基于口语语料库的汉语口语自动化考试词表的研制[1][2]

张文贤 路云 李晓琪
北京大学对外汉语教育学院 100871
axxian@163.com, yunlu@pku.edu.cn, lixiaoqi@pku.edu.cn

摘要：口语词表研制是汉语口语自动化考试研发前期的一项重要工作。词表的研制完全建立在汉语口语语料库的基础之上，充分考虑了口语中的常用词汇与汉语学习者学习到的词汇这两方面的因素。依据频率与功能等原则，词表最终选词 5186 个。该词表几乎涵盖了所有的基础词汇，并突出了考察口语能力这一特点。

关键词：汉语口语考试 词表 语料库

A Corpus-Based Vocabulary Lists Study For Automated Test of Spoken Chinese

ZHANG Wenxian LU Yun LI Xiaoqi
School of Chinese as a Second Language, Peking University, 100871

Abstract: Vocabulary list compiling for the Spoken Chinese Test is an important job at the beginning of the whole test development. This study is based on a oral Chinese corpus, and fully considers the oral common vocabulary and vocabulary for learners of Chinese as a second language. The word list was built according to the selection principles of frequency and function, and contains a total of 5186 words, which almost covers all the Chinese basic words, especially those that can test oral competence.

Key words: Spoken Chinese Test, vocabulary list, corpus

0 引言

北京大学与美国培生集团 2010 年 6 月至 2012 年 8 月成功合作研发了汉语口语自动化考试。该考试摆脱了人工评分，考生使用电脑或者电话就可以参加考试，考试时间一般为半个小时，考试结束数分钟后，即可以得到成绩报告单。本文介绍的是考试研发前期的一项重要工作——词表研制。词表为试题编写提供词汇来源，控制试题的词汇难度。研制出专门的汉语口语自动化考试词表，才能更好地为试题编写服务，进而加强考试的信度。

汉语常用词词表的研究由来已久，已经得到的成果也较为丰富，但是现有词表与汉语口语自动化考试的目的不符，均不能满足本项目的需要。词表编制不是一劳永逸的工作，需要随着时代的变化而不断修改。汉语学界早期编制的词表已经受到了时代的限制，如1962年文字改革委员会汉字组编制的《普通话3000常用词表》，1964年北京语言学院编写的《外国学生用四千词表》，1981年的《外国人实用汉语常用词表》(3040词)、1986年的《现代汉语频率词典》(常用词为8548词)、1986年的《对外汉语教学常用词表》(4000词)，1991年的《北京口语调查》(常用词为6966词)， 1990年北京师范大学现代教育技术研究所的《中小

[1] 本词表是为了编制考题而研制的，考题中的词汇全部在词表中可以查到，没有超纲词汇。当然，考生的应答使用的词汇可能超出词表的范围，这时，由专门的 fixtrans（审定加词）人员将这些词添加到机器词典中，以便计算机能够识别。

[2] 感谢匿名审稿人所提出的修改意见，在此深表谢忱。

学汉语常用词表》等。这些词表都不同程度地包含了某些过时的词汇，而且都不是面向学习汉语的外国学生的。

针对不同的目的，可以编制出不同的词表。以使用对象为出发点，可以编制国别化词表，如甘瑞媛（2004）《国别化"对外汉语教学用词表"制定的研究：以韩国为例》（10052词），孙红（2009）《面向泰国汉语教学"国别化"词表的研制》(常用词为8107)。目前面向综合汉语水平考试的词表有1992年的《汉语水平词汇与汉字等级大纲》（8822词）、2009年《新汉语水平考试大纲》（5000词）。以某一阶段的教学为目的的词表有1999年杨寄洲主编的《对外汉语教学初级阶段教学大纲》(词汇大纲，北京语言大学出版社)。展现当今的词汇状况的词表有2005年至今每年发布的《中国语言生活状况报告》中公布的"报纸、广播电视、网络高频词语表"等。但这些词表也均不能完全适用于汉语口语自动化考试。2010年《汉语国际教育用音节汉字词汇等级划分》（北京语言大学出版社）虽然词汇量大，词汇统计来源也包括口语语料，但此词表的研制目的与汉语口语自动化考试不同，其语料来源较多，不限于口语，所以也不便参考。

为了体现口语考试的特点，我们需要研制出有针对性的词表。汉语口语自动化考试的服务对象是来自世界各地的汉语第二语言学习者，这就决定了词表的研制需要考虑两方面的因素，一是汉语口语中的常用词汇，二是汉语学习者学习到的词汇。也就是说，词表的制定要考虑两个差异：汉语口语与书面语的词汇差异；第二语言学习者与汉语母语者的口语差异。因此，在选词的时候要充分考虑语料来源、词汇覆盖率等问题。

1 词表编制的过程

口语考试词表由中方和美方共同构建。先由中美双方独立根据各自的语料提取出口语词表，然后将这两个词表进行合并。双方词表以及最终合并词表的收词情况详见表1。

表1 词表收词情况

北大词表	培生词表	只在北大词表中出现的词	只在培生词表中出现的词	北大和培生词表中都有的词	北大和培生合并的词表
4434	3349	1837	781	2568	5186

美方所使用的语料有电话录音Callhome，汉语频率词典*A Frequency Dictionary of Mandarin Chinese：Core Vocabulary for Learners*(2009).(Richard Xiao, Paul Rayson,&Tony McEnery,New York：Routledge)，以及两套口语教材：*Integrated Chinese*,(3rd Edition) Y. Liu et alia, Boston：Cheng & Tsui与*Chinese Link-level 1* (2010)*and Level 2* (2008)(2nd Edition). S.Wu, Y. Yu,Y. Zhang, & W. Tian. New Jersey：Prenticel Hall。除此之外，还另外补充了一些常见词缀，比如"第、初、老、子、头"等。也补充了一些常见词，比如1－100的汉语数字；时间表达，如"秒、秒钟、分、分钟、时、小时、星期、礼拜"等；常见的姓，如"李、王、张、刘、陈、杨"等；常见的国家名、城市名，如"日本、韩国、新加坡、加拿大、英国、法国、北京、上海、广州、重庆"等；一些不能分析的词，即语块，如"随着"。美方词表中的词至少要在两个语料来源中出现，最后经过汉语专家干预确定下来。

下面主要介绍中方的词表研发过程，整个过程分为以下几个步骤：

第一步，创立口语语料库[3]。中方构建的语料库包括美方提供的电话录音Callhome，北京语言大学研制的汉语口语语料库以及重要的国内外口语教材。所谓重要的教材首先是指在中国国内或者海外使用范围广泛、时间较长，已经得到对外汉语教育学界的普遍认可。其次是课文语言不能过于书面化。口语教材特别是高级口语教材的课文选择存在书面语化的现象。有些高级口语教材的课文来源是报刊文章或广播录音，属于典型的书面语，选文

[3] 本口语语料库的字数约在 160 万以上。

围绕一个话题展开，完全不考虑交际场景，缺少口语交际的特点，这样的教材我们不选取。在充分调查国内外口语教材以及征求对外汉语教学专家意见的基础上，我们最终选定的教材详见表2、表3。

表2 国内教材

序号	书名	作者	出版时间	出版社
1	《大家说汉语 初级、中级、高级汉语口语(韩文注释本)》	初级、中级：苏瑞卿 高级：李丛	初级、中级：2005 高级：2008	北京大学出版社
2	《新生活汉语：中级口语(上、下册)》	连吉娥	2005	北京大学出版社
3	《初级、中级、高级汉语口语》(北大版新一代对外汉语教材·口语教程系列)	初级：戴桂芙等 中级：刘德联等 高级：任雪梅等	2006	北京大学出版社
4	《中级汉语听说教程》	胡晓清	2006	北京大学出版社
5	《汉语初级、中级、高级口语教程（上、下册）》	杨寄洲 等	初级：2007 中级：2010	北京大学出版社
6	《魔力汉语 初级、中级汉语口语（上、下）（英日韩文注释本）》	初级：姚晓琳,何薇,林齐 中级：林齐倩,何薇	2008	北京大学出版社
7	《短期培训系列 短平快汉语：初级口语（1）、（2）》	（1）：张新明 （2）：王励	2009	北京大学出版社
8	《你说·我说·大家说》初级、中级、高级系列	郑国雄、陈光磊	2002	北京语言大学出版社
9	《实用汉语口语课本》	陈若凡等	2003	北京语言大学出版社
10	《汉语口语教程》	戴悉心，王静	2004	北京语言大学出版社
11	《发展汉语 初级、中级、高级汉语口语》	初级：陈晨 中级：路志英 高级：李禄兴	2006	北京语言大学出版社
12	《说汉语》	吴叔平等	2008	北京语言大学出版社
13	《汉语口语速成（基础篇、入门篇、提高篇）》	马箭飞等	2008	北京语言大学出版社
14	《很好——初级汉语口语》	刘颂浩等	2008	北京语言大学出版社
15	《汉语口语教程》	陈光磊	2000	北京语言大学出版社
16	《体验汉语口语教程》	陈作宏	2010	高等教育出版社

表3 海外教材

序号	书名	作者	出版时间	出版社
1	德国 *Chinesisch——sprenchen, lessen schreiben (Sprach-und Schriftübungsbuch)*	Hans—Christoph Raab	2002	Julius Groos Verlag Tübingen
2	《汉语900句》	国家汉语国际推广领导小组办公室	2007	外语教学与研究出版社
3	美国《中文听说读写》（*Integrated Chinese, 2nd Edition*） level 1—level 2 textbook（共5册）	Tao-chung Yao & Yuehua Liu	2006	Cheng & Tsui Company
4	澳大利亚《汉语》for beginning students（1册), for intermediate students（3册), for senor students（1册)	Peter Chang Alyce Macherras Yu Hsiu Ching	1992	Longman

第二步，把教材中的对话部分输入电脑，以WORD文档存储。需要说明的是，教材输入的只是课文中的对话部分，有些课文有叙述部分，不被采用。对于教材中生词表中的词汇不做收集。因为各个教材所列出生词表的标准不同，所以我们不使用原教材中的词表。

第三步，把以WORD形式存储的材料转换成TXT格式。

第四步，用北京大学计算语言学研究所汉语词语切分与词性标注软件，对课文进行分词。然后提取出高频词汇。

第五步，将提取出来的词汇按照使用频率排列，然后进行人工干预。我们请四位从事对外汉语教学多年的专家进行人工删除。首先请各位专家分别干预，根据教学专家的语感以及教学经验，标记出个人认为需要删除的词汇，任何过时的词或者非口语常用词或者非词都标记出来，之后将四位专家的表格整合为一张总表。然后举行一个讨论会，中方专家充分讨论所标记出来的词，最终决定该词是否做删除处理，从而对词表内容进行适当的调整。

2 词汇确定的原则

2．1 词汇收录标准

词表编制，先要定性，即词表要收哪些词以及怎样收词。在确定好如何收词之后，再确定词表收词的数量。本词表中的“词”是广义的词，不是语言学中严格意义上的“词”，而是一个在口语中能说、能被理解的语义理解和表达单位，并且一定要符合汉语口语语感习惯。本词表中的词既包括传统意义上的单个词，也包括由多个词构成的作为一个整体来使用的语块。

在确定哪些词入选词表时，首先依据的是频率标准。以北大词表为例，语料来源有四块：Callhome、北语口语语料库、国内教材、国外教材。按照频率标准排序自然有四种第一顺序，比如下面的表格是以国内教材（词频）为第一顺序排列的词表。当专家在进行干预时，也要适当考虑在四类语料中的词频，最少要在两类语料中出现频率较高。当所提取的词汇达到5509个词时，在Callhome、北语口语语料库、国外教材三类语料中的词频都为0了，所以干预之前的词汇数是5509。专家干预时，如果只在一类语料中出现频率稍高，而在其他三类语料中都频率较低，而且也不符合专家语感的话，就会被删除。专家干预之后

的北大词表含4434条词语。

表4 以国内教材（词频）为第一顺序排列的词表

词语	国内教材（词频）	Callhome（词频）	北语口语语料（词频）	国外教材（词频）
的	0.035623	0.016961	0.025402	0.026438
我	0.025823	0.024513	0.018106	0.028884
了	0.021108	0.015403	0.016896	0.019447
是	0.017083	0.008800	0.01843	0.015318
不	0.016428	0.007217	0.012674	0.01223
你	0.014745	0.020707	0.007539	0.020245
有	0.008757	0.004567	0.007953	0.007598
人	0.007535	0.001956	0.004831	0.00451
一	0.007421	0.002679	0.007561	0.005465
也	0.007197	0.005930	0.011645	0.004129
就	0.007122	0.011264	0.015935	0.004788
在	0.006424	0.005147	0.006306	0.008344
说	0.005917	0.005782	0.007558	0.003383
这	0.005801	0.005850	0.011904	0.004545
好	0.005705	0.003797	0.004307	0.010617

仅仅依靠频率来判断是否收词是不充分的，还要综合考虑必要性、难易度以及词语对语料的覆盖率。McCarthy（1999）提出口语交际基本词的9个范畴：modal items（情态词）, delexical verbs（虚化动词）, interactive words（互动词）, discourse markers（话语标记）, basic nouns（基本名词）, general deictic items（一般指示词）, basic adjectives（基本形容词）,basic adverbs（基本副词）, basic verbs（基本动词）。在确定北大词表的过程中，特别是在处理该将哪些语块收录到词表当中时，我们充分考虑了这9个范畴。

2. 2 对于特殊形式的处理

对于语块、重叠、儿化等具体的问题，我们做了如下处理：

如果两个词常常在一起出现，并且这两个词组合在一起又意义确定，或者具有一定的交际功能，就将之作为语块出现在词表中。这样的词有“挺好、打电话、打球、那天、十几、各种、各国、各人、各地、各位、慢慢来、是不是、你好、不能不、说真的、年前、这两天、男女平等、好吧、够了、中小学生、什么的、也就是说”等。两个单音节副词总是连用，如“才能、只能、再不”，也作为一个词处理。以“不”、“这”、“那”开头的词收录的比词典广泛，“不少、不够、不大、这样子、这时候、那时候、那是”等也作为词收到词表。从这些词中能较明显地反映出原始语料高频词的特点：意义明确，结构固定，凝固性强，口语化明显。

对于某些常用重叠词，比如“帮帮、个个、家家、轻轻、唱唱、吹吹、随随便便、早早、痒痒、歇歇、简简单单、远远、深深、咬咬、安安静静、干干净净、恰恰、含含糊糊、方方面面、次次、甜蜜蜜、冷冷清清、考考、痛痛快快”，尽管出现频率较高，但是仍然作为原词的变化形式，不作为一个词条，不收入词表。

对于儿化词，词表提取出来的词有两种情况，带“儿”或者不带“儿”，带“儿”不带“儿”没有意义差别，只是文字记录的问题，这就没必要列两个词条。处理这类词时，我们遵照原语料，原语料有“儿”，词表中就保留“儿”，原语料没有“儿”，不再补充上“儿”。比如：一点一点儿，一块一块儿，这边这边儿，有空有空儿，没事没事儿，那边那边儿，

那会那会儿，好玩好玩儿，面条面条儿，一会一会儿，字字儿，眼镜眼镜儿，男孩男孩儿，女孩女孩儿。

3 汉语口语自动化考试词表分析

3. 1 词类分析

表 5 词类分析[4]

词性	数量（个）
名词	2270
动词	1677
形容词	669
语块	380
副词	285
代词	87
其他（量词、数词、指示词、叹词、助词等）	301

从上表可以看出，词表中语块的数量是除了名、动、形这些实词外最多的一类，比如“总的来说、总而言之、只不过、这样子、也就是说、想不到”等等。如前所述，本词表中的“词”是广义的词，不是语言学中严格语法定义上的“词”，而是一个符合汉语口语习惯的语义理解和表达单位。这些语块带有浓重的汉语口语表达色彩，在我们所搜集的汉语口语语料库中出现的频率高且结构固定。将这些语块收入我们的词表，从一个侧面体现出汉语口语自动化考试与传统汉语考试的一个最大的不同，即侧重考察被试汉语口语的实际运用能力。

3. 2 词表中的反义词

汉语口语自动化考试在初期研发阶段有一个题型是“反义词”，即要求考生说出听到的反义词。这一题型的词汇全部出自词表。因此，这就要求词表有足够的反义词备选。词表中的反义词有 135 对，那么能够用于考题的就是 135×2＝270 对，比如“大－小”这对反义词，考试的时候既可以问“大”的反义词是什么，也可以问“小”的反义词是什么。这些反义词词对包括不少单音节反义词。通过口语自动化考试 A 阶段的测试，我们知道，单音节反义词的测试效果并不理想。比如当考生听到“nán”时，并不知道是要说“南”还是“难”的反义词。如果考题是“南边”，则答案会更清晰。因此，在考试研发的 B 阶段，我们取消了单音节反义词，全部变为双音节反义词。B 阶段研发结束后，我们对试题进行了重新评估，发现若去掉反义词后不会影响到考试的信度。因此在正式推出汉语口语考试时，未采用反义词这一题型。

3. 3 汉语口语自动化考试词表与新HSK词表的对比分析

3.3.1 重合词

通过与新HSK（5000词）词表的对比分析，可以看到，汉语口语自动化考试（5186词）中与新HSK考试完全重合的部分一共有2659个词，整体重合比例为53.2%。

表6 口语词表与新HSK考试词表的重合词

新HSK级别	重合数量（个）	新HSK词表数量（个）	重合比例
1-3级	592	600	98.7%
4级新增词	559	600	93.2%

[4] 词表词性统计中的兼类词，按其具体词性分别计数。

5级新增词	965	1300	74.2%
6级新增词	543	2500	21.7%
合计	2659	5000	53.2%

通过与新HSK每级词表的具体对比分析，可以看到，汉语口语自动化考试词表在新HSK1-5级中的重合比例都非常高，特别是在HSK1-4级中的重合比例都高于90%以上，在1-3级中的重合比例更高达98.7%，这也说明汉语口语自动化考试词表中几乎涵盖了所有的基础词汇。在HSK6级新增词中重合比例降低，从另一侧面体现了该考试考查口语能力的特点。

3.3.2 新 HSK 中有而汉语口语自动化考试词表中没有的词

表7 只在新HSK考试词表中出现的词

新HSK级别	数量（个）	举例
1-3级	8	打篮球、第一、黑板、踢足球、行李箱
4级新增词	37	放暑假、极其、性别、填空、做生意
5级新增词	324	不耐烦、彩虹、朝代、翅膀、充电器
6级新增词	1980	安居乐业、百分点、霸道、斑纹、报仇
合计	2259	

新HSK中有而汉语口语自动化考试词表中没有的词主要集中于HSK6级词汇当中。其原因主要有：

第一，在新HSK1-4级中收录的一些动宾结构词汇在汉语口语自动化考试中没有，如“打篮球”、“踢足球”、“做生意”，但汉语口语自动化考试词表可能已经分别涵盖了“动词”和“宾语”的部分，比如词表中含有“打”和“篮球”，但没有收录“打篮球”。

第二，新HSK6级中收录的部分成语词汇在汉语口语自动化考试中没有，如“安居乐业”、“半途而废”、“饱经沧桑”、“博大精深”等，这些成语因带有强烈的书面语色彩，与汉语口语自动化考试的测试目的不符，故在词表研制之初时就已经被排除在外，所以在该词表中没有体现。

第三，新HSK5-6级中部分词汇可能因在口语教材或其他口语材料中使用频率不高而没被收录在汉语口语自动化考试词表中，如“彩虹”、“朝代”、“充电器”、“斑纹”等。

4 遇到的问题

本词表涉及到汉语口语和书面语词汇的差异，而这方面的研究还处在初步探索阶段。在汉语本体研究当中，焦点多集中在对比口语与书面语的语体差异方面，而对汉语口语词汇的系统研究却一直比较薄弱。比如常敬宇（1986）将汉语口语词汇的特点总结为形象性、实感性、叠音词占很大比重，王福生（2002）提出口语和书面语词汇等级的划界问题，张文贤等（2012）分析了语体差异大的同义词等。正如曹炜（2003）在《现代汉语口语词和书面语词的差异初探》一文中指出的那样：“关于现代汉语口语词和书面语词的状况，我们所知极少。几乎所有的现代汉语教材对这两种词汇现象均不予讨论。我们甚至至今无法获知口语词和书面语词在音节结构、词义架构、内部构造上所具有的不同的或相同的特征，更遑论其他。”这也是我们在词表研制过程所需要克服的一个困难。所以，为了使词表能最大程度地真实反映汉语口语的特点，我们采用了口语语料库选词和人工干预相结合的筛选办法，严格把握语料来源，充分调研了词汇的覆盖率，打破了语言学上传统的“词”的界限，确保词表所选语汇充分体现汉语口语语体的特点。

本词表建立在具有相当规模的汉语口语语料库的基础之上，语料库可以便于研究者在大规模样本的基础上进行定量分析，以获得客观可靠、科学全面的结果。目前该口语语料库的组成部分主要有Callhome电话录音、北语口语语料库以及国内外教材。口语语料库的

收集和建设是一个长期的过程，语料库越大，可以提取的词就越多，那么也就能更加真实地描写汉语口语词汇系统的全貌，从而使词表内容更加精准。若进一步加强汉语口语语料库建设，则能使词表更加精准和丰富。

在汉语口语自动化考试词表的研制过程中，还存在词性确定和词性标注的问题。比如目前的词表同音词和多义词只按照书写形式并为一个词条，在今后的词表制定中，应对此进行改变。如“只”在词表中只作为一个词条收录，而其有量词和副词的用法，读音也不相同。今后我们将按照其用法和读音对其进行区分。

参考文献

[1] David Y. W. Lee . Defining Core Vocabulary and Tracking Its Distribution across Spoken and Written Genres: Evidence of a Gradience of Variation from the British National Corpus. Journal of English Linguistics, 2001(29 / No. 3)： 250–278

[2] Junko Shirato and Paul Stapleton. Comparing English vocabulary in a spoken learner corpus with a native speaker corpus: Pedagogical implications arising from an empirical study in Japan. Language Teaching Research,2007(11): 393–412

[3] McCarthy. M. What constitutes a basic vocabulary for spoken communication? Studies in English Language and Literature, 1999(1)：233–249

[4] Wallace Chafe and Deborah Tannen . The Relation between Written and Spoken Language. Annual Review of Anthropology, 1987(16)：383–407

[5] 曹炜. 现代汉语口语词和书面语词的差异初探.语言教学与研究，2003（6）：39－44

[6] 常敬宇. 汉语口语词汇的特点.逻辑与语言学习，1986（4）：36－27

[7] 甘瑞媛. 国别化“对外汉语教学用词表”制定的研究：以韩国为例，北京语言大学博士学位论文. 2004

[8] 马清华. 唯频率标准的不自足性——论面向汉语国际教育的词汇大纲设计标准. 世界汉语教学，2008（2）：119－134

[9] 苏新春. 对外汉语词汇大纲与两种教材词汇状况的对比研究. 语言文字应用，2006（2）：103－111

[10] 孙红.《面向泰国汉语教学“国别化”词表的研制》，暨南大学硕士学位论文，2009

[11] 王福生. 对外汉语教学活动中口语和书面语词汇等级的划界问题. 国际汉语教学学术讨论会论文集（赵金铭主编），北京：北京大学出版社，2002：47－59

[12] 张文贤.邱立坤.宋作艳.陈保亚. 基于语料库的汉语同义词语体差异定量分析. 汉语学习. 2012（3）：72－80