现代维吾尔语统计分析技术研究

艾孜尔古丽[1,2]，李晓[1]，玉素甫.艾白都拉[2]

[1](中国科学院新疆理化技术研究所 乌鲁木齐 830011)

[2]（新疆师范大学　乌鲁木齐　830054）

ysp2002@126.com

摘要：随着社会的发展变化，语言生活也在不断地发展变化。为了切实掌握小学初中维吾尔语文教材中用词情况，对小学、初中维吾尔语文教材作为研究对象，用词概况进行研究。在本文中首先陈述研究使用的语料；其次介绍统计系统研究概况；三、小学、初中维吾尔语文教材用词研究包括研究总词次、总词种数、总词干种数；四、讨论与分析词频与词种、词种覆盖率、词种。

关键字：维吾尔语文、小学、初中维吾尔语文教材、词语、统计分析

中图分类号：TP391　　**文献标识码：**A

Research of Modern Uyghur language statistical analysis technology

Azragul ,Alim Murat,Yusup Abaydulla

(Xinjiang Normal University, Urumqi, 830054)

Abstract: With the development of our society, the languages are also constantly evolving. In order to master the use of the Uyghur language teaching material terms conscientiously and to grasp the information of the Uyghur language term, we can make a study of elementary and junior high school material term. In this article, first of all, to state the corpus which is being used; Secondly introduces the general situation of the study statistical system; Third I describe the total term times, the total word species and the total stem species; Forth, I discus and analyze word frequency and word type, word type coverage rate, word type.

Key words：Uyghur language; new curriculum elementary and junior high school teaching material; word; statistics analyze;

1 研究对象

语文教材是语文课的重要教学依据，其用词状况直接关系到语文教学的效果。本次主要研究了小学初中维吾尔语文教材中的用词实态状况进行研究。本次研究对象选取了由新疆维吾尔自治区九年义务教育教材审定委员会审定，新疆教育出版社出版的九年义务教育新课程标准普通班实验教科书(简称普通班版)的维吾尔语文教材和九年义务教育新课程标准双语班实验教科书(简称双语班版)的维吾尔语文教材。由于一年级教材主要以识字为主，词汇量非常小，所以本次调查选取了小学二年级到初中三年级共8个年级的两种版本的语文教材。每个年级的教材分上下两册，共16册，两套教材合计32册。

2 现代维吾尔语统计技术

为了进行维吾尔语文教材语料中用词实态情况的研究，利用统计技术、计算机技术和人工智能技术开发了“现代维吾尔语语料统计系统”，该系统是由现代维吾尔语电子语料文件格式转换模块、现代维吾尔语文本调整模块、现代维吾尔语文本校对模块、现代维吾尔语语料统计模块、现代维吾尔语语料管理模块 5 个模块组成。

2.1 现代维吾尔语电子语料文件格式转换模块

本模块是把 DOS 环境下的 8 位北大方正系统文件格式（简称 BDFJ）的语料转换成维文 UKIJ 格式，这一项工作也是国家自然科学基金 60463005 项目的成果，经过改善和相对调整后，在本项目中应用解决了现代维吾尔语生语料的选取和录入、整理工作和编辑工作。

2.2 现代维吾尔语文本调整模块

打字员录入过程中经常产生词汇中间多余空格、标点符号、段落符号、换行符等乱敲键

盘等错误，但是在排版和现实过程中肉眼是看不到这些错误。因此统计系统也会在统计过程中把一个词语错误地统计成两个词语，以致于不能获得正确的统计结果。为了获得正确的现代维吾尔语文本统计结果，利用模式匹配、统计和机器学习结合技术，研究实现现代维吾尔语文本调整模块。该模块主要解决调整少数点符号、句子末尾的点符号与文章小标题序号方边的点符号、文章末尾产生的乱回车等问题。

2.3 现代维吾尔语文本校对模块

为了保证语料质量，需要解决现代维吾尔语文本调整模块不能解决的现代维吾尔语文本中出现的正字正音问题。若不处理此问题，语料校对环节的工作量会大大增加，人工校对的方式已无法适应迅速增长的电子文本的数量。为了提高语料校对效率和减轻人员低层次的重复劳动负担，本课题组研制了一种音节统计、单词统计和语法规则相结合的,以音节匹配为主要手段的维吾尔语文本校对系统，解决现代维吾尔语自动校对问题。

2.4 现代维吾尔语料管理模块

语言资源建设是非常庞大的系统工程，其中非常重要的环节是大规模语料的管理问题。为了管理好大规模语料，研制了现代维吾尔语语料库管理模块，它是由文本管理模块、语料类型分类模块和子语料库生成模块等三个子模块组成。该模块主要解决文本入库、文本删除、文本查询、文本分类（平面媒体、网络媒体等）和按某一特殊需要，从文本当中选取文本中的片段已生成子语料库。

2.5 现代维吾尔语统计模块

根据现代维吾尔语用词调查项目的需求，需要考察语料中出现的每一个词汇、该词汇的频次、频率、词语长度、文本数等项目，本课题组研制了现代维吾尔语统计模块。众知维吾尔语属于黏贴性语言，词语与词语之间空格来区分，但社会发展的需要维吾尔语部分词语应用汉语或外来语借词，另外维吾尔语词语长度有 1 字符（只有一个）、2 个字符、3 个字符、4 个字符、5 个字符词语等不同的长度。在汉族人名撰写成维吾尔语时，汉族人名的姓和名之间用空格来分割，每一个撰写汉族人的姓长度 2 个字符、3 个字符、4 个字符不等。例如：جاڭ زېمىن （张泽民）中的جاڭ不是维吾尔语中的词语，这对统计工作带来词语判断歧义。为解决此问题，本课题组采取以下措施解决词语判断歧义问题。

1）通过大量语料统计，并过滤词语长度 1-5 字符的词汇，建立一个样本数据库。

2）统计过程中遇到长度 1-5 字符的词语时，先把该词与样本数据库中的词语进行比较，如果此词在数据库中存在，则认定为维吾尔语词语，否则认定为非维吾尔人名词语，并且将该词后面空格和紧跟的词汇一起添加在非维吾尔人名词库中。例如：جاڭ زېمىن （张泽民）。在以上原理基础上，采用的单词统计和构词规则相结合技术、人机交互技术和机器学习技术保证了统计数据的可靠性和精确性。以下图 1 所示语料统计模型结构图。

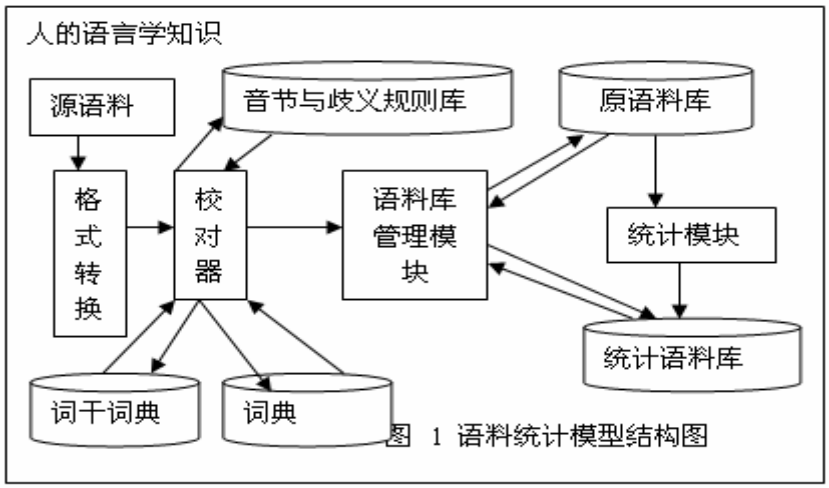

图 1 语料统计模型结构图

3 小学初中维吾尔语文教材用词研究

3.1 基本数据

本次用词研究项目有词在语料中出现的频次、频率、词语长度、文本数等。维吾尔文的词可由单个或多个字母组成，且字母在书写时必须以词为单位连体书写。词与词之间以空格隔开，标点符号紧跟在词语之后。两套教材总词次：540,294 次；两套教材总词种数：68,979 个；两套教材总词干种数：24,568 个。

表1 普通班版与双语班版小学、初中语文教材基本数据

教　材	词次	词种	词干种	共用词干	独用词干
普通班版	293 268	56 174	22 171	18 728	3 443
双语班版	247 026	50 249	21 125		2 397
两套教材相差	46 242	5 925	1 046		1 046

表 1 可见，普通班版与双语班版词次相差 46,242，词干种数量相差 1,046，词种数量相差 5,925，说明两套版本教材语料的量、词种数和词干种数具有不小的差异。

3.2 长词使用分析

两套教材出现的最长词长度为 27 个字母。下面列出 10 个长词的例子：

مەركەزلەشتۇرەلمەيدىغانلارنى（不能集中的）、**چىقىرىۋاتقانلىقىڭلارنىمۇ**（就连被你们赶出去）、**بويسۇندۇرالايدىغانلىقىغا**（去征服）、**ئايرىلالمايدىغانلىقىمىزنى**（我们不能离开彼此）、**مۇجەسسەملەشتۇرۇۋالغاندەك**（好像明白的样子）、**مۇئەييەنلەشتۇرمەكچىمەنكى**（我准备确定）、**ئالماشتۇرۇۋېلىنغانلىقىنى**（被交换）、**يوشۇرۇنالمايدىغانلىقىنى**（无法被隐瞒）、**ئايلاندۇرالايدىغانلىقىغا**（可以转动（运转或周转））、**مەسلىھەتلىشىۋالغاندەكلا**（就像提前商量好了似的）。

其中第一个词长度为 27 个字母、第 2-5 个词长度为 24 个字母、第 6-10 个词长度为 23 个字母。但是这些词使用频率非常低，其中 3 个词出现过 2 次，其他的只出现过 1 次。

3.3 每册用词情况分析

下面是每册教材用词情况统计。

表2　教材每册用词情况统计

册号	普通班版				双语班版			
	词次	词种	词干种	新增词干种	词次	词种	词干种	新增词干种
3	2 242	1 444	1 246	1 246	1 594	973	837	837
4	4 651	2 452	1 996	1 475	4 217	2 281	1 835	1 494
5	7 823	4 077	2 966	1 797	6 720	3 557	2 790	1 853
6	8 052	3 783	2 941	1 367	7 140	3 524	2 726	1 327
7	13 668	5 967	4 395	1 940	12 530	5 745	4 299	2 056
8	11 778	5 582	4 190	1 428	11 591	5 566	4 213	1 578
9	14 166	6 517	4 863	1 630	14 645	6 618	4 917	1 661
10	1 8701	7 682	5 532	1 587	19 182	8 278	5 806	1 810
11	18 371	8 001	5 861	1 420	16 263	7 493	5 534	1 453
12	16 723	7 144	5 200	973	16 614	6 934	5 079	968
13	22 089	9 427	6 663	1 422	20 414	8 561	6 026	1 231
14	34 856	12 416	8 087	1 608	21 078	8 647	6 064	1 112
15	35 257	12 885	8 456	1 436	28 894	11 001	7 398	1 165
16	31 100	11 879	7 894	1 128	19 198	8 677	6 136	852
17	27 595	10 895	7 383	951	28 432	11 131	7 595	1 095
18	26 196	10 274	6 959	763	18 514	7 608	5 415	633
合计	293 268			22 171	247 026			21 125

表 2 显示，两套教材的词干学习从小学 2 年级第 3 册开始到初中二年级第 16 册基本呈现平缓增长的趋势。这种安排符合学生的学习规律和认知能力。

4 词频与词种分析研究

4.1 词种与频次的关系

频次与词种的关系能从一定角度反映词在教材中的使用情况，统计结果见表 5。

表 3 词频分布

频次段	词种数	占总词种数的比例(%)
1	26 405	38.28
2	18 210	26.4
3	4 999	7.25
4	4 119	5.97
5	2 276	3.30
6—10	5 741	8.32
11—20	3 440	4.99
21—100	3 153	4.57
>100	636	0.92
合计	68 979	100.00

表 3 显示，两套教材中，仅出现 1 次的词有 26 405 个，占总词种数的 38.34%，频次为 5（包括 5 次）以下的词为 56,009 个，占词种数的 81.20%，说明维吾尔语词种数量庞大，用词分散。频次为 6（包括 6 次）以上的词有 12,970 个、词干有 9,157 个，占总词种数的 18.80

%，其中频次为 11（包括 11 次）上词有 7 229 个、词干 5 541 个，占词种数 10.48%。

4.2 词种覆盖率分析研究

覆盖率是反映词语影响力的一个重要指标。研究结果见表 4。

表 4　不同覆盖率的词种数

覆盖率(%)	词种	频　次	占总词种数的比例(%)
50	1 487	270 146	2.16
60	3 031	324 191	4.39
70	6 076	378 218	8.81
80	12 574	432 240	18.23
90	28 762	486 265	41.70
95	42 269	513 279	61.28
96	47 367	518 682	68.67
97	52 770	524 085	76.50
98	58 173	529 488	84.33
99	63 576	534 891	92.17
100	68 979	540 294	100.00

表 4 显示，高频词中，覆盖所有语料的 50%、80%、90%时，分别使用的词种为 1,487 个、12,574 个、28,762 个。

4.3 词种分布分析研究

词种分布是指词种在课文中的分布状况。两套教材共有课文 1,270 篇，按课文数来统计词语的分布，可以说明词语的课文使用范围。分布最广的词语是**بىر**（壹，995）、**بىلەن**（与或和，958）、**بۇ**（这，907）；有 28,666 个词种只在一篇课文中出现。研究结果见表 5。

表 5　词的课文分布情况统计

课文数	词种	比例(%)
1	28 666	41.56
2—5	29 684	43.03
6—10	4 950	7.18
11—50	4 709	6.83
51—100	589	0.85
101—200	241	0.35
201—300	76	0.11
301—400	35	0.05
401—500	13	0.02
≥501	16	0.02

表 5 显示，分布在 51 篇课文以上的有 970 条词语，占词语总数的 1.41%。分布在出现 1 次的词语达 28,666 条，占词语总数的 41.56%，1~5 次的词语占词种数 84.59%。可见，教材词语数量大，但大多属低频词、课文分布数量低，词语使用分散。

5 结束语

一、就所调查的教材来看，维吾尔语语文教学从小学到高中共涉及 33,248 个词干，其中小初阶段学习词干 24,568 个，高中阶段学习词干 26,124 个。本次统计了小学、初中、高中教材总词条 109,043 种，其中提取小学、初中、高中教材总词干条 33,248 种，与在《中国语言生活报告 2009》发布的“现代维吾尔文网站用词调查” 中的总词干条 31,452 种比较，共用词干条 23,493 种，占整个网站总词干种的 74.70%、占小学、初中、高中教材总词干种的 70.66%。说明中小学教育阶段维吾尔语语文教学能满足网络媒体资料的阅读、内容理解、描述要求。

二、小初教材词干与高中教材词干中共用词干 17,444 个，占各自总词干的 71.00%、66.77%，可以看出高中教育是对小学、初中基础教育所学知识的巩固、延伸和扩展。

三、小初教材词长与高中教材词长相对稳定，词长分布比例以及频次分布比例基本一致，都是 7——10 个字母构成的词种最多，占全部词种的 53%-54%，其中多数是“词干+词尾”组成的黏合词语；5 字母构成的词使用频次最高，占总频次的 20%左右，体现出语言的省力原则。短词中多义词多，长词往往是单义词，表达意思比较完整，超越一般词语本身描述功能，体现出黏着语特点。

四、两套高中教材中，必修课教材从语料规模到词种、词干种数量都小于选修教材，体现出两套教材不同的教学目的，具有互补性。

参考文献

1、哈米提.铁木尔著 《现代维语语法》，民族出版社出版 ，1987 年 6 月。
2、阿不利孜.牙库甫等 《现代维语详解词典》，中央民族出版社，1991 年 11 月。
3、王铁琨 张普等， Language Situation in China:2005（下） The Commercial Press.
4、王铁琨 张普等， Language Situation in China:2006（下） The Commercial Press.
5、王铁琨 张普等， Language Situation in China:2007（下） The Commercial Press.
6、王铁琨 张普等， Language Situation in China:2008（下） The Commercial Press.
7、易坤琇 高士杰编著，维吾尔语语法 中央民族大学出版社 1998 年 2 月
8、程适良等， 现代维吾尔语语法 新疆人民出版社 1996 年 9 月

本文获得国家自然科学基金项目（批准号：61063036,61262066）、国家教育部社科基金（10YJA740121）、国家语委科研规划项目（YB115-38）、国家自然科学基金委重点项目（批准号：61132009）、国家语委“十二五”科研规划项目（YB125-45）、国家科技部科技支撑计划项目（2009BAH41B00）等项目的支持。

作者：艾孜尔古丽 助教，在读博士； 研究方向：自然语言处理；
李晓 研究员，博士生导师； 研究方向：自然语言处理；
通信作者：玉素甫.艾白都拉 教授，硕士生导师；研究方向：自然语言处理；
手机：13999272491