现代维吾尔语常用词统计关键技术研究

艾孜尔古丽[1]，努尔艾合买提[1]，玉素甫·艾白都拉[1]
（1.新疆师范大学，新疆维吾尔自治区 乌鲁木齐市 830054）

摘要：本文研究了构建现代维吾尔语言语料库的关键技术与方法,特别是现代维吾尔语言语料库的构建，并对现代维吾尔语语料预处理技术，现代维吾尔语语料统计技术，现代维吾尔语词干提取技术，现代维吾尔语数据分析技术进行了研究;研制了现代维吾尔语常用词候选表，从词语的使用频度和词语的分布两方面对词语进行了基本考察，将维吾尔语词语的“词种数、频次、频率、文本数、词长”作为常用词候选表的依据。

关键词：现代维吾尔语；语料库；常用词候选表；计量分析
中图分类号：TP391　　**文献标识码：**A

Research On Key Technology Of Modern Uyghur Language Common Words

Azragul[1], Nurahmat[1], Yusup Abaydula[1]
(1.Xinjiang Normal University, Urumqi, Xinjiang, 830054, China)

Abstract: This paper studies key technologies and methods of the modern Uyghur language corpus construction, in particular the construction of modern Uyghur language corpus, and the modern Uyghur corpus pretreatment technology, modern Uyghur corpus statistical techniques, modern Uyghur stem segmentation technology, modern Uyghur data analysis techniques were studied; To developed a modern Uyghur common words candidate list, the author thinks that carry out investigation mainly from two aspects: word use frequency and distribution, The Uyghur words′ ″the word species, frequency , frequency, number of texts, word length ″ will be regarded as the foundation.

Key words: Modern Uyghur language; Corpus; Common words lexicon; quantitative analysis.

1 前言

现代维吾尔语常用词计量研究是少数民族语言信息处理领域急需研究的重要课题。维吾尔语常用词汇表的欠缺，是影响维吾尔语词汇学、计算语言学和维吾尔语信息处理工作质量的重要因素,迫切需要研制具有代表性、可靠性、权威性的维吾尔语常用词汇表，促进维、哈、柯等阿尔泰语系的新疆少数民族自然语言理解跨越式发展。

为确保收集语料的可靠性、代表性和权威性，本文重点对语料来源、语料范围、语料载体等进行了研究，以保证常用词候选表的权威性和代表性。

---

* **收稿日期：**　　**定稿日期：**

基金项目：新疆维吾尔自治区自然科学基金（2014211A045）；教育部人文社会科学一般项目（14YJC740001）；新疆维吾尔自治区高校科研计划青年教师科研启动基金(20140706213103147)；国家自然科学基金重点项目(No.61132009)；国家自然科学基金项目(No.61262066)；国家语委“十二五”科研规划项目(YB125-45)。

（1）在现有的语料库资源基础上，系统、持续地进行收集、整理、加工和处理现代维吾尔文平面媒体、教育教材媒体、有声媒体、网络媒体语料，构建现代维吾尔语语料库，相比之前的语料库，本语料库语料来源更广、语料领域更宽、各个领域比率控制适当。

（2）对构建现代维吾尔语言语料库的关键技术与方法进行进一步优化与完善，新增了人名识别和数据自动分析技术。对词语使用频次及其词汇文本数进行基本考察，从词语的使用频度和词语在文本中出现的次数两方面加以考虑，提出了现代维吾尔语常用词候选表。

本研究不但为维吾尔语等少数民族自然语言理解及处理工作提供了基础，也可为阿尔泰语系的少数民族语言的规范化、教材设计、中小学语文教育、扫盲教育、双语教育和辞书编纂提供服务。

2 现代维吾尔语言语料库的资源建设研究

为了做好收集媒体语料的可靠性、代表性和权威性，对语料来源、语料范围、语料载体等方面进一步研究，保证常用词候选表的权威性和代表性，根据现有语料具体情况，以传播媒体作为筛选依据。

本语料库是由平面媒体（主要文学作品和经典名著为主，代表文学语言）、教育教材媒体（新疆教育出版社、新疆科技出版社、新疆人民出版社、美术出版社等正规出版社出版的正规出版物，代表科学技术、文化、金融、工业生产多领域的文学和生活语言）、有声媒体（新疆电视台每天播出的 30 分钟新疆新闻和 30 分钟的新闻联播文本语料，代表新闻报道语言）、网络媒体（十多家比较正规网站，主要网络语言生活为主）组成的总语料。它基本代表维吾尔族人政治、经济和社会生活的方方面面。

本语料资源是由国家语言资源监测中心少数民族分中心“维吾尔语文研究基地”、新疆师范大学“网络信息安全与舆情分析重点实验室”（以下简称实验室）提供。

2.1 平面媒体

本语料是以解放以来国家正式出版社出版的文学作品为主组成的平面媒体语料的收集作为研究对象，语料容量 188MB，总语料容量中的比率 26.81%。

2.2 教育教材媒体

本语料是等科普性和教育性较强的正式出版物组成的教育教材媒体语料的收集作为研究对象，语料容量 173MB，总语料容量中的比率 24.67%。

2.3 有声媒体

研究所使用收集的语料来源于新疆电视台每天 30 分钟播出的新疆新闻和每天的新闻联播 30 分钟的文本语料。采集的语料时间跨度为 2010 年 1 月至 2012 年 12 月，共 1080 天的 1080 小时播放时间的文本语料。栏目领域国际、国内、新疆、教育、文化、经济、健康、生活等各行各业发生各种各样的新闻事件，语料容量 171.2MB，总语料容量中的比率 24.42%。

2.4 网络媒体

研究网络媒体网站对象是新疆政府网、昆仑网、天山网等 18 家网站。收集语料时间跨度 2006.4--2012.12 之间，栏目领域国际、国内、新疆、教育、文化、经济、健康、生活、计算机等社会发展的各个领域，语料容量 169MB，总语料容量中的比率 24.10%。

3 现代维吾尔语常用词关键技术与方法研究

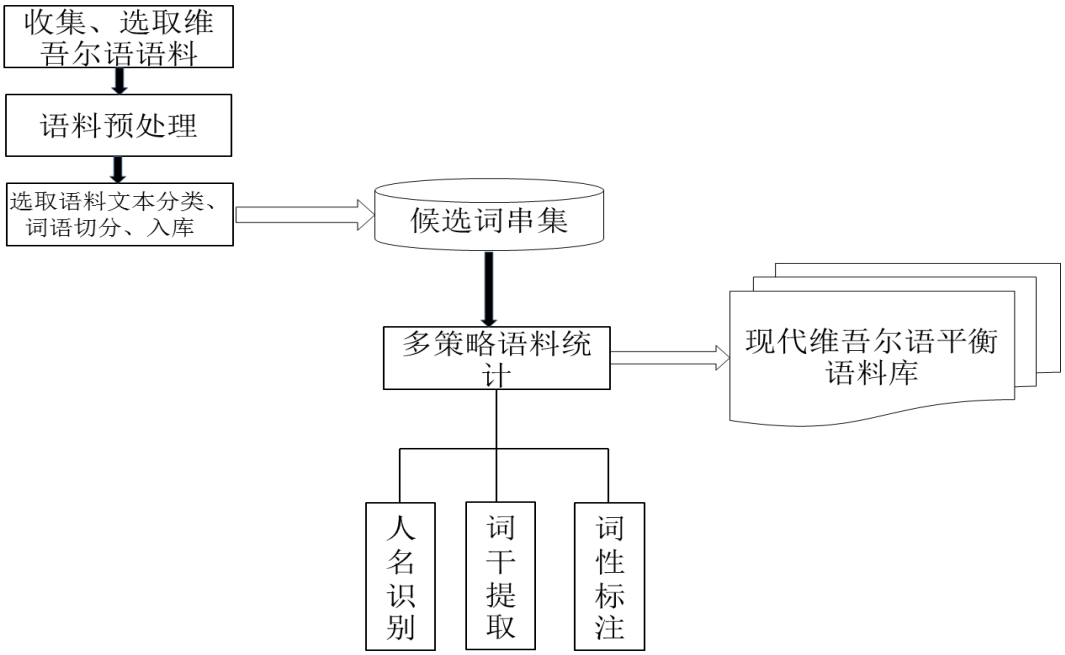

图 1 基于平衡语料库的现代维吾尔语多策略统计模型

⑴ 现代维吾尔语语料预处理技术：收集语料，对语料进行预处理，并形成文本文件。

⑵ 现代维吾尔语语料统计技术

①对调查语料统计：研究项目包括词次、频率、词种、词长和文本数等项目，最终形成维吾尔语词频表。

②人名识别：根据维吾尔族人、汉族人、外国人姓名维吾尔语中描述特点，对已研究的识别技术进一步优化，确定识别规则，解决汉族人名中姓和名空格隔开描述问题，优化汉族、外国人姓名识别率。

⑶ 现代维吾尔语词干提取技术

词干提取词干中利用基于词典和人机交互技术结合方法提取词干。提取词干过程中，通过现代维吾尔语词干词典维护来发现提取词干过程中出现的新词干，并对机器词典中新词干的进行补充，增加机器学习等功能。

维吾尔语文词语的具体构词方式见图 2。

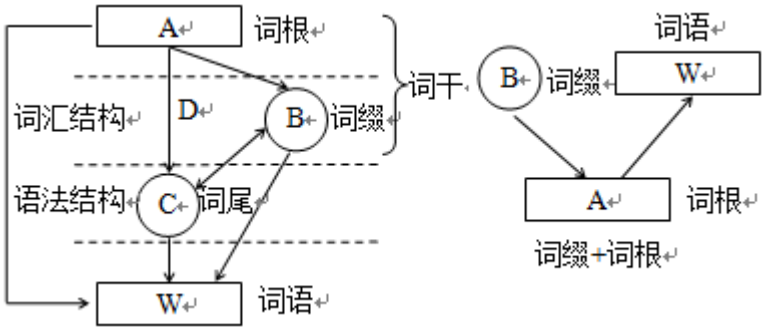

图 2　现代维吾尔语词法结构模型

图 2 中，A 表示词根，B 表示词缀，C 表示词尾，D 表示词干，W 表示词语。

⑷ 现代维吾尔语数据分析技术

本技术主要解决常用词、次常用词、部分常用词、独用词、词种分布、覆盖率等计算数据等几个部分的自动分析技术。

词次（频次）：每一调查对象的频次同其前调查对象频次的累加和。频次是一个具体的数字，它直观地反映了某个词语在语料中真实、原始的使用情况。

$$A_i = \sum_{k=1}^{i} n_k$$

其中：Ai 为调查对象 i 的累加频次，ni 为调查对象 i 的出现次数。

频率：每一调查对象的词次的累加和，与所有语料中调查对象总次数的比值，即：

$$Bi = \sum_{k=1}^{i} n_k / N \times 100\%$$

其中：Bi 为调查对象 i 的累加频率，ni 为调查对象 i 的出现次数，N 为所有语料中调查对象出现的总次数。

一般来说，频率愈高的词其常用程度愈高。这是最直观，且大多情况下都颇有成效的统计方法。

累加覆盖率：指所有词语的频率由高到低降序排列时，每一个词语与其前词语的频率之和在全部语料中所占的比重。

$$F_i = \sum_{i=1} n_i / N \times 100\%$$

其中：Fi 为调查对象 i 的覆盖率，ni 为调查对象 i 的出现次数，N 为所有语料中调查对象出现的总量。

累加覆盖率的作用是能清楚观察到每个词在由高到低的频率排序中在词语整体中所处的位置。

⑵词语领域通用度的定量分析

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。

本项目采用改进后的领域通用度计算步骤如下：

①计算领域类词语频度 Fx:

FX 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 UIX:

采用 A.Juilland 公式计算词语的文本使用度:

$$s_k = \sqrt{\sum_{i=1}^{n} (N_k^i - N_k)^2 / n}$$

$$D_k = I - S_k / (N_k \times (n-1)^{\frac{1}{2}}) \qquad (0 \leqslant D_k \leq 1)$$

词的文本使用度： $UI_K = D_K \times F_k$ (取整数值)

其中,N_k^i公表示 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：$DC_k = SMR/Mean \qquad (0 \le DC_k \le 1)$

SMR 及 Mean 分别定义如下：

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

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

K 号词语的领域通用度：$U_k = DC_k \times UI_k$

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

⑶词语时间通用度的定量分析

词语的时间通用度是词语在考察时间内通少日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 Fk:

Fk 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 Tk:

采用分布均匀度计算词语在考察时间内各月分布的均匀程度,计算公式为:

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

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

K 号词语的时间度通用度：$T_k = SMR/Mean \qquad (0 \le T_k \le 1)$

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；Fki;是词语在第 i 个月的词频度。

⑷词汇通用度的定量分析

词语通用度 Ok 是综合考虑词语的领域使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,以后在考虑较大地域范围流通语料时,应纳入地域通用度的考察。

词汇通用度的计算方法为:

词语通用度 $O_k = T_K \times U_k$

Tk 表示 k 号词的时间通用度,Uk 表示 k 号词的领域通用度。Ok 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 现代维吾尔语常用词候选表的研制研究

对词语进行基本考察，从词语的使用频度和词语的分布两方面加以考虑。维吾尔语词语的“词种数、频次、频率、文本数、词长”作为常用词候选表的依据。

在此基础上，提取出不同媒体语料库的高频词表，在四个词表中，筛选出不同媒体语料的共用词，作为现代维吾尔语常用词候选表；筛选出任意三个语料库的共用词，作为现代维吾尔语次常用词候选表；筛选出任意两个媒体语料库的共用词，作为现代维吾尔语部分常用词候选表；筛选出各媒体语料库的独用词，作为现代维吾尔语独用词候选表。

5 实验数据

5.1 基本数据

本文的研究语料涵盖平面媒体、有声媒体、网络媒体、教材媒体四种．共计 96,025 个文本文件，43,529,435 词次。现代维吾尔语语料采集的依据及选择详见本文第三部分。

本语料平面媒体（文学作品语料）、教育教材媒体（科普教材媒体）、有声媒体（新闻语料）、网络媒体（网络语料）组成的总语料。它基本代表维吾尔族人政治、经济和社会生活的方方面面。语料具体情况如下表 5-1 所示。

表 5-1 总语料的分布情况表

语料媒体	平面媒体	教育教材媒体	有声媒体	网络媒体	总语料
词次	11 879 662	12 195 468	10 587 381	8 866 924	43 529 435
词种数	350 760	273 230	216 021	323 660	703 669
词干种数	106 386	91 892	68 053	78 333	147 054

5.2 常用词汇与常用词干比较数据

为了保证常用候选代表性和权威性，现代维吾尔语词作为常用词或现代维吾尔语词干将作为常用词，需要进一步确认的问题。根据维吾尔语的特点和具体四大媒体语料为依据，对语料统计数据进行比较分析。

（1）现代维吾尔语词语基本数据

选取平面媒体、教育媒体、有声媒体、网络媒体等四大媒体的四个词表进行比较，提取四大媒体共用总词表。本表共收录了现代维吾尔语常用词语 62,330 个，具体情况如表 5-2 所示。

表 5-2 现代维吾尔语常用词语情况表

项目	词种数	占词种数比例（%）	频次	占词频次中比例（%）
现代维吾尔语常用词语	62 330	0.18	33 834 388	77.73

表 5-2 可以看，词种 62,330 个共用词对总语料覆盖率 77.73%。说明对总语料的覆盖率相对偏低，不能承担代表现代维吾尔语常用候选词角色。

（2）四大媒体词干基本数据

同样对四大媒体的四个词干表进行比较，提取四大媒体共用总词干表。本表共收录了现代维吾尔语常用候选词干 36,488 个，具体情况如表 5-3 所示。

表 5-3 现代维吾尔语常用词干情况表

项目	词干种数	占词干种数比例（%）	频次	占词干频次中比例（%）
现代维吾尔语常用词干	36 488	24.85	41 452 953	95.23

表 5-3 可以看，36,488 个共用词干占总语料覆盖率 95.23%。说明对总语料的覆盖率接近整个语料，能承担代表现代维吾尔语常用候选词角色。

5.3 现代维吾尔语高频词、高频词干基本数据

高频词是指在语料中词频累加覆盖率达到 90%的全部用词。根据这个定义，从每一种媒体语料中覆盖率达到 90％时提取高频词，高频词、词干种总语料中分布情况表 5-4 所示。

表 5-4 高频词、词干种总语料中分布情况

项目	词频	占总词次比例（%）	词种	总词种比例（%）	词干种	占总词干种比例（%）
平面媒体	10 691 698	24.56	43 901	6.24	12 224	8.31
教育媒体	10 975 954	25.22	24 233	3.44	9 561	6.5
有声媒体	9 528 624	21.89	12 794	1.82	4 595	3.12
网络媒体	7 980 238	18.33	29 398	4.18	8 165	5.55

总语料	43 529 435	90.00	703 669	15.68	147 054	23.48

表 5-4 可以看出，每一种媒体语料在总语料中分布情况。

以词干能代表现代维吾尔语常用候选词角色特点为依据，根据高频词在媒体中分布情况，确定现代维吾尔语共用词、部分共用词、准部分共用词和独用词等四个档次。计算时教育媒体定义为 A、平面媒体定义为 B、网络媒体定义为 C、有声媒体定义为 D。四大媒体（ABCD）共用部分叫做常用候选词（共用词）；任意三种媒体（ABC、ABD、ACD、BCD）和任意两个媒体中共用的（AB、AC、AD、BC、BD、CD）中共用的部分叫做次常用候选词（大部分共用词）只有一种（A、B、C、D）媒体中出现的词叫做独用词。经过四大媒体高频词干进行比较，提取常用候选词、次常用候选词和独用词。常用候选词和独用词的具体情况表 5-5 所示。

表 5-5 常用候选词和独用词表

项目	常用词种	占常用词种比例（%）	词次	占总词次比例（%）
常用候选词	3 186	19.9	30 468 709	77.77
次常用候选词	5 889	36.79	6 861 820	17.52
独用词	6 934	43.31	1 845 738	4.71
合计	16 009	100	39 176 267	100.00

表 5-5 可以看，常用词和独用词的分布情况。由于常用候选词和次常用候选词合并后占总高频词语料中的比例 90.20% 。这说明提取的常用候选词表对本次考察语料是可行的。

5.4 现代维吾尔语常用词候选表

表 5-6 给出词次 10 万次以上的 22 条高频常用候选词样例表。

表 5-6 词次 10 万次以上的 22 条高频常用候选词样例表

词汇	汉译	频次	文本数	词汇长度	分布
ئۇ	他、她、它	563 989	42 862	1	A,B,C,D
ۋە	和	542 404	63 428	2	A,B,C,D
بۇ	这、这个	508 802	64 100	2	A,B,C,D
بىلەن	与	470 366	60 264	5	A,B,C,D
بىر	一	436 498	54 179	3	A,B,C,D
قىلىش	做（将来时）	191 621	38 740	5	A,B,C,D
بولۇپ	做（过去时）	185 521	38 089	5	A,B,C,D
قىلىپ	做（过去式）	149 610	35 071	5	A,B,C,D
بولغان	做过	145 446	33 361	6	A,B,C,D
دەپ	说	137 666	19 444	3	A,B,C,D
شىنجاڭ	新疆	121 936	28 032	6	A,B,C,D
كېيىن	以后	121 102	29 219	5	A,B,C,D
دۆلەت	国家	119 765	30 888	5	A,B,C,D
ئىككى	二	119 575	27 166	4	A,B,C,D
كېرەك	必须	119 571	18 168	5	A,B,C,D
بولىدۇ	行、可以	119 300	17 380	6	A,B,C,D
ئۈچۈن	为	115 657	27 821	4	A,B,C,D
ئادەم	人	114 039	20 875	4	A,B,C,D
خەلق	人民	111 886	24 937	4	A,B,C,D

جۇڭگو	中国	105 568	26 138	5	A,B,C,D
ئاپتونوم	自治	102 111	23 752	7	A,B,C,D
شۇ	是、就是	101 740	21 616	2	A,B,C,D

6 总结

在维吾尔语基地相关研究的基础之上选取了更大规模的真实语料建成现代维吾尔语语料库，其语料库包括平面媒体、教材媒体、有声媒体、网络媒体等四类主流媒体。语料量43,529,435 词次。而现阶段，这些资源的合理、有效应用，对于深化与扩展语言资源的监测工作有重要意义，同时也是计算语言学服务于语言生活、语言教学、语言工程、辞书编纂等方面的重要体现与有益尝试。其中，四大媒体语言文字使用频率变化、频序排位相对变化反映了媒体对社会生活的关注点的变化。透过这些字词语的使用状况可以看到年度的社会生活、时事面貌。

参考文献

[1] 艾孜尔古丽等，中小学维吾尔语文教材用词数据分析方法与应用研究，计算机工程与应用（核心），P108-111 页，2014.2,第一作者，
[2] 艾孜尔古丽等，现代维吾尔文网络媒体用词研究，计算机应用与软件（核心），P67-68 页，2012.2，第一作者。
[3] 艾孜尔古丽等，基于网站用词调查的现代维吾尔语词干提取和应用，计算机应用与软件（核心），P32-34 页， 2012.3，第一作者。
[4] 艾孜尔古丽等，现代维吾尔语语言资源监测中数据分析技术研究 ，计算机应用与软件（核心），P36-39 页， 2013.4，第一作者。
[5] 玉素甫等，基于网站用词调查的现代维吾尔语词尾切分和应用研究，计算机应用与软件（核心），P13-15 页， 2012.4，第二作者。
[6] 玉素甫等，基于网站用词调查的现代维吾尔语词长研究，计算机应用与软件（核心），P32-34 页，2012.5，第二作者。
[7]玉素甫等，信息处理用现代维吾尔语词干类标记集研究，信息技术与标准化，P45-48 页，2011.6，第三作者。
[8] 苏新春，《汉语词汇计量研究》，厦门大学出版社，2001
[9] 苏新春，杨尔弘，2005 年度汉语词汇大规模统计的分析与思考，厦门大学学报，2006 年 06 月
[10]赵小兵，基于动态流通语料库的现代汉语基本词汇自动识别与提取方法研究，博士学位论文，2007.6

作者简介：艾孜尔古丽（1987——），女，讲师，主要研究领域为计算语言学、自然语言处理。Email:Azragul2010@126.com；努尔艾合买提（1984——），男，研究生，主要研究领域为计算语言学、自然语言处理；通讯作者：玉素甫·艾白都拉（1958——），男，教授，主要研究领域为计算语言学、自然语言处理。Email:ysp2002@126.com。

第一作者：

通讯作者：