Overseas English Testing: Pedagogy and Research 国外英语考试教学与研究, 2021, 3(2), 71-91
Published Online May 2021 in Hans. http://www.hanspub.org/journal/oetpr
https://doi.org/10.12677/oetpr.2021.32010

Hans汉斯

基于语料库工具AntConc对学生雅思写作词汇丰富性和搭配强度研究

李丰贤

广州新东方学校，广东 广州
Email: lifengxian5@xdf.cn

收稿日期：2021年5月1日；录用日期：2021年5月24日；发布日期：2021年5月31日

摘　要

过去，语料库曾经是少数专业人士如语言学或语言测试学专家使用的工具；但现在随着电脑技术的广泛普及和教学理念的更新，基于语料库的研究方兴未艾。其中，语料库软件AntConc的基本功能使得量化学生语言水平变为可能。本文旨在探索语料库建设的方法，主要从词汇和搭配两个方面，对学生的写作能力进行数据分析和给学生提出可量化的评估建议。

关键词

语料库，语言测试，写作，词块

A Corpus-Based Evaluation of IELTS Test-Takers' Writing Lexical Range and Collocational Strength via AntConc

Fengxian Li

Guangzhou New Oriental School, Guangzhou Guangdong
Email: lifengxian5@xdf.cn

Received: May 1st, 2021; accepted: May 24th, 2021; published: May 31th, 2021

Abstract

Corpus, a one-time unfamiliar concept restricted to only a certain number of professionals, has been gaining mounting momentum and revolutionizing the linguistic and education field, with the advent

文章引用: 李丰贤. 基于语料库工具 AntConc 对学生雅思写作词汇丰富性和搭配强度研究[J]. 国外英语考试教学与研究, 2021, 3(2): 71-91. DOI: 10.12677/oetpr.2021.32010

of computer technology and new teaching methods. This paper aims to explore the methodology of corpus construction for the purpose of analytical studies of students' writing competence mainly in terms of vocabulary and collocation. Basic elements like the concordance and cluster sections of corpus software AntConc demonstrate the possibility to quantify the performance of students' language proficiency. The figures and statistics will be employed for further interpretation and suggestions.

Keywords

Corpus, Language Tests, Writing, Collocation

1. 绪论

1.1. 研究背景目的

自上世纪 90 年代开始，语料库语言学以及电脑技术蓬勃发展。语料库曾经是一片无人熟知的领域，而现在无论在语言学还是外语教学，语料库都占有一席之地，且势头越演越烈。语料库研究属跨学科的研究，涉及到语料库的建库方法以及学习者语料库的分析研究，其理论基础涉及到语言学和教学法的相关理论。如今，现在便捷的电脑技术，辅以语料库相关知识和软件，可以使任意一名普通的语言研究者能够更好地量化教学效果。借助语料库他们可以对学生的问题有更全局的了解，以量化的方式评估和测量。值得一提的是，基于语料库的研究和应用对外语教学带来了许多革新性的影响[1]。本文旨在讨论语料库的建库方法，分析学生在英语书面语的表现情况，基于语料库数据分析来提供相应的解决方案。

1.2. 语料库的定义

从字面上来看，语料库“corpus”一词在拉丁文原意为“主体，躯体”。如今 corpus 指的是经文本处理，以电子形式储存的大型文本库。值得一提的是，语料库的文本是根据特定的采样标准，为了达到特定目的而建设[2]。通过对这些样本的分析研究，本文旨在揭示目的语言的内在规律特点，而这些特点都是可延伸，可扩展到语言本体的。

总的来说，语料库软件是教师分析语言的利器。比如说单语语料库软件 AntConc。AntConc 由日本早稻田大学的 Laurence Anthony 教授开发。AntConc 原本只运行于 Windows 和 Mac 系统，用于简单共现(concordance)。现已发展为世界知名的多功能语料库软件。[3]

2. 语料库在写作研究的具体应用

2.1. 关于自建语料库

2.1.1. 如何建立学生语料库

明确学生的目标

根据目前现状来看，学生年级、背景以及相应的英语写作水平差异化较大，面临的考试也比较分散。从大学四级考试，专业四级考试，或者是雅思考试都有一定比例。虽然各类标准化考试的写作评分标准有不少重叠，但是为了得到更准确的数据，必须制定详细的标准衡量学生的作文。在本文中，雅思写作考试的 TASK 2 被选为学生的目标，也就是说本次研究的语料库是以雅思写作 TASK 2 为导向的。

准备和收集

学生采样材料可以来自他们日常的练习，考试的试卷等，主要分为两大类别。一是他们之前产出过的材料，比如说写过的英语短文、英语作业等。另一方面则是由导师根据一定要求加以引导所产出的材料。

根据标准的流程，学生会收到导师事先布置的写作题目，然后将在上交期限内以电子文档的形式提交相关的作文。具体的流程如下：

1) 通过微信建立起师生交流群；
2) 布置雅思写作题目并设置期限；
3) 收集并组织学生的上交材料；
4) 用这些材料建立相关语料库；
5) 利用 AntConc 和 Treetagger 等软件分析文本特征。

随着现代科技的蓬勃发展，远程教育和在线课堂已经成为常态。考虑到学生和受众的地理差异性，通过社交软件来建立联络机制是较为优先的选择。微信的普及使教师和学生沟通更加方便，高质量的语言信息简化了交流互动繁琐的过程。通过该有效便捷的社交软件，每位成员可以保持联络，处于活跃状态中。

另一方面，电子邮件也是另一关键媒介。和微信的两大特点——随意性以及信息过分泛滥性相比，电子邮件更加正式，能够有效地传播严肃的信息，如写作要求，上交期限等。而且邮件的附件可以传送更多信息比如说 PDF 格式的作文等。当然，每位学生的作文都必须通过邮件上交，要求是以 Word 文档的格式以便收集。

Word 文档向来以编辑方便和易于标记而著称。然而作为本次研究关键点的 AntConc 只能和 txt 文档兼容，也就说所有的 word 文档必须转换为 txt 文档。大多数情况下，txt 文档可以轻松地消除 word 文档中的冗余信息，格式不统一等情况。

在转换文档的过程中，文档格式错误是值得注意的问题。尤其是汉语和英语的符号混用问题必须小心处理，文档的格式错误会损害最终数据的准确性。在作文数量不大的情况下，人工剔除和转换是可接受的。但在应对规模较大的语料库时，教师可以借助风林开发的 *TextEditor* 软件来处理(图 1)。

Figure 1. The basic interface of TextEditor
图 1. TextEditor 的基本界面

通过 *TextEditor* 轻松点击鼠标便可解决文本中信息冗余问题。再者，在实际运用中，一个完备的语料库和精密编码和标记密不可分，同时这些编码和标记也是作文进一步分类的前提。在本研究中所采用的编码系统是 *Treetagger* 的编码系统(表 1)。[1]

Table 1. A list of parts of speech with their tags in Treetagger [4]
表 1. Treetagger 中的各类词性及标记[4]

POS Tag	Description	*and, but, or, &*
CC	coordinating conjunction	*1, three*
CD	cardinal number	*the*
DT	determiner	*there is*
EX	existential there	*d'œuvre*
FW	foreign word	*in,of,like,after,whether*
IN	preposition/subord. conj.	*that*
IN/that	complementizer	*green*
JJ	adjective	*greener*
JJR	adjective, comparative	*greenest*
JJS	adjective, superlative	(1),
LS	list marker	*could, will*
MD	modal	*table*
NN	noun, singular or mass	*tables*
NNS	noun plural	*John*
NP	proper noun, singular	*Vikings*
NPS	proper noun, plural	*both the boys*
PDT	predeterminer	*friend's*
POS	possessive ending	*I, he, it*
PP	personal pronoun	*my, his*
PP$	possessive pronoun	*however, usually, here, not*
RB	adverb	*better*
RBR	adverb, comparative	*best*
RBS	adverb, superlative	*give up*
RP	particle	*?, !, .*
SENT	end punctuation	*@, +, *, ^, |, =*
SYM	symbol	*to go, to him*
TO	*to*	*uhhuhhuhh*
UH	interjection	*be*
VB	verb *be*, base form	*was|were*
VBD	verb *be*, past	*being*
VBG	verb *be*, gerund/participle	*been*
VBN	verb *be*, past participle	*is*
VBZ	verb *be*, pres, 3rd p. sing	*am|are*
VBP	verb *be*, pres non-3rd p.	*do*
VD	verb *do*, base form	*did*
VDD	verb *do*, past	*doing*

Continued

VDG	verb *do* gerund/participle	*done*
VDN	verb *do*, past participle	*does*
VDZ	verb *do*, pres, 3rd per.sing	*do*
VDP	verb *do*, pres, non-3rd per.	*have*
VH	verb *have*, base form	*had*
VHD	verb *have*, past	*having*
VHG	verb *have*, gerund/participle	*had*
VHN	verb *have*, past participle	*has*
VHZ	verb *have*, pres 3rd per.sing	*have*
VHP	verb *have*, pres non-3rd per.	*take*
VV	verb, base form	*took*
VVD	verb, past tense	*taking*
VVG	verb, gerund/participle	*taken*
VVN	verb, past participle	*take*
VVP	verb, present, non-3rd p.	*takes*
VVZ	verb, present 3d p. sing.	*which*
WDT	wh-determiner	*who, what*
WP	wh-pronoun	*whose*
WP$	possessive wh-pronoun	*where, when*
WRB	wh-abverb	*;, -, --*
:	general joiner	*$, £*
$	currency symbol	*and, but, or, &*

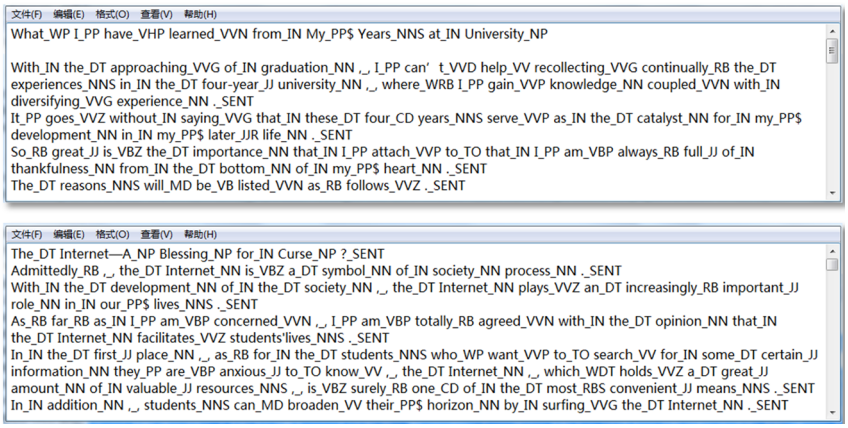

Figure 2. Samples from students' essays (tagged)
图 2. 学生作文样本(已标记)

通过借助精密的编码标记系统，教师可获取一系列附有词性标记的文本材料，如图 2 所示。在接下来的章节中我们会详细讨论和研究。总结起来，学生语料库由两部分组成：一是经过精确处理的电子文本，二是综合的编码标记。

2.1.2. 如何建立范例语料库

从技术的层面上来说，范例语料库的建立方法和学生作文语料库基本一致，而最根本的区别在于范例语料库的文本来源。鉴于学生语料库是以雅思为导向的，几个关键的范例文本来源展示如下。

广义的范例语料库

理论上，几乎前文提及的所有通用语料库都可作为学生打磨语言的良好工具。事实上，另一广义上的语料库则是谷歌搜索引擎。谷歌搜索引擎每天都触及到虚拟世界中数以百万计的文本，而这些文本充满了最紧贴时代潮流的语言和表达。在这信息的汪洋大海中也不乏权威新闻文本的身影。所以谷歌可以说是现今最大的原始语料库。虽然如此，由于本研究主要以雅思为导向，为了更好的获取相关信息，教师需要一个更准确和相对限制性的框架来指导我们范例语料库的建立。

狭义的范例语料库

在建立狭义的范例语料库时，教师可以首先考虑剑桥雅思的官方出版物——《剑桥雅思考试全真试题集》系列。主要原因是这个系列的材料是雅思考试典型的阅读材料。该部分材料在提升参加考试者的读写能力上有着不可替代的作用，尤其因为其文章经过精挑细选，质量极高，因此在范例语料库占有一席之地。

此外，其他补充材料包括高质量的新闻杂志如《经济学人》、《时代杂志》等。其中《纽约时报》的社论栏目“*Room for Debate*”值得一提。该栏目中绝大部分的社评高度凝练，围绕中心，平均字数为280 词。相比较而言，雅思写作 TASK 2 主要是 250 词以上的议论文。鉴于两者题材和字数上的相似程度，将“*Room for Debate*”收录在范例语料库内也很有必要。

精心打造好学生语料库和范例语料库之后，基本准备工作已完成。接下来本文主要讨论基于语料库详细的操作运用。

2.2. 语料库的操作运用

2.2.1. 分析验证学生语料库中的搭配

词汇丰富度

本文在先前的章节已介绍了语料库的基本构造。根据雅思写作 TASK 2 的写作评分标准，“词汇使用”是四项评分的一大关键，其他三项分别为“任务完成度”、“连贯与衔接”、“语法准确性和多样性”。[5]通过已标记的学生文本，教师可清楚地看到他们的词汇使用偏好以及作文中相应的瑕疵。接下来我们提取三位学生在讨论禁止手机的写作话题时所使用的动词列表作为对比研究对象。以下是如何利用 AntConc 提取相关动词列表的步骤：

1) 用 AntConc 读取已标记的学生文本(图 3)；

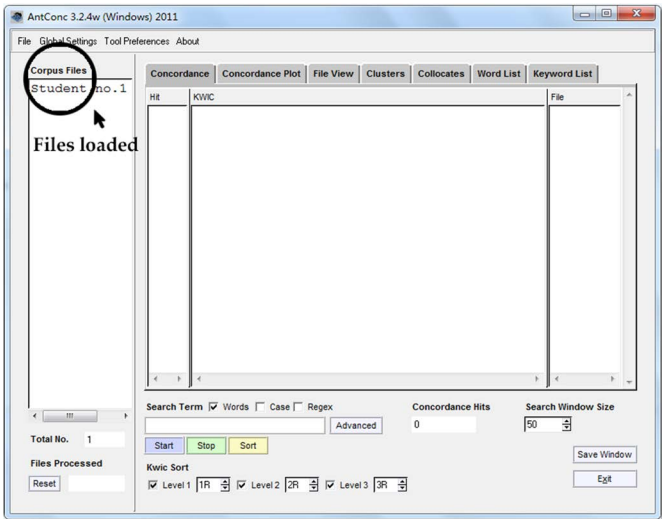

Figure 3. The Concordance interface with loaded files in Antconc
图 3. AntConc 读取文本后的基本界面

2) 选择 AntConc 的“词簇(Cluster)”板块(图 4);

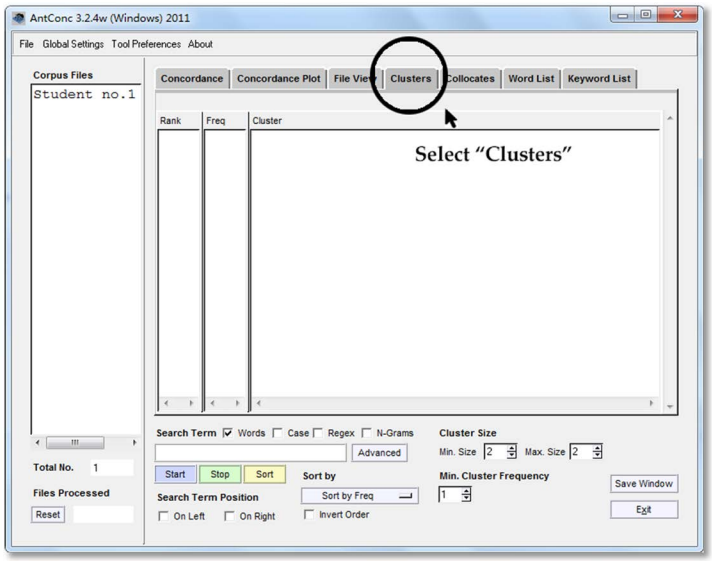

Figure 4. The Cluster interface in Antconc
图 4. AntConc 的“词簇(Cluster)”界面

3) 调整右下方的“最大词簇量(the maximum cluster size)”为 1 (图 5);

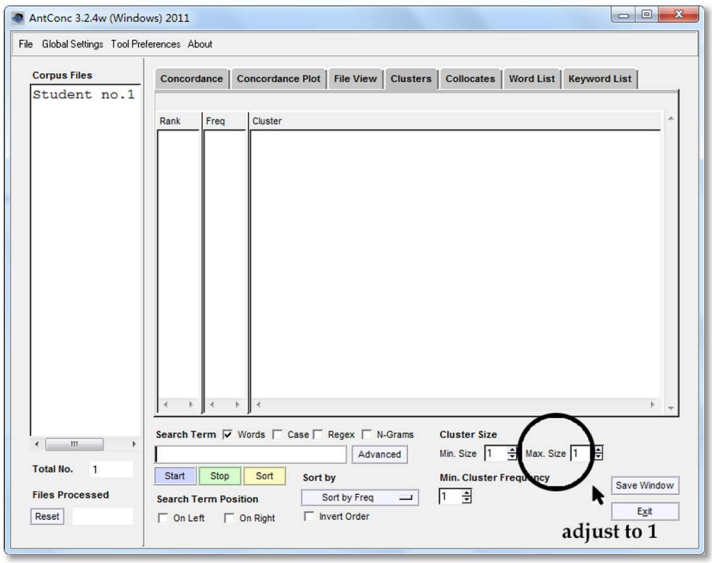

Figure 5. The Cluster interface with maximum cluster size as “1” in Antconc
图 5. AntConc 的“词簇(Cluster)”调整后界面

4) 在搜索栏输入代码“*_VV*”(意为“任何形式的动词”) (图 6);

5) 点击开始，生成词表(图 7)。

经过五步后，教师就成功获得了一份半成品，但还有一些数据有待处理。因为词表中含有屈折词根的词和该词的原型会被处理为两个不同的词，如“caused”和“cause”被分开统计。所以教师需要合并同类项，保证数据的精准度。最终三位学生的动词词表如下表 2。

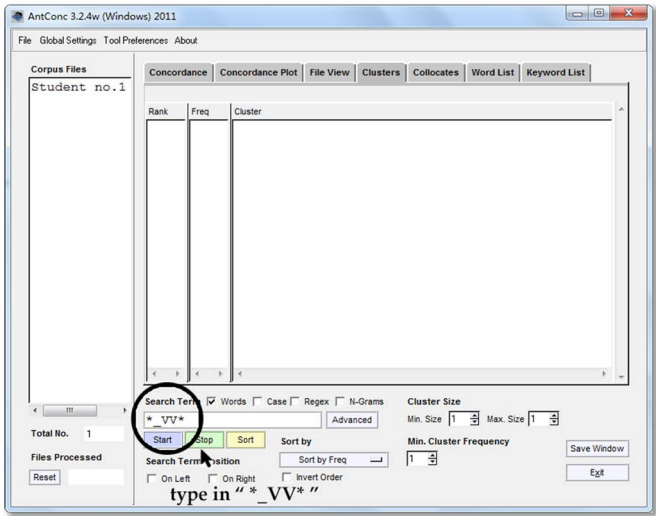

Figure 6. Input the code "*_VV*" in the search bar of Antconc
图 6. 在 AntConc 的搜索栏输入代码 "*_VV*"

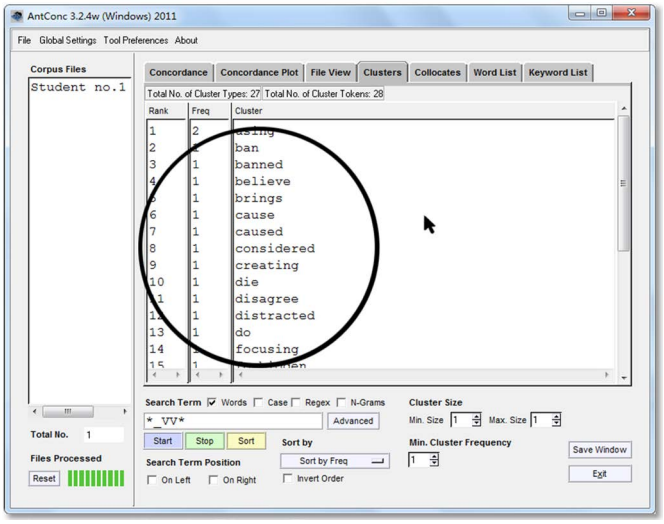

Figure 7. The raw verbal list of student no.1
图 7. 一号学生的原始动词词表

Table 2. Excerpts of verbal wordlists of three students
表 2. 三位学生动词词表节选

1			2			3		
rank	freq	words	rank	freq	words	rank	freq	words
1	5	take	1	3	use	1	2	take
2	5	ban	2	2	ban	2	1	annoy
3	3	want	3	2	cause	3	1	characterize
4	2	make	4	2	lead	4	1	demonstrate
5	2	leave	5	2	treat	5	1	detest
6	2	get	6	1	believe	6	1	encapsulate
7	2	find	7	1	bring	7	1	erupt
8	2	do	8	1	consider	8	1	exert

Continued

9	2	bring	9	1	create	9	1	impede
10	2	benefit	10	1	die	10	1	imprison

如图所示，三个表格中的最大动词词频依次从 5 降至 2。细节上来看，一号学生倾向于重复使用单音节词如“take”、“ban”、“want”、“make”等，最高频率高达 5 次。相对比而言，二号学生喜欢使用“use”、“ban”、“cause”等词，但最高频率只有 3。另外，三号学生多采用多音节词如“annoy”、“characterize”以及“demonstrate”等，但最高频率只有 2。

从语言学的角度来看，学生的心理词汇量(mental lexicon)存在着较大差异。而一个学生的最大词频数可以作为衡量用词是否多样的重要指标。换言之，作文中重复出现的表达和词语可视为学生的用词偏好，同时也可作为词汇匮乏的证据，上面一号学生和二号学生就是如此。虽然学生间的滥用词汇有不少重叠，但在整个词汇的分布序列中仍存在不同的用词特点。

寻觅同义词

教师可以鼓励并引导学生学会辨别作文中存在的滥词，然后在语料库中搜寻合适的同义表达，进而创造出符合个性化发展需求的词汇多样性加强体系。此外，这种操作可以避免学生被动接受老师准备的表达和用词，减少填鸭式的灌输。

COCA 语料库在该方面有大量语料，教师可以提供多样化的同义表达。因此，学生可根据自己的滥词在 COCA 中搜寻相应的同义替换。以“ban”一词为例，我们在 COCA 搜索栏输入代码“[=ban]”即可获得一系列同义词，如下图 8 所示。获得相关信息后，教师也需要对学生进行正确引导，和学生强调并非所有同义词都可直接替换，必须浏览其相关语境综合考虑后方可使用如，下图 9 所示。

Figure 8. The search results for the synonyms of “ban”
图 8. “ban”同义词在 COCA 语料库的搜索结果

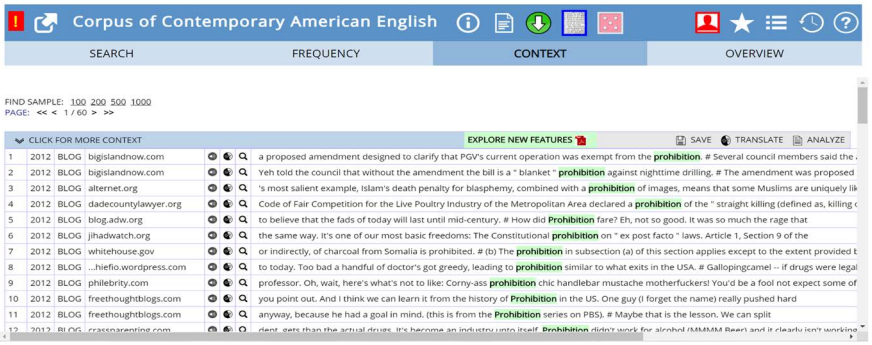

Figure 9. The context of the word “prohibition” [6]
图 9. “prohibition”一词在 COCA 语料库的语境共现[6]

挖掘特色措辞并验证其搭配的地道性

另一方面，除了关注高频词汇和同义词，文章的搭配也是决定语言质量的关键。而一个人措辞中最有特色的部分往往落在较低频的词汇上，在一篇 250 词的作文尤其如此。接下来我们分析四号学生的特色措辞及其搭配(图 10~12)。

在 AntConc 的共现界面中我们看到“assemble”以动名词的形式和“family”搭配，“elevate”和“communication”搭配(图 13)。

但这两个搭配在自建的范例语料库以及大型通用语料库如 COCA 中并未找到相应的记录。可以推断这两个表达都不太合适，四号学生需要寻找更为准确地道的表达。

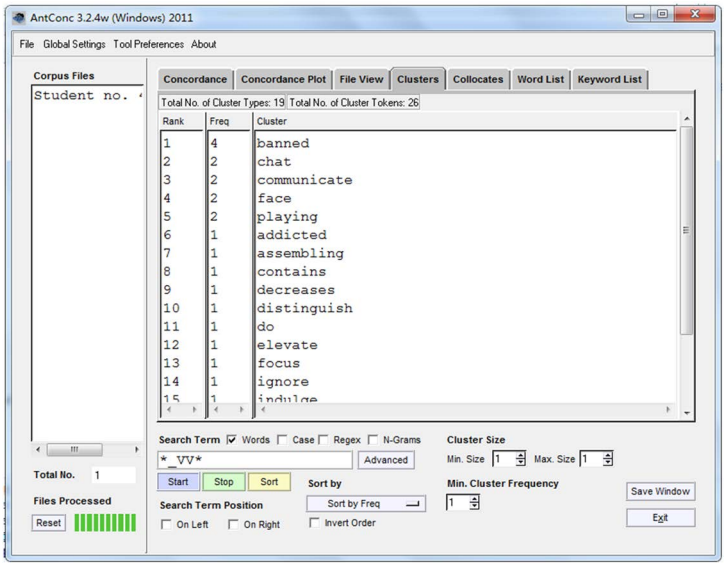

Figure 10. The verbal wordlist of student no.4 (excerpt)
图 10. 四号学生的动词列表(节选)

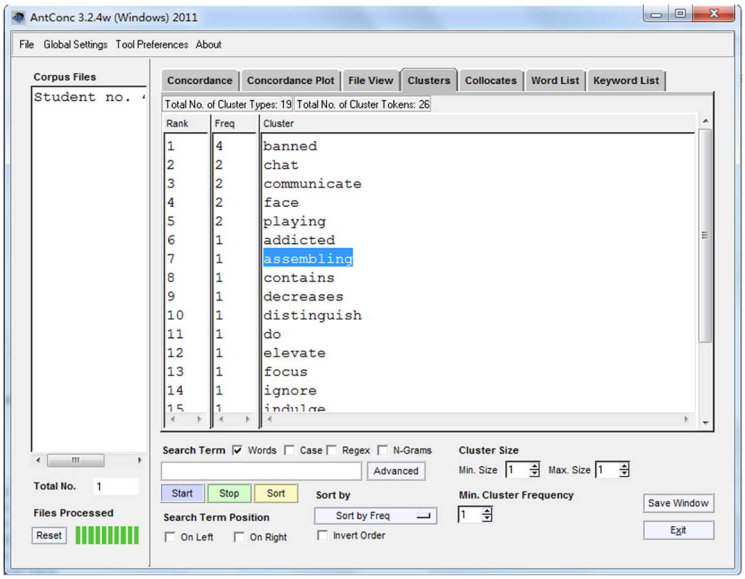

Figure 11. Distinctive diction e.g. “assemble”
图 11. 特色措辞“assemble”

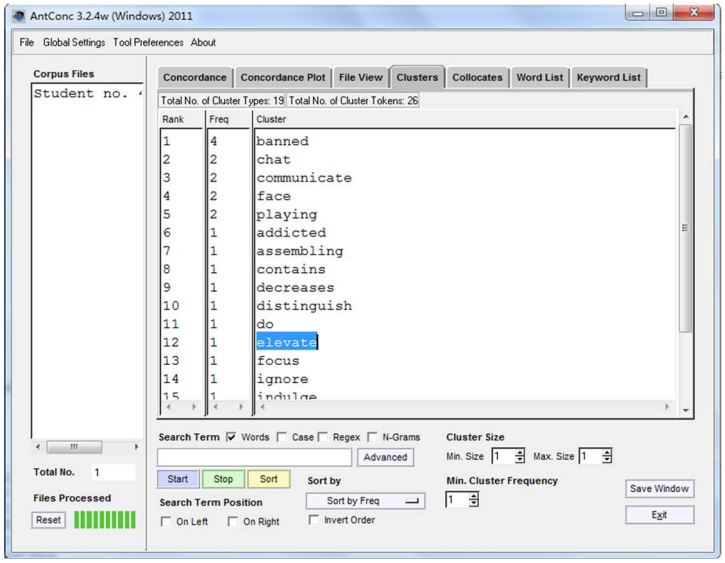

Figure 12. Distinctive diction e.g. “elevate”
图 12. 特色措辞“elevate”

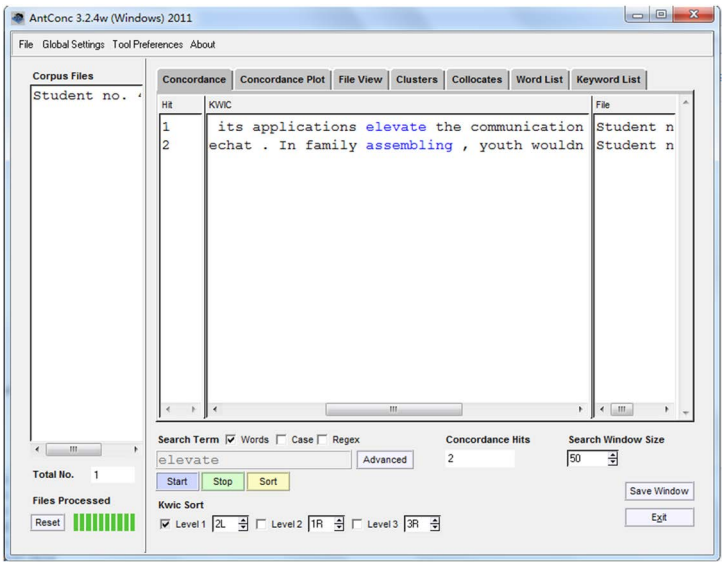

Figure 13. The concordance interface of “elevate” and “assemble”
图 13. “elevate”和“assemble”在共现界面的结果

2.2.2. 指导学生在语料库中寻找搭配范例

针对错误或低频搭配，教师可以引导学生在范例语料库中找到合适的搭配，下面以“elevate the communication”为例。学生可以在 COCA 语料库中自行使用代码“elevat* the *.[n*]”和“*.[v*] the communication”查找 elevate 和 communication 高频地道的搭配(图 14，表 3)。

根据以下搜索结果，“elevate”通常搭配抽象性名词如“importance”和“status”，排除了和“communication”搭配的可能性。另一方面，和“communication”搭配的动词多为“improve”和“facilitate”等属于同一语义场的动词。利用同样的方法，我们可得出和“family”的地道搭配为“family gathering”而非“family assembling”。最后，我们引导学生得出结论“improve/facilitate the communication”和“family gathering”才是恰当合适的搭配。

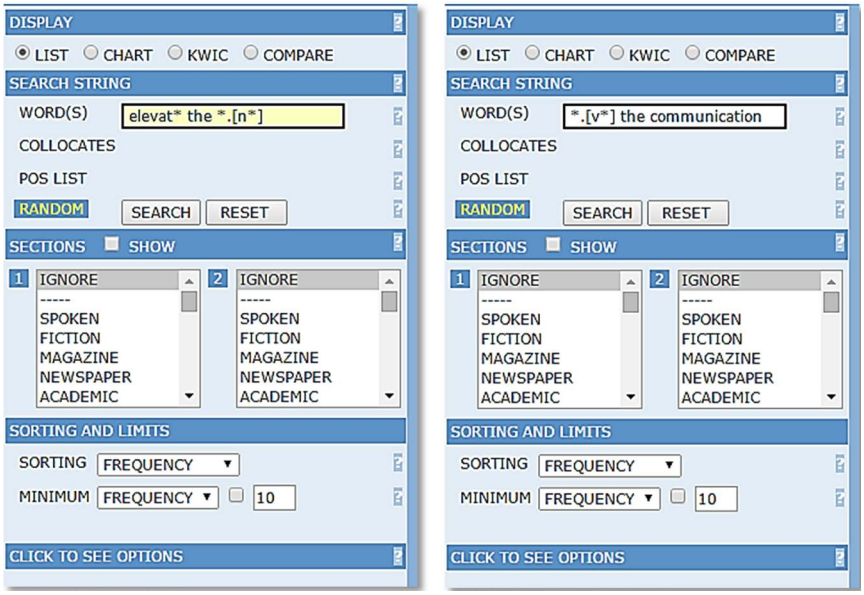

Figure 14. The respective codes in the search bar of COCA
图 14. 在 COCA 中的两个搜索代码

Table 3. Excerpts of search results of collocations based upon the codes above
表 3. 根据以上代码所得的搜索结果(节选)

Rank	Result	Freq	Rank	Result	Freq
1	elevated the status	11	1	improving the communication	6
2	elevate the status	9	2	improve the communication	6
3	elevate the importance	8	3	bridging the communication	5
4	elevate the level	8	4	bridge the communication	5
5	elevated the importance	8	5	increase the communication	3
6	elevating the head	8	6	open the communication	3
7	elevating the stock	8	7	disrupt the communication	2
8	elevate the quality	7	8	explain the communication	2
9	elevate the role	6	9	facilitate the communication	2
10	elevating the status	6	10	meet the communication	2
11	elevate the debate	5	11	overcome the communication	2
12	elevate the spirit	5	12	practicing the communication	2
13	elevate the discussion	4	13	process the communication	2
14	elevate the body	4	14	protect the communication	2
15	elevated the concept	4	15	receive the communication	2
16	elevate the head	3	16	think the communication	2
17	elevate the rights	3	17	thought the communication	2
18	elevate the tone	3	18	lost the communication	1
19	elevate the women	3	19	link the communication	1
20	elevate the wound	3	20	let the communication	1
	...			...	

2.2.3. 布置任务让学生仿写语料库的句子

除词汇的灵活运用外，语法准确度和多样性也是获取高分的另一重要方面。为了使学生免遭陈腐句型模板的折磨，更好地提升句子语法多样性，范例语料库可以提供很多实际的帮助。从范例语料库中我们可提取自然地道的句子结构，满足广大学生的语法句式训练需求。由于英语语法整体框架过于庞大，我们优先考虑功能性语句的训练如表达观点、举例、强调、以及名词化的句子。以下拿名词化举例。

名词化的定义

“语言学上，名词化是指将动词、形容词、副词作为名词或名词化短语使用，且带有一定程度上的词形转变。”[7]虽然两句所表达的意思相近，但经过名词化处理的句子比原句更为正式和简洁。在学术类英语写作中，正式程度的高低也是一大关键点。

e.g. 1:

If we more closely examine the admittedly small sample of data Dr. Rockwell collected, we would explainwhat is happening in two aspects.

A closer examination of the admittedly small sample of data Dr. Rockwell collected suggested two explanations for what is happening.”

——经济学人 2013-06-01 [8]

e.g. 2

Whether the participants are involved or not has been essential to the development of relevant programs.

The involvement of the participants has been essential to the development of relevant programs.

—*Reading Passage 1, Test 3,* 剑桥雅思真题系列 4 [9]

对比而言，我们可以清楚地意识到上述例子中，名词化的句子(句二)都比原句更正式。句子的正式程度和句子的名词化程度成正比。根据一贯的学生为本原则，我们可以从学生偏好的词汇下手，作为名词化的种子。

下面以二号学生使用的“consider”为例：

“If we carefully consider the noise and disturbance of mobile phone, we can see the distraction it exerts in resting places.”

接下来我们把“consider”作为名词化的种子，在范例语料库中搜寻相关例句如下：

“Careful consideration of our system of numeration leads to the conviction that…”

—*Reading Passage 3, Test 2,* 剑桥雅思真题系列 6 [10]

二号学生将被给予指引，观察两句话的结构，并融合两句话的特点，仿写得出名词化句子如下：

“Our careful consideration of the noise and disturbance of mobile phone illustrates the distraction it exerts in resting places.”

总的来说，有了范例语料库和相关指引，理论上学生能够通过这个方法对自己的词汇入手，修改自己所写的句子，提高其句子的名词化程度，成为具有个人特色的写作者。

2.3. 基于语料库对写作的量化评估

2.3.1. 基于语料库的搭配强度评估流程

全面综合的写作评估是十分复杂和多层次的任务，因此在本章节我们主要讨论搭配强度(collocational strength)这一关键因素。在本章节中 AntConc 和其他大型语料库的数据将在评估过程中扮演关键的角色。

通常情况下，COCA 语料库能满足大部分研究需求。但在该项研究中我们需要多个语料库协同，尤其是更大型的语料库如谷歌图书语料库(Google Books Corpus)、谷歌搜索引擎等。主要原因是验证搭配的

地道性需要多方验证，不能只限制在一个语料库中，否则会造成数据的误差和缺失等情况。前文已涉及到谷歌搜索引擎的功能，此处不再赘述。而谷歌图书语料库则是具有 1150 亿词的语料库，收录了海量正式出版物的电子文本。通过三个语料库的协同作用，我们可以建立起较为完备的短语搭配评估分级体系。在该研究中，各大语料库的词频和数据是决定搭配强度的核心。具体流程图展示如下图 15。换言之，短语根据词频和数据可以进行归类，有明确的搭配强度分级额具体得分，如表 4 所示。

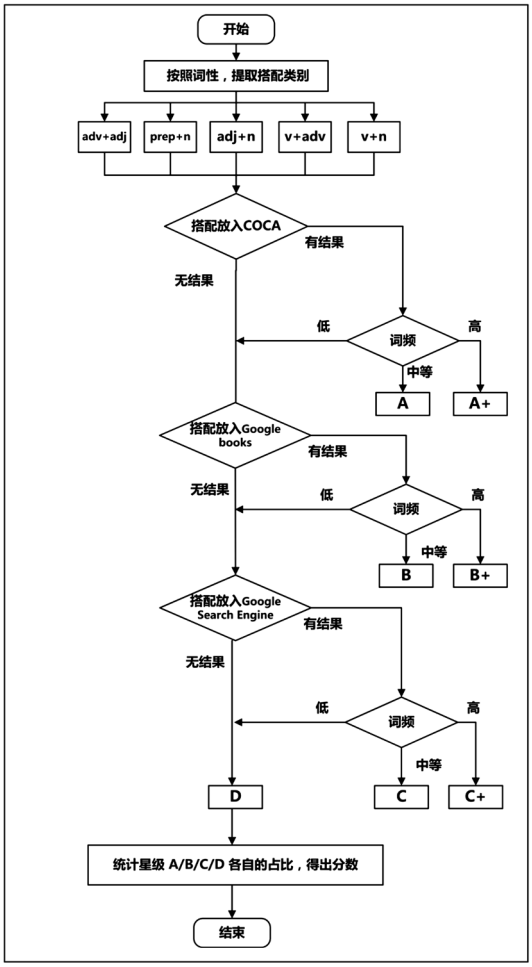

Figure 15. The flowchart of corpus-based rating of collocational strength
图 15. .基于语料库的搭配强度流程图

Table 4. Hierarchical structure of phrases and collocations
表 4. 短语和搭配的强度分级表

分类	子类	地道性	使用频率	得分
A	A+	√	★★★★★	5
	A	√	★★★★☆	4.5
B	B+	√	★★★★	4
	B	√	★★★☆	3.5
C	C+	√	★★☆	2.5
	C	√	★★	2
D	D	×	★	1

首先我们可以根据词性对学生的短语搭配进行分类，如动宾短语，形容词性短语等。若想获得学生的动宾搭配列表，我们可以在 AntConc 的词簇界面输入以下代码“*_VV*”并设置最小词簇数为 6。以下为四号学生的动宾短语列表(图 16)。

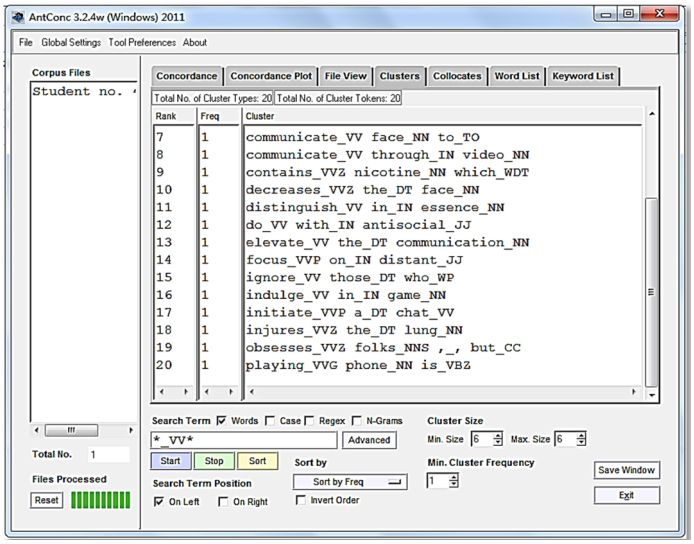

Figure 16. The search result of verbal collocations of student no. 4
图 16. 四号学生的动宾搭配搜索结果

四号学生使用的两个搭配为“elevate the communication”和“initiateachat”。在章节 2.2.2 我们已讨论过“elevate the communication”并非地道表达，在三大语料库中都不存在词频数据，属于 D 类搭配。而“initiateachat”在 COCA 中频率为 1，在谷歌图书语料库中为 96，在谷歌引擎中为 7499。换言之，“initiateachat”是地道的 C 类表达——也就是一种较为低频的用法，却具有一定的个人特色。基本上所有类别的搭配都可通过上述流程进行分类评估，得到一个综合搭配强度的总分。这样的评级和得分可以帮助我们清晰地观察到学生滥用或误用的表达，教师就可以分析其背后的根本原因，提供个性化的表达优化方案。

2.3.2. 学生个案分析

在该部分我们以五号学生为例，以上述的分类评估系统作为衡量标准。五号学生是早期首批接触语料库的学生之一，具有较好语料库技能和分析能力。我们先来关注她在不同时期的动词列表情况。下面左边的动词列表 5-1-1 是提取自未经修改的文章，右边则为修改后的版本(表 5)。

Table 5. Two complete verbal wordlists of student no. 5 at an earlier stage
表 5. 五号学生早期的两份完整动词列表

5-1-1			5-1-2		
rank	freq	words	rank	freq	words
1	5	apply	1	3	apply
2	4	do	2	2	hunt
3	4	become	3	2	interviewing
4	2	choose	4	1	agree
5	2	find	5	1	approaching
6	2	interviewing	6	1	attach
7	2	know	7	1	attend

Continued

8	1	agree	8	1	become
9	1	appreciates	9	1	claim
10	1	approaching	10	1	converse
11	1	attend	11	1	coupled
12	1	claim	12	1	craved
13	1	comes	13	1	create
14	1	confirm	14	1	do
15	1	dare	15	1	elect
16	1	following	16	1	emphasize
17	1	getting	17	1	empowered
18	1	given	18	1	erupted
19	1	hunting	19	1	exert
20	1	imagine	20	1	failed
21	1	lose	21	1	find
22	1	make	22	1	following
23	1	offered	23	1	frown
24	1	opt	24	1	get
25	1	overcoming	25	1	illustrated
26	1	prove	26	1	imagine
27	1	provided	27	1	inquired
28	1	put	28	1	lacked
29	1	recruit	29	1	longed
30	1	regard	30	1	lost
31	1	seems	31	1	managed
32	1	seen	32	1	opt
33	1	select	33	1	overcoming
34	1	strengthened	34	1	presented
35	1	succeeded	35	1	prove
36	1	take	36	1	provided
37	1	tested	37	1	put
38	1	treat	38	1	recruit
39	1	used	39	1	seems
40	1	wanted	40	1	select
			41	1	showcased
			42	1	sift
			43	1	strengthened
			44	1	supplied
			45	1	take
			46	1	utilize

根据观察，五号学生减少了她个人的高频词汇“apply”和“do”的使用，将最高动词词频从 5 降至 3，与此同时添加了原有动词的同义表达，如“utilize”相对“apply”，“manage”相对“succeed”，“crave”相对“want”等。此外，有不少同义词和原有动词是相应地共同使用的，比如“supply”和“provide”，“long”和“crave”等。总而言之，修改后的版本可体现出作者在动词词汇使用的丰富度和多样性。

在她后期的作品中。其最高动词的词频稳定在 2 到 3 之间，且使用同义词数目不断上升，总体呈现出良好趋势，如下表 6 所示。

Table 6. The recent verbal wordlists of student no. 5
表 6. 五号学生近期完整动词列表

5-2-1 (original)			5-2-2 (revised)		
rank	freq	words	rank	freq	words
1	3	do	1	3	give
2	3	give	2	2	do
3	2	consider	3	1	acknowledged
4	1	acknowledged	4	1	agree
5	1	agree	5	1	aid
6	1	aid	6	1	arise
7	1	arise	7	1	assisting
8	1	assisting	8	1	building
9	1	building	9	1	claim
10	1	claim	10	1	consider
11	1	coupled	11	1	coupled
12	1	dared	12	1	dared
13	1	develop	13	1	deem
14	1	establish	14	1	develop
15	1	existing	15	1	establish
16	1	fall	16	1	existing
17	1	feel	17	1	fall
18	1	follows	18	1	feel
19	1	forget	19	1	follows
20	1	frown	20	1	forget
21	1	fueled	21	1	frown
22	1	go	22	1	fueled
23	1	gush	23	1	go
24	1	helping	24	1	gush
25	1	illustrated	25	1	helping
26	1	imagine	26	1	hinge
27	1	isolated	27	1	illustrated
28	1	let	28	1	imagine
29	1	lift	29	1	isolated
30	1	listed	30	1	let
31	1	lose	31	1	lift
32	1	needs	32	1	listed
33	1	owing	33	1	lose
34	1	prefer	34	1	necessitates
35	1	provides	35	1	owing
36	1	pulled	36	1	planted
37	1	put	37	1	prefer
38	1	refrained	38	1	provides
39	1	saw	39	1	pulled

Continued

5-3-1 (original)			5-3-2 (revised)		
rank	freq	words	rank	freq	words
1	2	face	1	2	serves
2	2	plays	2	2	plays
3	2	serves	3	1	think
4	2	give	4	1	tend
5	2	increase	5	1	teaching
6	1	acknowledged	6	1	strengthening
7	1	arise	7	1	solving
8	1	belonging	8	1	share
9	1	came	9	1	set
10	1	climb	10	1	render
11	1	cultivate	11	1	put
12	1	discussing	12	1	promoted
13	1	feeling	13	1	nurturing
14	1	follows	14	1	lose
15	1	forming	15	1	listed
16	1	foster	16	1	let
17	1	frown	17	1	leading
18	1	fueled	18	1	insist
19	1	gathering	19	1	increase
20	1	go	20	1	imagine
21	1	gush	21	1	illustrated
22	1	ignite	22	1	ignite
23	1	illustrated	23	1	gush
24	1	imagine	24	1	go
25	1	insist	25	1	give
26	1	leading	26	1	gathering
27	1	let	27	1	fueled
28	1	listed	28	1	frown
29	1	lose	29	1	foster
30	1	nurturing	30	1	forming
31	1	promoted	31	1	follows
32	1	put	32	1	feeling
33	1	set	33	1	face
34	1	share	34	1	enhance
35	1	solving	35	1	discussing
36	1	strengthening	36	1	cultivate
37	1	teaching	37	1	confront
38	1	tend	38	1	climb
39	1	think	39	1	came

在短语搭配的分类评级方面，我们主要针对其在作文 5-1-1 和作文 5-1-2 中使用的名词性短语进行了分析评估。具体表格如下表 7~8。

Table 7. The noun phrases in essay 5-1-1 and their grading based on frequency in three corpora
表 7. 作文 5-1-1 中的名词性短语及其基于语料库频率的分级

5-1-1				
grade	words and phrases	freq		
		COCA	Google Books Corpus	Google
B+	academic qualifications	41	7914	/
A+	better way	1949	/	/
B+	communicative skills	17	7783	/
A	educational background	222	/	/
B+	elite schools	79	5503	/
C+	excellent employees	2	585	132,000
B+	face interview	0	6093	/
B	fair method	4	2718	/
A	first reason	177	32,193	/
A+	good choice	781	/	/
B+	heated discussion	91	23,141	/
A	high ability	236	/	/
B	latter viewpoint	5	2040	/
C+	low qualifications	2	421	39,600
A+	positive attitude	748	/	/
C	promising employees	1	431	15,000
A	proper way	385	/	/
B+	psychological qualities	6	4114	/
B+	qualified employees	17	6077	/
B+	several interviews	62	16,634	/
A	sharp increase	172	50,354	/
B+	shy girl	36	4537	/

Table 8. The noun phrases in revised essay and their grading based on frequency in three corpora
表 8. 修改后的作文 5-1-2 中的名词性短语及其基于语料库频率的分级

5-1-2				
grade	words and phrases	freq		
		COCA	Google Books Corpus	Google
B+	academic qualifications	41	7914	/
B	better vehicle	20	2227	/
A+	certain degree	560	/	/
A	educational background	222	/	/
C+	educational degrees	7	1677	145,000
C+	elite students	10	836	112,000
B+	established fact	39	60,571	/
B+	excellent ability	2	2758	/
B+	excellent candidates	23	3928	/
C+	excellent employees	2	585	132,000
B+	fair standard	2	4928	/
A	great importance	424	/	/

Continued

B+	heated discussion	91	23,141	/
A	interpersonal skills	345	/	/
B	latter viewpoint	5	2040	/
A	positive attitude	748	/	/
C	promising employees	1	431	15,000
B+	psychological qualities	6	4114	/
B+	qualified employees	17	6077	/
B+	several interviews	62	16,634	/
A	sharp increase	172	50,354	/
B+	shy girl	36	4537	/

将上述数据转换为饼状图后，我们可得到各类短语的占比情况如下图 17~18。

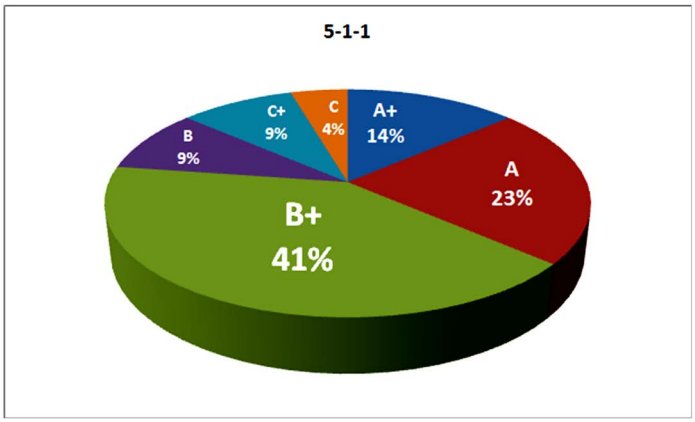

Figure 17. The proportion of nominal collocations of different types in essay 5-1-1
图 17. 作文 5-1-1 中各类名词性短语的占比情况

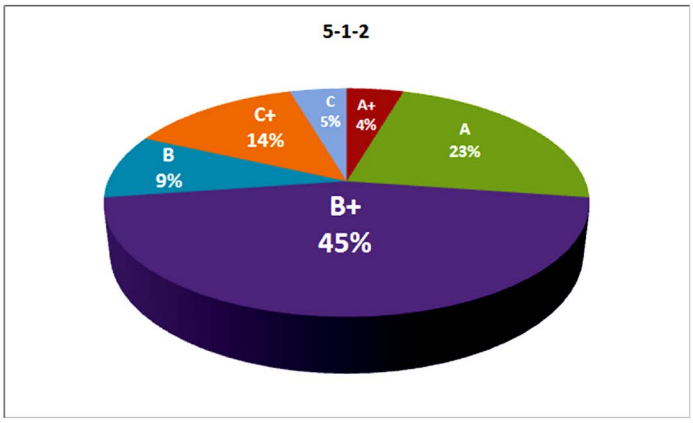

Figure 18. The proportion of nominal collocations of different types in revised essay
图 18. 作文 5-1-2 中各类名词性短语的占比情况

在两篇作文中，占比最大的皆为 B+类短语，超过了 40%；紧接着的是 A 类短语，在两篇作文中稳定地占比 23%；同样占比稳定的有 B 类短语(9%)。而相比而言，在第二篇作文中 C+类短语上升了 5%，而 C 类短语上升了 1%。反而 A+类短语从 14%降至 4%。

通过这个案例，我们可观察到的结论是：在修改过程当中，五号学生在名词性短语这一块不仅保持了较高的地道性，而且还努力尝试加入虽然较为低频但具有特色的表达方式。总体来说是呈上升趋势。

3. 语料库应用的优势和局限

总而言之，对于教师和研究者而言，使用语料库工具的优势主要在于：

1) 量化评估。使用者可以通过数据来对学生的词汇丰富度和搭配强度进行精准有效的评估，评估之后这些数据以可视化的方式展示给学生，从而让学生清楚地了解自己写作用词和搭配上的优点和缺点，同时教师可以根据这些结果对学生进行写作反馈，引导学生扬长避短，提高写作整体用词和搭配的质量；

2) 数据留存。通过积累和留存学生写作的数据，教师和研究者可以预测之后的学生学习轨迹，从而制定更有针对性、个性化的学习路径，使用更符合学生水平的材料作为写作范本，提高教学效率。

虽然语料库给写作研究带了不少便利，但其缺点和限制也同样明显。

1) 操作复杂。作为一门新兴学科，基于语料库的写作研究无疑是跨学科的，不仅需要语言学的相关理论，还需要相关的电脑理论和操作等。如此复杂的理论基础和操作阻碍了语料库在师生之间的普及。建立小型语料库至少需要数月的投入和努力，且需要良好的计算机操作能力，更不用说更大型语料库的建设。进入语料库获取相关信息也需要较高的数字和编码能力；

2) 规模有限。目前我们可针对搭配和词汇分析，但还缺乏更方便有效的方法理论来探索更深层次的内部文本属性，目前尚且无法用 AntConc 对句子或者段落为单位进行分析和研究。这些问题有望在将来随着科技的发展得到进一步探索和解决。

参考文献

[1] 梁茂成, 李文中, 许家金. 语料库应用教程[M]. 北京: 外语教学与研究出版社, 2010.
[2] http://en.wikipedia.org/wiki/Corpus
[3] http://www.antlab.sci.waseda.ac.jp/software.html#antconc
[4] https://courses.washington.edu/hypertxt/csar-v02/penntable.html
[5] https://takeielts.britishcouncil.org/sites/default/files/ielts_task_2_writing_band_descriptors.pdf
[6] https://www.english-corpora.org/coca/
[7] http://en.wikipedia.org/wiki/Nominalization
[8] https://www.economist.com/science-and-technology/2013/06/01/sacred-geese
[9] 剑桥大学考试委员会. 剑桥雅思官方真题集 4: 学术类[M]. 剑桥: 剑桥大学出版社, 2005.
[10] 剑桥大学考试委员会. 剑桥雅思官方真题集 6: 学术类[M]. 剑桥: 剑桥大学出版社, 2007.