料库内容。

3 维吾尔语词性标注集

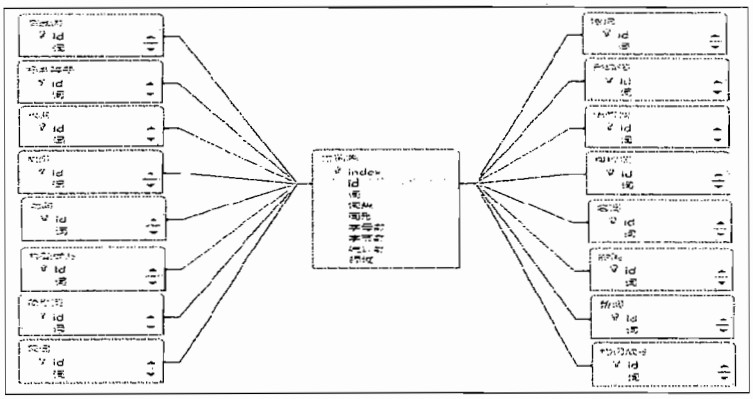
维吾尔语语法信息词典中，词语初步被分为 15 类：名词、动词、代词、连词、量词、数词、形容词、副词、语气词、简称词、模拟词等。语法信息词典的数据库设计时，数据库要给词典提供所有词性，根据维吾尔语语法标注集中的 15 种词类，本文设计的数据库将使用的维吾尔语词性标注集是新疆大学制定的维吾尔语信息处理用词类标记集。它由维吾尔语基本词类标记集、附加成分标记集以及复合标记集等三个部分组成。与传统词性一样包括名词、形容词等基本语法词类。将词类分为一级、二级、三级三个级别的词类。一级标记有 15 类（名词、动词等 12 个基本词类加上标点符号、单词型附加成分、拉丁文等三个特殊类）；二级标记有 71 类（对于一级标记的细分）；三级标记有 51 类（对于二级标记的细分）并且表示由两个或两个以上单词构成的短语型单词标注符号两个，一共有 139 个词类。

4 数据库设计

数据库包括总库、名词分库、动词分库、形容词分库、副词分库、代词分库、数词分库、时位词分库、后缀词分库、连词分库、构形附加成分分库等。对应每个库建立一个数据表，共建立了 17 个数据表。

根据使用对象有个人版和网络版，个人版用 Access 作为数据库管理平台，网络版则使用 SQL Server 2005。基础词性表相同，个人是网络版的导出结果，主要是已经标注好的使用频率较高的维吾尔文词语。用于个人基本查询、统计、分析时使用。通过网络版进行参数调整可以设置个人版的数据范围和类型。网络版具有更复杂和全面的查询、比较、统计、处理功能。设计时还考虑处理其他兄弟民族语言的语法信息(目前是哈萨克、柯尔克孜)和区分自动标注和人工标注(主要用于测试自动标注的准确性)。具有较完善的用户管理、任务管理、权限管理、文档管理、词条管理、语法属性管理、数据备份何还原等功能。为了减少上层管理程序的开发和维护周期，提高响应速度，减少开发语言的依赖性和提高数据安全性将大量的基础功能通过存储过程、触发器、自定义函数在数据库内嵌入式完成。目前已包括总词条(已经标注和未标注的)1200 万，其中没有重复词条约 29 万。

总库是语法信息词典的核心，其中收录了常用维吾尔语单词词干 11 万，设置了 9 个通用属性字段。各个分库的基本字段(包括编号、维吾尔语词条、拉丁词条、词类等)来自总库，其他内容根据各类词的语法特征分别设置。



词典主词表（总库）与分表（分库）关系图

目前，已经完成分类和属性设置的是：有 115938 余词条和 9 个属性字段的《总库》；55918 个词条、14 个属性字段的《名词分库》；48199 个词条、18 个属性字段的《动词分库》；6462 个词条、15 个属性字段的《形容词分库》；1806 个词条、16 个属性字段的《数词分库》1261 个词条、16 个属性字段的《代词分库》；787 个词条、16 个属性字段的《副词分库》；575 个词条、26 个属性字段的《量词分库》；270 个词条、18 个属性字段的《语气词分库》；228 个词条、16 个属性字段的《模拟词分库》；140 个词条、15 个属性字段的《后置词分库》；76 个词条、15 个属性字段的《连词分库》；68 个词条、12 个属性字段的《构词附加成分分库》；186 个词条、15 个属性字段的《叹词》；30 个词条、15 个属性字段的《标点符号》等。

5 结语

维吾尔语语法信息词典目前已经初步实现了总库、名词、动词、形容词、数词、代词、副词、量词、语气词、模拟词、后置词、连词和标构形附加成分分库的构建。随着分库数量的增加，如、简称词、标点符号、构形附加成分等分库，同时根据用户的实际需求，将继续完善和扩大词典功能，如属性字段的扩充，如拉丁字符维吾尔文对应词条字段、汉译词条字段以及句法、语义等方面的属性字段等。

参 考 文 献

[1] 梅家驹, 竺一鸣, 高蕴琦,殷鸿翔. 同义词词林. 上海辞书出版社, 1996.

[2] 阿比达• 吾买尔,吐尔根• 依布拉音. 维吾尔语句子边界识别的研究与实现[J]. 新疆大学学报. 2008 年 5 月：P1-4.

[3] 俞士汶, 朱学锋等. 现代汉语语法信息词典详解. 第二版. 清华大学出版社, 2003 年 2 月: P19-26.

[4] 刘珉. 汉维共时对比语法. 新疆人民出版社, 1991 年 9 月: P9-17.

[5] 力提甫• 托乎提. 电脑处理维吾尔语语音和谐律的可能性[A]. 中央民族大学学报，2004 年第五期.

[6] 阿依克孜• 卡德尔, 开沙尔• 卡德尔, 吐尔根• 依布拉音.面向自然语言信息处理的维吾尔语名词形态分析研究[J]. 中文信息学报, 2006, (3): P43-48.

[7] 哈密提• 铁木尔. 现代维吾尔语语法[M]. 北京民族出版社, 1987 年.

[8] 段慧明, 松井久仁於, 徐国伟, 胡国昕, 俞士汶. 大规模汉语标注语料库的制作与使用. 语言文字应用, 2000 年 2 月: P72-77.

[9] 俞士汶, 段慧明, 朱学锋, 孙斌. 北京大学现代汉语语料库基本加工规范.《中文信息学报》, 2002 年第 16 卷第 5 期: 49~64 和第 6 期: P58-65.