哈萨克语通用词汇自动提取方法研究与实现*

王雅莉，古丽拉·阿东别克
新疆大学 信息科学与工程学院，乌鲁木齐 830046
E-mail: wyl870915@gmail.com

摘　要：以哈萨克语通用词汇自动提取为目标，实现了哈萨克语词汇通用度统计系统。主要介绍了哈萨克语通用词汇自动提取技术，基于通用词汇的三大特征：领域通用性、地域通用性、时间通用性，采用统计的方法考察哈萨克语词汇的通用程度，在哈萨克语词频统计的基础上实现了哈萨克语词汇的通用度统计，根据词语通用度 OK 值提取哈萨克语通用词汇。实验结果表明此方法可行有效。

关键词：通用词汇；哈萨克语；词汇通用度；领域通用度；时间通用度

Study of Automatic Extraction Methods and Implement of Kazakh Common-used Words

Wang Ya-li, Gulila·Altenbek
College of Information Science and Engineering, Xinjiang University, Urumqi 830046
E-mail: wyl870915@gmail.com

Abstract: With automatic extraction of Kazakh Common-used words for the goal, implement the statistical system of Kazakh lexical general degrees. Mainly introduces automatic extraction technique of Kazakh Common-used words. Based on the three properties of Common-used words: Filed generality, Regional generality, Time generality; use statistical methods to investigate the general degree of Kazakh words. On the basis of frequency statistics of Kazakh words, implement the statistics of Kazakh lexical general degrees. Extract Kazakh Common-used words in terms of OK value of lexical general degrees. Experimental results show that the method is feasible.

Keywords: common-used words; Kazakh; lexical general degrees; filed general degrees; time general degrees

1　前言

语言是人类进化的产物，是信息的高级载体，是思想与交际的工具[7]。从一个民族的语言系统来说，词汇是承载语言信息的基本载体，它是语言系统中最活跃、最具生命力的元素[2]。人类社会正在从工业社会迈向信息社会，信息的主要载体是自然语言。自然语言研究如何让计算机理解人类语言并开发有关的适用系统，然而作为自然语言当中的通用词汇是一个民族的语言系统中最常见、使用频率高的那些词汇，在某一时段内，通用词汇是一个相对稳定而又开放的集合[8]。

新疆是多民族的地区，少数民族占总人口的 60%，哈萨克语是仅次于维吾尔语的通用的 6 种少数民族语言文字之一，而且是跨境语言(哈萨克斯坦)[7]。近年来对哈萨克语文学语言和方言的多方研究，对哈萨克语的使用与规范、文学语言的推广和普及起到了积极的作用。

随着社会的持续发展、科学技术的不断进步，尤其是计算机技术的大力推广和运用对少数民族语言的研究和应用创造了良好的人文和科技环境，同时对传统的语言研究提出了许多新的要求。哈萨克语有着丰富的词汇，哈萨克语通用词汇的研究是顺应社会发展，对大量不同语体的书面语和口语材料进行统计分析，运用计算机技术对哈萨克语词汇进行科学研究，得到哈萨克语通用词汇。如果对哈萨克语的统计分析采用手工的方式进行，其工作量是惊人的，正确率也是难以保证的。

* 本文承国家自然科学基金（No.60763005）和国家教育部、国家语委民族语言文字规范标准建设及信息化科研项目(No.MZ115-92)的资助。

本文根据通用词汇的三大特征，即领域通用性、时间通用性和地域通用性，用量化的“词汇通用度”指标来表示词汇通用的程度，用统计的方法考察大众媒体所使用哈萨克语词汇的通用程度，进而实现基于哈萨克语主流报纸媒体语料上的哈萨克语通用词汇自动提取的目标。

2 哈萨克语通用词汇的界定及其特征

2.1 哈萨克语通用词汇的界定

通用词汇的词是一个民族的人民日常都在使用、不容易变化、比较稳固的词语。通用词汇中的词是语言词汇的核心，它们表达的是与人们世世代代的日常生活关系非常密切的事物，如自然现象，家畜、亲属名称，人的肢体、器官名称，表示方位、时令、数目、劳动工具等词汇；从古代部落到现代哈萨克民族，哈萨克人一直从事畜牧业生产，过着游牧的生活，因此哈萨克语的通用词汇中还包含着大量有关畜牧业的词汇。

通用词汇作为哈萨克语词汇中重要的组成部分，并非很容易就能从整个词汇成员中划分出来。简单地给出几个难以精确衡量的标准对词汇成员做硬性划分既不科学，也没必要，同时还会遇到难以解决的问题。因此，迄今为止，尚没有一个科学量化的哈萨克语通用词汇衡量标准。

不可否认，通用词汇在具有历史稳固性的同时，也是在缓慢变化的。然而这并不妨碍我们在哈萨克语语料库基础上研究哈萨克语词汇的全民常用性、时间稳定性及构词能力等特征，考察在大众媒体中词汇的真实使用状况，为哈萨克语通用词汇的自动提取提供依据。

为此，参考汉语通用词汇的界定[2]，对哈萨克语通用词汇进行如下定义：哈萨克族人民在日常交流的词汇使用中，主要涉及的那些使用频率高、在各领域内、各地区间及各时间段中通用程度高的词汇。

2.2 哈萨克语通用词汇的特征

根据哈萨克语通用词汇的定义，可以得出通用词汇具有全民通用性特征，主要表现在如下三个方面：

（1）领域通用性：通用词汇具有在各领域、各行业普遍使用的特性；

（2）地域通用性：通用词汇具有在使用哈萨克语进行交流的不同地域的人们普遍使用的特性；

（3）时间通用性：通用词汇具有时间上使用稳定的特性。

基于通用词汇的上述三大特性，我们用“词汇通用度”这一指标来量化的描述这些特性，用统计的方法考察大众媒体所使用的哈萨克语词汇的通用程度，以实现基于哈萨克语主流报纸媒体语料上的哈萨克语通用词汇自动提取的目标。

3 哈萨克语通用词汇自动提取方法研究

如前所述，哈萨克语通用词汇具有领域通用性、时间通用性及地域通用性特征，词语的通用程度可以分别用与这些特征相对应的领域通用度、时间通用度及地域通用度量化指标来衡量。

3.1 领域通用度的定量分析方法研究

词语领域通用度是用来衡量词语在语言各流通领域的通用程度，即词语常用程度的量化指标。其计算公式不仅应该考察词汇的词频，同时还应该考虑词语在不同文本及不同领域和分领域的分布是否均匀。主要包括定性及定量两种考察方式：

（1）定性考察：依靠领域词汇相交，获得各领域及所有领域的共用词汇。

（2）定量考察：根据词汇在各领域出现的词语频度及词语分布是否均匀等情况，计算词汇的领域通用度。

定性考察方式相对比较简单，对词语的领域分布情况也是一目了然，其缺点是忽略了衡量词语常用程度的重要指标——词频，所以这种考察方式只是作为辅助方式提供参考，本文对词语领域通用度的分析将采用定量计算的方式。

领域通用度计算步骤如下：

（1）计算领域类词语频度 F_k

F_k 为 k 号词语在领域类语料中出现的总频次。

（2）计算 k 号词语文本使用度 UL_k

采用 A.Juilland 公式计算词语的文本使用度：

$$S_k = \sqrt{\sum_{i=1}^{n}(N_k^i - N_k)^2 / n} \qquad D_k = 1 - S_k /(N_k \times (n-1)^{\frac{1}{2}})$$

词的文本使用度 $UL_k=D_k \times F_k$。（取整数值）

其中，N_k^i 表示 k 号词在第 i 类领域中出现的相对频度，N_k 表示 k 号词在所有类中出现的平均相对频度，n 是语料的文本总数，D_k 表示 k 号词的散布系数，F_k 表示 k 号词的词频。

（3）计算 k 号词语的领域通用度 U_k

采用分布均匀度(Distributional Consistency，英文简称 DC)计算词语在各领域类分布的均匀程度，计算公式为：

分布均匀度 DC_k =SMR/Mean。（$0 \leqslant DC_k \leqslant 1$）

SMR 和 Mean 的定义分别如下：

$$SMR = (\sum_{i=1}^{n}\sqrt{FK_i} / n)^2 \qquad Mean = (\sum_{i=1}^{n} Fk_i / n)$$

K 号词语的领域通用度 $U_k=DC_k \times UL_k$。

上式中，n 表示领域类数，要求各领域类语料库语料等量；Fk_i 是词语在第 i 领域类 k 号词的频度，UL_k 表示 k 号词的文本使用度，DC_k 表示 k 号词的领域类分布均匀度。

3.2 时间通用度的定量分析方法研究

词语的时间通用度是词语在考察时间内通用程度的量化衡量指标。它需要观察词语在考察期内使用是否稳定，即词语词频在各月分布的均匀程度。

时间通用度计算步骤如下：

（1）计算词语月频度 F_k

F_k 为 k 号词语在各月语料中出现的总频次。

（2）计算 k 号词语的时间通用度 T_k

采用分布均匀度(Distributional Consistency，英文简称 DC)计算词语在考察时间内各月分布的均匀程度，计算公式为：

$$SMR = (\sum_{i=1}^{n}\sqrt{FK_i} / n)^2 \qquad Mean = (\sum_{i=1}^{n} Fk_i / n)$$

K 号词语的时间通用度 T_k= SMR/Mean。（$0 \leqslant T_k \leqslant 1$）

上式中，n 表示考察时间内月个数，要求各月中语料库语料等量；Fk_i 是词语在第 i 个月的词频度。

3.3 地域通用度的特征描述

词汇的地域通用度从共时的角度观察词语在不同地域的媒体使用情况，即在考察时间内，词汇在不同地域媒体中使用的稳定程度。

地域通用度类似于时间通用度，它们均是考察词汇在不同分类体系中分布的均匀程度，即使用的稳定程度，区别在于时间通用度按年度中月份进行分类，而地域通用度则按不同地域的媒体进行分类。

由于目前实验所用的原始语料来源只有《新疆日报》哈萨克语版，地域的代表性不足，所以地域通用度对词汇通用度的影响暂时没有纳入到考虑范围。

3.4 词汇通用度的计算方法

如前所述，目前所描述的“词汇通用度 O_k”是综合考虑词语的领域通用度及时间稳定度而提出的，并未考虑地域通用度对词语通用度的影响。

词汇通用度的计算公式为：

词汇通用度 $O_k = T_k \times U_k$。

T_k 表示 k 号词的时间通用度，U_k 表示 k 号词的领域通用度。O_k 表示词语的通用程度，该值越大，k 号词的常用性特征及考察时间内使用稳定性特征表现就越好。

4 哈萨克语通用词汇自动提取系统

4.1 语料的选取及预处理

实验中，采用《新疆日报》哈萨克语版 2008 年度一年的电子文本数据进行统计。

为了考察词汇的领域通用度，需要对原始媒体语料进行分类，本文将原始语料分为“政治”、“经济”、“教育”、“生活”、“体育”5 类。

原始语料为媒体报纸网页格式，需要按“年月日”将原始语料转化为纯文本格式语料，同时应该滤除网页格式中的垃圾信息，只保留有效的文本信息内容，转换后文件格式为 TXT 文件。

4.2 系统结构

为了计算哈萨克语词的词汇通用度，首先将预处理后的纯文本格式语料入库，接着对入库语料进行提取处理，然后对提取出来的词进行词频统计，得到词频后便可计算出词汇的领域通用度和时间通用度，最后计算出哈萨克语词的词汇通用度。系统操作流程图如图 1 所示。

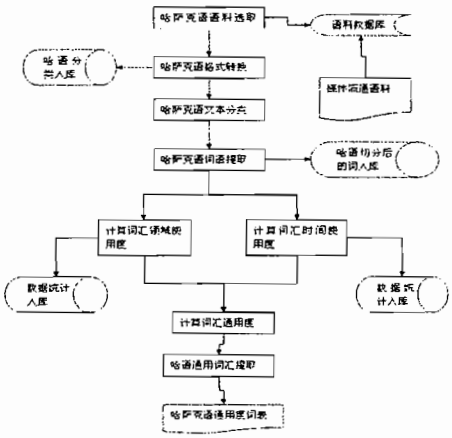

图 1　哈萨克语通用词汇提取流程图

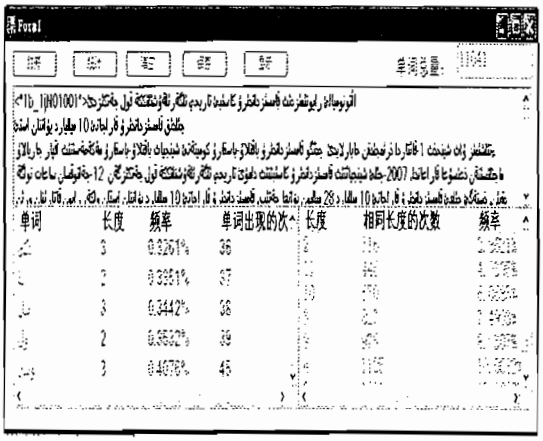

图 2　哈萨克语词频统计界面

4.3 哈萨克语词汇通用度的计算步骤

哈萨克语词汇通用度的计算过程及系统界面如下：

（1）计算每个月内哈萨克语“词”的词频 F_k 及文本数

Step1：计算各领域类、每月中每个文本中“词”的词频数。

Step2：计算各个领域类、每月中所有文本中“词”的词频数、文本数。

“哈萨克语词频统计”系统界面如图 2 所示。

（2）计算各领域类、每月中所有文本“词”的散布系数 D_K（值 0~1）及文本散布度 UL_k。

（3）合并各领域类表。

（4）计算一个月类各领域类的总词次。

（5）计算词频，对数据进行归一化处理。

归一化处理的目标，使得与分类语料的大小无关，且使数据平滑。

（6）计算每月中所有领域类 “词”的文本散布度 UL_k 和领域通用度 U_k。

“哈萨克语词汇领域通用度统计”系统界面如图 3 所示。

（7）计算年度哈萨克语词时间通用度 T_k 和词汇通用度 O_k。

“时间通用度和词汇通用度统计”系统界面如图 4 所示。

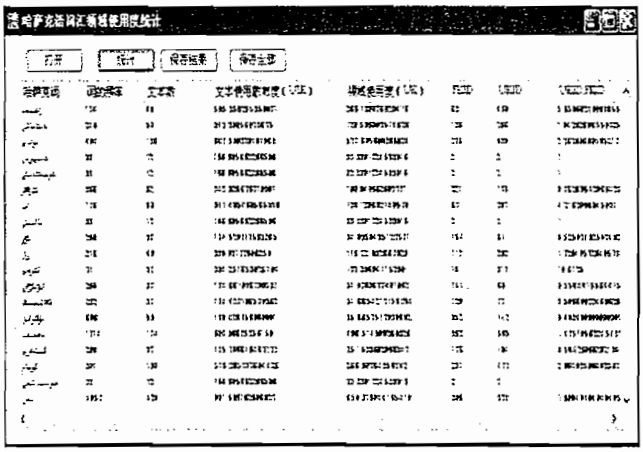

图3　哈萨克语词汇领域通用度统计界面

图4　时间通用度和词汇通用度统计界面

4.4 哈萨克语通用词汇的提取

如前所述，词汇通用度表明了词语在大众流通媒体中的通用程度，它融合了词语词频、文本散布、领域分布、使用时间稳定等统计特征，是这些能力的综合表现程度的量化衡量指标。所以，提取哈萨克语通用词汇，参考提取汉语通用词汇的方法[2]，按 2008 年度《新疆日报》哈萨克语版所选语料词的“词汇通用度 O_k”的值从大到小进行排序，取覆盖总语料 85%~90%的词作为通用词汇。

5 实验结果及分析

根据哈萨克语词汇通用度的计算方法，采用 C#语言进行系统开发，实现了哈萨克语词汇通用度统计系统，按照哈萨克语通用词汇的提取规则，成功提取了哈萨克语通用词汇。哈萨克语词汇通用度部分实验结果如表 1 所示。

表中 OKID 表示按“词汇通用度 O_k”的值进行排序后的位置顺序编号。

从实验结果来看，提取方法基本令人满意，但词汇通用度的准确性还有待于进一步提高，影响准确性的主要因素有：

（1）本实验中，将原始语料分为“政治”、“经济”、“教育”、“生活”、“体育”5 类，而这五类无法囊括所遇到的报纸媒体的各类文本。

（2）目前“词汇通用度”没有考虑地域通用度对词语通用度的影响。

（3）目前实验所用的只有《新疆日报》哈萨克语版 2008 年度一年的语料，语料规模上受到很大限制。

表 1 哈萨克语“词汇通用度”部分实验结果

哈萨克语词	词频 FK	文本数	文本使用散布度 ULK	领域通用度 UK	时间通用度 TK	词汇通用度 OK	OKID
جان和	356	374	8941.6452	8539.7060	0.95187	8128.7041	737
بىر一个	291	386	8892.6146	8711.3119	0.75388	6567.3362	736
دە	189	261	8825.5426	8605.3388	0.72414	6231.4522	735
كەرەك应该	140	187	8700.0821	7210.1816	0.74866	5397.9969	734
بۇتىن整个	68	125	8582.6442	7483.6931	0.54400	4071.1290	730
لاۋلى	83	100	6933.5912	4877.2432	0.8300	4048.1118	729
ءار每	140	275	8708.4042	7800.9534	0.50909	3971.3944	728
دەپ说了	97	173	8552.3111	6051.5284	0.56069	3393.0535	722
ءۇشىن为了	120	261	8707.4002	8650.4981	0.39080	3380.6544	721
تيىس该	109	186	6980.2351	5176.0691	0.58602	3033.2878	716

6 结论及展望

本文通过对哈萨克语通用词汇的界定，得到哈萨克语通用词汇的三大特征：领域通用性、时间通用性及地域通用性，用与这些特征相对应的领域通用度、时间通用度及地域通用度等量化指标来衡量词汇的通用程度。本文中的“词汇通用度 O_k”是综合考虑词语的领域通用度及时间稳定度而提出的，并未考虑地域通用度对词语通用度的影响。根据哈萨克语词汇通用度的计算方法，采用 C#语言进行系统开发，实现了哈萨克语词汇通用度统计系统，并成功提取了哈萨克语通用词汇。

为了进一步提高哈萨克语词汇通用度的准确性，增加统计数据的说服力，下一步的工作应该是扩大语料库的规模，加大文本分类的数目，而不是仅限于目前的 5 个领域，还应考虑较大地域范围流通语料，比如增加哈萨克斯坦的媒体语料等，地域通用度纳入“词汇通用度”的考察范围。

参 考 文 献

[1] Qiangjun Wang, Isabella Park, Pu Zhang. Automatic Extraction of The Unlisted Terms In The Field of Information Technology Based on The Dynamic Circulation Corpus. Proceedings of IEEE, 2003. pp.452-458.

[2] 赵小兵. 基于 DCC 的现代汉语基本词汇自动识别与提取方法研究[D]. 北京: 北京语言大学, 2007.

[3] Jirapa Vitayapirak, Phornsuk Ratiroch-anant. Computational Approach for Processing of Control Engineer Text: Applications for Corpus Lexicography. Proceedings of IEEE, 2006.

[4] 韩秀娟. 基于动态流通语料库的通用词语用字研究及字词语关系考察[D]. 北京: 北京语言大学, 2007.

[5] 王灿辉, 张 敏, 马少平. 自然语言处理在信息检索中的应用综述[J]. 中文信息学报, 2007, 21(2): 35-45.

[6] 嘎日迪, 赵小兵, 马红旭等. 蒙古文自动处理系统研究[J]. 中文信息学报, 1999, 13(4): 57-62.

[7] 古丽拉·阿东别克, 达吾勒·阿布都哈依尔, 木合亚提·尼亚孜别克等.现代哈萨克语词级标注语料库的构建研究[J]. 新疆大学学报, 2009, 26(4): 394-401.

[8] 唐长宁. 基于现代汉语动态流通语料库的通用词汇自动提取方法研究[D]. 呼和浩特: 内蒙古师范大学, 2008.

[9] 毕丽克孜. 语料库语言学的应用和维吾尔语语料库词频统计的意义[J]. 新疆师范大学学报, 2005, 26(2): 226-228.