语料库技术在专四专八阅读中的应用—以词汇为例

严振羽
（华北理工大学 外国语学院，河北 唐山 063210）

摘要：本篇研究将会借助日益成熟的语料库技术，借助 range 词汇分析软件中内置的两类词表对专四专八考试中的阅读材料进行分析，观察专四专八阅读中词汇的使用特点，得出应对专四专八阅读时需要的词汇量。本研究不仅可以帮助学生备考专四和专八考试，同时在提高学生学术词汇意识和学术词汇积累方面也有推动作用。

关键词：range 软件；语料库；专四专八；词汇分析

国外研究学者提出根据阅读任务确定词汇量需求的观点，专四专八考试作为英语专业的学生必须参加的两门等级考试，分数高低往往用来衡量英语专业学生的英语水平，因此，努力在考试中获取更高分数对他们来说至关重要，专四专八考试中，阅读部分的分值比重较大，为了在阅读部分获取更高分数，学生要具备理解阅读材料所需的词汇量。词汇量的确定主要是根据文本内容的词汇覆盖率决定的。词汇覆盖率是成功阅读理解的第一道门槛。专四和专八考试阅读材料的难度不同，那么构成两类考试阅读材料的词汇在选择上的特点也不尽相同。

一、理论框架

（一）词汇覆盖范围

词汇覆盖范围是制定和评估单词列表的一个基本概念。它指的是读者所知道的文本中单词的百分比（Nation，2006）。研究发现了成功的阅读理解的词汇覆盖率。Laufer（1989）发现，需要读者了解文本 95% 的词汇，以确保对文本的合理阅读理解。Hu 和 Nation（2000）研究得出结论，98% 是充分理解的必要的词汇覆盖率。Schmitt（2011）在最近的一项研究中，也认为 98% 是充分理解学术文本的最佳覆盖率。

（二）Range 词汇分析

本研究使用的语料库分析软件是 RANGE。RANGE 软件作为一个语料库工具，其基本原理是把研究对象文本与公认的权威词汇表进行对比，基于文本的测试方法和词频的词汇量化统计分析得出一组数据，显示出对象文本中出现在词汇表中的词汇及其出现的频率以及占总词汇的比率，通过分析这些数据，可以得出不同文本中的共有词汇以及各自的词汇使用情况。

（三）Range 词表

RANGE 内置三个基础词：BASEWORD1、BASEWORD2、BASEWORD3。BASEWORD1 包含约 1000 个最常用的词族，BASEWORD2 包含约 1000 个次常用词族， BASEWORD3 包括前两个词表之外的、高中的和大学各科教材中最常用的学术词汇。这三个基础词表包括基本单词和衍生单词，词族（Families）是指中心词加上其屈折形式和衍生形式，类符（Type）是指单词计数单位，形符（Token）是指的是单词的标记例。例如：“To be or not to be，that is a question”中，共有 7 个词族，8 个类符，10 个形符。

其次是 Nation 根据英语国家语料库（BNC）编制的 14 个词表。在 1 ~ 14 词表中，每个词表包括 1000 个词族。第 1 个词表为词频最高、词分布最广的 1 000 个词族，其他 13 个词表中的词族其频率和词分布依次降低。针对二语学习者，认识文本中 95% 的词汇，基本能保证阅读理解，如果想充分理解，则需要认识 98% 以上的词汇。

二、研究方法

（一）研究内容

本篇研究将会以专四专八考试的阅读材料作为 range 软件的分析材料。笔者将会选取从专四专八考试改革后 2016–2019 年的阅读考试材料，共 24 篇阅读文本，自建语料库 C1 和 C2，C1 包含专业八级考试的阅读材料，C2 包含专业四级考试的阅读材料。分析结果将借助图表展示。

（二）研究问题

通过 range 软件对两个语料库的文本内容进行分析，旨在回答两个问题：

a：学生应对专四专八阅读文本需要多大的词汇量？

b：专四和专八阅读材料在词汇使用方面有什么特点，有什么差异？

（三）研究过程

数据处理之前要进行一系列的预处理，首先是把我们找到的阅读材料进行文本转换，转换成文本 txt 格式，方便 range 软件的处理，其次，把文章标准化，删除每篇文章中的数字、图表、书目、图表和其他一些组件，以消除与词汇分析无关的因素，并确保存储在语料库中的文章由计算机软件可读，以减少错误。

三、研究结果与分析

根据表 1 我们可以得出，基本理解专八的阅读材料，需要达到 95% 的词汇认识率，因此，由图表得出具备 7000 词族的 94.52% 再加上专有名词等的 1.28%，在专八考试的阅读文本中的覆盖率达到了 95.8%，如果充分理解阅读材料，学生要具备 14000 词族的 96.25% 再加上专有名词等的 1.28%，才可以达到专八阅读文本的 98.8% 的覆盖率。

表 1 Range 对 C1 的词汇分析结果

WORD LIST	TOKENS/%	TYPES/%	FAMILIES	字符百分比累加（不包含 / 包含专有名词和感叹词）
1	8828/78.23	1424/44.40	783	78.23/79.51
2	868/ 7.69	597/18.62	451	85.92/87.2
3	352/ 3.12	250/ 7.80	215	89.04/90.32

4	290/ 2.57	202/ 6.30	170	91.61/92.89
5	143/ 1.27	118/ 3.68	109	92.88/94.16
6	84/ 0.74	63/ 1.96	55	93.62/94.9
7	101/ 0.90	64/ 2.00	59	94.52/95.8
8	49/ 0.43	37/ 1.15	36	94.95/96.23
9	38/ 0.34	33/ 1.03	33	95.29/96.57
10	33/ 0.29	31/ 0.97	30	95.58 /96.86
11	25/ 0.22	24/ 0.75	24	95.8/97.08
12	19/ 0.17	17/ 0.53	17	95.97/97.25
13	17/ 0.15	6/ 0.50	15	96.12/97.4
14	15/ 0.13	13/ 0.41	11	96.25/97.53
15	143/ 1.27	87/ 2.71	87	97.52/98.8
16	1/ 0.01	1/ 0.03	1	
not in the lists	278/ 2.46	230/ 7.17		

根据表 2 我们可以得出，基本理解专四的阅读材料，需要达到 95% 的词汇认识率，因此，由图表得出具备 4000 词族的 94.23% 再加上专有名词等的 1.12%，覆盖率达到了 95.35%。如果想充分理解文本的话，9000 词族的 96.95% 再加上专有名词等的 1.12%，可以达到专四阅读文本的 98.07% 的覆盖率，因此，要想完全理解专四阅读的文本材料所需要的词汇数量是 9000。

表 2 Range 对 C2 的词汇分析结果

WORD LIST	TOKENS/%	TYPES/%	FAMILIES	字符百分比累加（不包含 / 包含专有名词和感叹词）
1	6583/81.72	1142/53.87	693	81.72/82.84
2	600/ 7.45	379/17.88	308	89.17/90.29
3	282/ 3.50	180/ 8.49	152	92.67/93.79
4	126/ 1.56	92/ 4.34	85	94.23/95.35
5	72/ 0.89	52/ 2.45	47	95.12/96.24
6	31/ 0.38	30/ 1.42	29	95.5/96.62
7	73/ 0.91	37/ 1.75	31	96.41/97.53
8	30/ 0.37	19/ 0.90	19	96.78/97.9
9	14/ 0.17	13/ 0.61	13	96.95/98.07
10	9/ 0.11	9/ 0.42	9	97.06/98.18
11	20/ 0.25	14/ 0.66	14	97.31/98.43
12	4/ 0.05	4/ 0.19	4	97.36/98.48
13	13/ 0.16	8/ 0.38	8	97.52/98.64
14	6/ 0.07	5/ 0.24	5	97.59/98.71
15	90/ 1.12	53/ 2.50	53	
16	0/ 0.00	0/ 0.00	0	
not in the lists	103/ 1.28	83/ 3.92		
Total	8056	2120	1470	

根据表 3 我们可以得出，专八考试的阅读文本材料里，共出现 3207 个词，其中 1375 个词属于基础词表 1，约占总词汇的 42.87%，451 个类符属于基础词表 2，约占总词汇的 14.06%；352 个类符属于基础词表 3，约占总词汇的 10.98%；另有 1029 个词在 3 个基础词表之外，约占总词汇的 32.09%。

表 3 对 C1Range 统计结果

WORD LIST	TOKENS/%	TYPES/%	FAMILIES
one	8713/77.22	1375/42.87	776
two	646/ 5.72	451/14.06	354
three	540/ 4.79	352/10.98	258
not in the lists	029/32.09	1385/12.27	1385/12.27
Total	Total	3207	1388

根据表 4 我们可以得出，专四阅读材料的词汇 Range 统计结果表明，专四阅读试题中共出现 2120 个词，其中 1141 个属于基础词表 1，约占总词汇的 53.82 %；302 个类符属于基础词表 2，约占总词汇的 14.25%；156 个类符属于基础词表 3，约占总词汇的 7.36%；另有 521 个词在 3 个基础词表之外，约占总词汇的 24.58%

表 4 对 C2 Range 统计结果

WORD LIST	TOKENS/%	TYPES/%	FAMILIES
one	6627/82.26	1141/53.82	706
two	441/ 5.47	302/14.25	249
three	249/ 3.09	156/ 7.36	125

关于应对专四专八考试所需要的词汇量，专八考试的要求的词汇储备是最好是 10000，并且不能少于 8000，由此看来，8000 词汇的词汇量要求可以基本帮助我们理解文本材料，要完全理解阅读材料，学生要努力扩充自己的词汇量，达到 13000 词左右。

专四考试的要求的词汇储备是不能少于 8000 词，根据分析结果，8000 的词汇量可以基本帮助我们理解文本材料，如果要完全理解文本的含义，学生要达到 9000 词左右的词汇储备。

专八阅读的词汇组成在数量上明显高于专四阅读，其次大多数词汇主要是普通词汇，但是学术词汇也被广泛使用，所谓学术词汇就是指常出现在各学科学术文章中的词汇，因此，我们要注意提高自己的词汇量，还有学术词汇的积累，平时多读一些科学类的文章，而非仅仅是出现在阅读文本和教科书里的普通英语词汇。

四、总结

本研究只选取了专四专八改革后 2016–2019 年的阅读文本进行文本分析，语料库方面还有待进一步的充实，但是通过这个研究，我们可以得出，学生在备战专业四级和专业八级考试时，除了具备考试要求的基本词汇量，还要努力拓展词汇范围和词汇数目，以轻松应对专四专八的阅读部分。其次，专四专八考试涉及学术词汇的选择，因此，学生在日常学习中要多多加强科学类学术类文本的阅读，更多积累学术词汇。另外老师在开展英语教学课程之前有意识地为英语教学准备相关的学术单词列表，提高学生对学术词汇的学习意识。

参考文献：

[1] 鲍贵，王霞 .RANGE 在二语产出性词汇评估中的应用 [J]. 外语电化教学，2005（8）：54–58.

[2] 程实 . 语料库工具 Range 在文体研究中的应用 [J]. 语文学刊 · 外语教育教学，2019（9）：42–46.