Modern Linguistics 现代语言学, 2024, 12(10), 51-62
Published Online October 2024 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2024.1210859

Hans汉斯

基于语料库的大学英语四、六级考试词表测评研究

黄宇歆，常晓莹，罗卫华

大连海事大学外国语学院，辽宁 大连

收稿日期：2024年8月21日；录用日期：2024年9月30日；发布日期：2024年10月11日

摘　要

本研究利用COCA5000词表、CETC与iWriteBaby语料库作为参照对象考查大学英语四、六级考试词表(CET词表)。CET与三个参照对象的对比结果表明，CET词表包含大部分三个对比组的共有高频词，但其中还有CET_ONLY词汇，即在三个参照对象中都从未出现的词汇以及共有词中还有部分低频词。基于此并结合词表特点，本研究针对CET词表的编制优化提出了相应的参考建议，以期对我国大学生英语学习者的词汇学习提供更为便利的条件。

关键词

大学英语四、六级词表，基于语料库，词汇覆盖率

A Corpus-Based Research on the Assessment of CET Vocabulary List

Yuxin Huang, Xiaoying Chang, Weihua Luo

School of Foreign Languages, Dalian Maritime University, Dalian Liaoning

Received: Aug. 21st, 2024; accepted: Sep. 30th, 2024; published: Oct. 11th, 2024

Abstract

This study uses the COCA5000 vocabulary, CETC and iWriteBaby corpus as reference objects to examine the CET vocabulary list. The results of the comparison between CET and the three reference objects show that the CET vocabulary contains most of the common high-frequency words of the three comparison groups, but there are also CET_ONLY words, that is, words that have never appeared in the three reference objects and some low-frequency words in the common words. Based on this and characteristics of the vocabulary list, this study puts forward corresponding reference

文章引用: 黄宇歆, 常晓莹, 罗卫华. 基于语料库的大学英语四、六级考试词表测评研究[J]. 现代语言学, 2024, 12(10): 51-62. DOI: 10.12677/ml.2024.1210859

suggestions for the compilation and optimization of the CET vocabulary list, in order to provide more convenient conditions for vocabulary learning of English learners in Chinese colleges.

Keywords

CET Vocabulary List, Corpus-Based, Lexical Coverage

1. 引言

全国大学英语四、六级考试(College English Test)，简称“CET”，是“由教育部主办、教育部教育考试院主持和实施的大规模标准化考试，其目的是促进我国大学英语教学工作，对大学生的英语能力进行客观、准确的测量，为提高我国大学英语课程的教学质量提供服务。”在我国，大学英语四、六级考试被用于测试中国在校大学生的综合英语能力，且能否通过该测试这一标准也被广泛用于各高校中，作为能否取得毕业证的门槛，可以说，不管在毕业升学还是求职等各方面，能否通过大学英语四、六级考试以及能否在其中取得好成绩，对我国在校大学生来说至关重要。

CET 全方位考察大学生的英语综合能力，而在英语学习中，词汇是公认的英语听、说、读、写能力发展的基础，因此在备考大学英语四、六级考试时，词汇学习的重要性不言而喻。为了便于考生备考，2016 年版的大学英语四、六级考试大纲新增了四、六级考试词表。一般来说，词表由一定数量的词汇构成，其中包含的词汇通常按照特定原则选定，因此不同研究人员开发的词表都各不相同。英语词表能帮助二语学习者提高词汇学习效率和自主学习能力，对二语学习者词汇学习具有重要作用，此外成熟科学的词表还能作为英语教学和测评的依据，成为教学大纲甚至教学材料的参照。

根据大学英语四、六级考试大纲所言，大学英语四、六级词表的编制参考了多方词表、词典及词库中的词汇，并且“词目的选择遵循‘以定量分析为主，定性分析为辅’的原则”。通过查阅文献可知，国内对于大学英语四、六级词表的研究屈指可数，受国外词表相关研究的启发，本研究从本族语语料库、试题语料库和中国大学生英语学习者语料库三类语料入手考查大学英语四、六级词表，研究问题包括：1) 大学英语四六级考试词表与各个参照对象的对比结果如何？2) 上述对比结果对 CET 词表编制有何启示？

2. 研究设计

2.1. 研究对象

大学英语四、六级词表来源于全国大学英语四、六级考试大纲(2016 版)，也是近些年最新版本，包含 5418 个词目。由于全国大学生英语考试分为四级和六级两个等级，与考试等级相对应，词表也对其中词汇进行了四级词汇和六级词汇的分类标注。如上所述，该词表在一定程度上代表期望中国大学生英语词汇学习应达到的水平。

COCA5000 词表来自于由学者 Mark Davies 创建的美国当代英语语料库(Corpus of Contemporary American English，简称 COCA)，包含 COCA 中总频次排序前五千的词目及相关一系列数据。COCA 由总计约 10 亿词、共八个类型且分布基本均匀的文本构成，总体上文本类型广泛，词量庞大，且该库语料逐年更新，在本族语词汇方面具有代表性。

大学英语四六级考试试卷语料库(College English Test Corpus)，简称 CETC，是本研究自建语料库，其中包括 2016 年至 2024 年期间大学英语四、六级考试的卷面文本，总计超三十八万词。需要特别说明的是，由于听力原文材料以语音输出，而非词形，所以在本研究中暂时不做处理。

iWriteBaby Chinese Learner English Corpus，即 iWriteBaby 中国学习者英语语料库(以下简称 iWriteBaby 语料库)，由北京外国语大学许家金教授团队加工整理，最初发布于 2019 年，而后在 2022 年进行了更新，总规模超八百万词，攘括全国二十三个省份、四十八座城市、六十九所不同水平大学的一百五十四个专业学生的写作文本，入库的作文题超一千个。该语料库词汇容量大，在地域、年级及主题等各类因素上均取样广泛，在中国大学生的用词总体特征方面有一定代表性[1]。

2.2. 数据处理

大学英语四、六级词表直接取自全国大学英语四、六级考试大纲(2016 版)。由于本研究在词目化(lemmatization)的实际操作中只考虑词形，最后产出的词目表中的词目实际上属于“modified lemma (Stoeckel, 2019)”[2]，因此对于该词表中分列的一词多义词，本研究将其合并只保留其一个词形，如“bank1”表“银行”和“bank2”表“堤坝”会全都归并为“bank”，最后处理完的词表包含 5403 个词目。

COCA5000 词表分开放置词形相同却拥有不同词性的词汇，同上，在本研究中利用 Microsoft Excel 中的“数据–合并计算”功能将这类词全部归并处理，最后该词表中包含 4380 个词目。

大学英语四、六级考试试卷语料库(College English Test Corpus)是本研究自建语料库，在利用 Microsoft Word 中“插入–对象–文件中的文字”功能将 2016 年至 2024 年期间大学英语四、六级考试的各个试卷文档合并后，去除中文字符及标点，而后使用 Perl 软件对余下的文本进行清洁，生成 CETC 的类符词汇表。将该类符词汇表导入 Treetagger 软件进行赋码处理，而后生成 CETC 词目表。

iWriteBaby 语料库词表通过 BFSU CQPweb 平台导出，经 Microsoft Visual FoxPro 进行符号清理后，借助 Treetaggger 软件进行词性标注，而后形成词目表。

为进一步确保以上操作生成数据的准确性，本研究对生成的词目表进行人工查验。此外为实现数据计算的统一化，借助 Microsoft Excel 人工计算各词表每百万词标准化词频并全数据应用，得出最终归拢的所有词目的每百万词标准化词频。

最后，利用 Microsoft Excel 进行词表对比。通过 Excel 中“开始–条件格式–突出显示单元格规则–重复值”这一功能，将 CET 词表分别与 COCA5000 词表、CETC 词表和 iWriteBaby 语料库产出的词表两两进行比较，得出各两词表之间重复的词汇，即共有词。在此基础上，利用“开始–筛选”功能筛出两表共有词，而后便可生成两表各自独有词。通过以上一系列操作，最终得出的数据表包括 CET-COCA，CET-CETC 和 CET-iWriteBaby。

3. 分析与发现

3.1. CET-COCA 对比结果分析

本研究选用的 COCA5000 词表来源于 COCA 语料库网站，其中包含美国当代英语语料库中排名前 5000 的词目(lemma)，经统一化处理后词目总计 4380 个。CET 词表和 COCA 词数对比结果如表 1 所示，其中两词表的共有词分别占两词表总词目数的 58.19%和 71.78%，独有词则分别占两词表总词数的 41.81%和 28.22%。

通过查看 COCA5000 词表可以知道，COCA5000 词表中所有词目的累计词频占 COCA 总库词量的约 83%，而 CET 词表和 COCA5000 两词表的 3144 个共有词目的累计词频就已经占了 COCA 总库的约 77%。结合表 2 分析，CET 词表与 COCA5000 词表相重合的词汇在 COCA 各个频段中的数量依次递减。CET 词表中有 903 个词目位于 COCA 的第一个频段，即前一千词中，仅这一部分词的总体频次就占了 COCA

总词量的近 70%。

Table 1. Comparison results of lemmas in CET and COCA5000
表 1. CET 与 COCA5000 词目对比结果

	词目总数	共有词数	独有词数
CET	5403	3144	2259
COCA	4380		1236

Table 2. The number of CET lemmas in each frequency band in COCA5000
表 2. CET 词目在 COCA5000 各频段数目

COCA 频段	COCA 词目排名	CET 词目在各频段总数
1	1~1000	903
2	1001~2000	744
3	2001~3000	650
4	3001~4000	630
5	4001~4380	217

由此来看，参照 COCA5000 而言，虽然 CET 词表中这些词目总体上在 COCA 总库中体量少，总频次占比大，但是结合 CET 词表的总体词量考虑，其中只有不到 60%的词目在英语本族文本中的使用频率居于前列，剩下的 2259 个词目并未出现在 COCA 前 5000 个词目中。经初步推断，这部分词的选用原因之一可能是为了满足中国大学生英语学习者更高水平词汇学习的需求，因为 2259 个 CET 独有词中有 1084 个都属于大学英语六级词汇，而这部分词汇相较于上述位于前列的高频词汇，恰恰不管在词形还是词义方面都更为复杂，属于中国英语学习者词汇学习中的“高级词汇”。由于 COCA5000 词表的词量限制，更为实际具体的原因需要利用其他参照对象进一步考察。

3.2. CET-CETC 对比结果分析

本部分内容就 CET 词表和 CETC (College English Test Corpus)的对比结果进行分析。CETC 中的语料是大学英语四、六级考试的所有卷面内容，经处理有 10,872 个词目，词数总计 382,778 个。经与 CET 词表对比，两者共有词数分别占两表词目总数的 81.95%和 40.74%，两表独有词则分别占各表的 18.05%和 59.26%，两表词目总数对比具体数据如表 3 所示。

Table 3. Comparison results of lemmas in CET and CETC
表 3. CET 与 CETC 词目对比结果

	词目总数	共有词数	独有词数
CET	5403	4429	975
CETC	10,872		6443

CET 词表与 CETC 的共有词仅占 CETC 词目总数的不到 50%，导致这一情况的原因有一部分是 CETC 本身的词量较之 CET 词表就要大一倍。除此之外，从上表能观察到的结论似乎很有限。但当在词目数基础上加上各词目的累计频次会发现，尽管 CET 词表对 CETC 的词目覆盖率很低，但总体词目频次覆盖率却能达到 85.15% (见表 4)，这在一定程度上表示两词表大部分共有词在 CETC 中的出现频次都相当高。

Table 4. CET's lemmas coverage of CETC
表 4. CET 对 CETC 的词目覆盖状况

				CET	覆盖率(%)
CETC	词目总数	10,872	已覆盖词目数	4429	40.74
	词目累计总频次	382,778	已覆盖词目总频次	325,952	85.15

当把 CET 词表和 CETC 的共有词按频次降序排序后进行观察，显而易见地是两表共有词中排序前一百的词绝大部分都是功能词，如冠词 a、an 和 the；代词 you、we 和 they 等；介词 to、of、in、on 和 from 等；连词 and、but 和 so 等(见图 1)。这与先前的研究结果相呼应，在各类英语语料库位于前列的高频词中，往往都有部分功能词占据重要分量。这类词虽然在英语中数量有限，但却在英语词句的衔接、指代及限定等方面都发挥了重要作用。

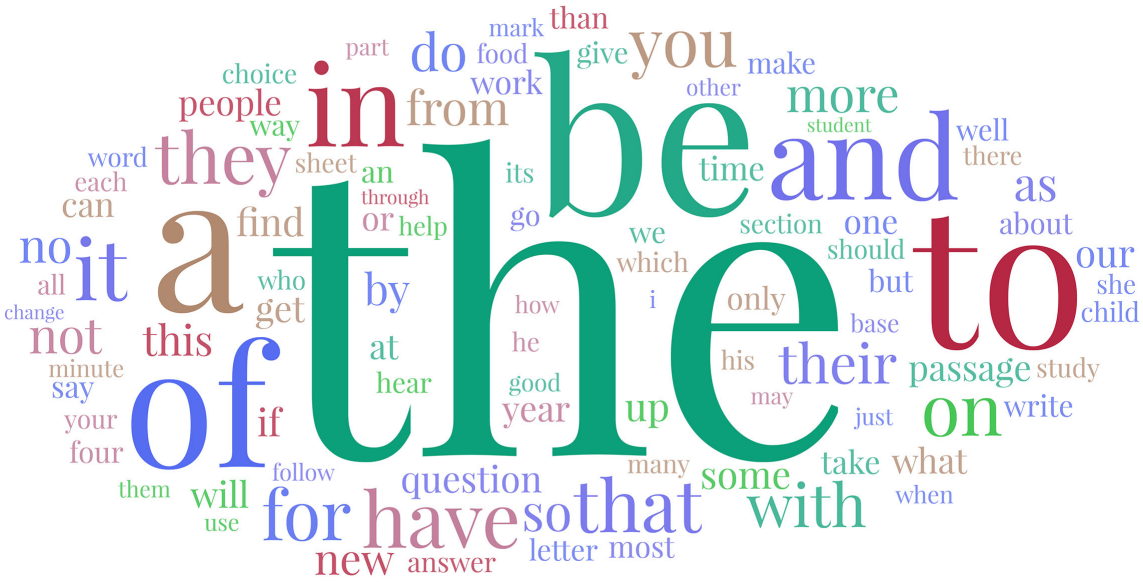

Figure 1. CET-CETC first 100 shared words
图 1. CET-CETC 前一百共有词

而除了上述功能词外，前一百词中还有部分实义词。而这些实义词中，有一部分词受题干影响反复出现，如 question、answer、sheet、section、mark、choice、letter、passage 和 base 等。经查看不受题干影响或者说受题干影响较小且使用频次依旧居高的实义词有 people、say、make、good、time、take、change 和 go 等。以上这些词汇也常出现于其他词表的高频词段中，因其普遍具有包括通用性、释义和替代能力强及搭配构词能力强等诸多词汇特征，也通常被称为核心词汇(Carter, 1987) [3] [4]。

除了上述高频词外，CET 词表和 CETC 词表共有词中还存在一部分频次更低的词汇。图 2 中是两表共有词中频次 ≤10 的词汇及各词频下的词汇总量。为便于统一对比，图 2 中所用频次是基于原频次计算而得的每百万标准频次。十分明显地是这些低频词的分量并不小，简要计算后可知，频次 ≤10 的词汇总数有 1568 个，占共有词目总数的 35.40%，而这 1568 个词的累计词频总量仅有 8775，占共有词每百万总频次的 1%，占一百万词量的约 0.09%。CETC 中涵盖了 2016~2024 年，共计九年的大学英语四级和六级试卷卷面内容，而大学英语四、六级考试一年两次，每次考试有三套试题，总计超 100 套试题，每套试题大约有 4000 左右个词，而这些低频词在这些文本中的出现频率屈指可数。值得注意地是，CET 词表与

CETC 词表的 4429 个共有词中有 756 个六级词汇，而在这 1568 个低频词中，六级词汇有 592 个，这些词汇在词表中的存在意义需要进一步考量。

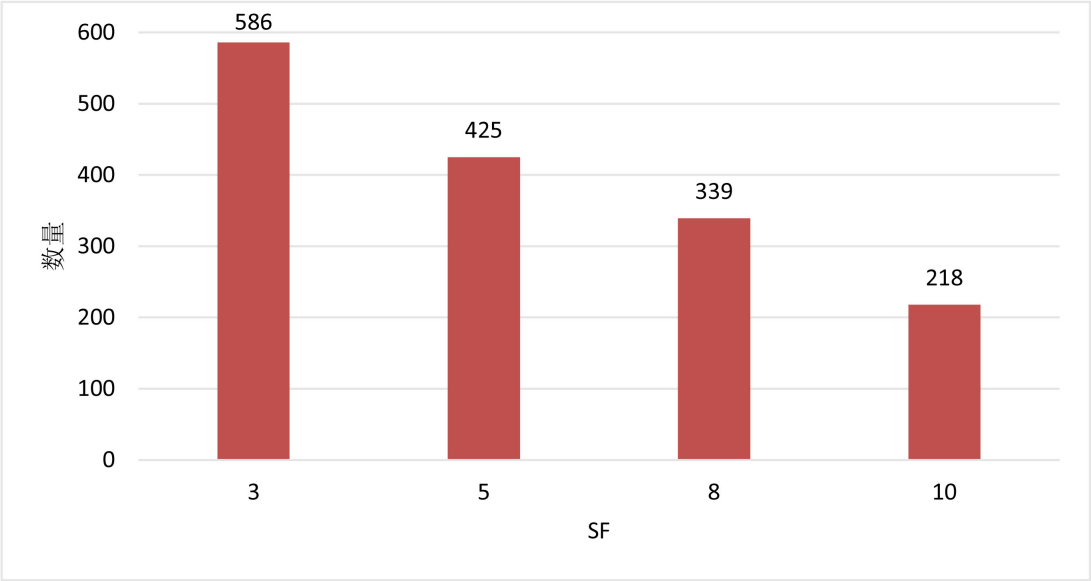

Figure 2. The number of shared words whose SF per million are less than 10 in CET-CETC
图 2. CET-CETC 共有词中每百万标准词频 ≤ 10 的词量

此外，经 CET 词表和 CETC 对比后也产出了双方的独有词。一方面，CET 词表中的独有词即表明这部分词在 CETC 中从未出现过，而这部分词被编入 CET 词表是否合理还需要和本文其他部分的对比结果相结合分析后才能做出判定，因为 CETC 部分的词汇仅仅是大学英语四、六级考试的卷面文本，这些语料词汇具体来说更倾向于输入性词汇，总体上并不能代表大学英语四六级学习所需词汇。

另一方面，CETC 独有词中更加值得注意的是其中的高频词汇。由于 CETC 中的语料是卷面文本，因此最后生成的词表中有小部分被高频用作选项符号的英文字母。当把它们排除，剩下的高频词主要由派生词和国家名称构成，一般来说国家名称并不会列入词表，更多地是作为专有词汇列入国家地理类的专有名词词表。而为避免行文累赘，关于 CETC 独有词中的高频派生词，本文将在后半部分统一讨论。

3.3. CET-iWriteBaby 对比结果分析

表 5 是 CET 词表和 iWriteBaby 语料库词表的对比结果。两表的共有词分别占两表总词目数的 95.17%和 28.67%，而两表独有词则占有各自词表总词目数的 4.83%和 71.33%。与上一部分关于 CET 词表和 CETC 的对比结果的讨论相类似，造成 CET 词表和 iWriteBaby 语料库词表对比结果差异较大的主要原因还是两表的总词量差异。

Table 5. Comparison results of lemmas in CET and iWriteBaby
表 5. CET 与 iWriteBaby 词目对比结果

	词目总数	共有词数	独有词数
CET	5403	5142	261
iWriteBaby	17,936		12,794

iWriteBaby 语料库的总词量是 CET 词表的三倍多，也是因此，CET 词表对 iWriteBaby 语料库的已覆盖词目总数比之 CET 词表对 CETC 的已覆盖词目数要更多。结合 CET 词表与 CETC 的对比分析结果来看，CET 词表对 iWriteBaby 语料库的词目覆盖率更低，仅有 28.67%，但词目总频次的覆盖率却高得多，超 90% (见表 6)。这一方面是因为 iWriteBaby 语料库词量大且语料话题更广泛，另一方面，与上一部分的内容相呼应，两表共有词大部分在 iWriteBaby 语料库中都被高频使用。

Table 6. CET's lemmas coverage of iWriteBaby corpus
表 6. CET 对 iWriteBaby 语料库的词目覆盖状况

				CET	覆盖率(%)
iWriteBaby	词目总数	17,936	已覆盖词目数	5142	28.67
	词目累计总频次	8,046,643	已覆盖词目总频次	7,440,665	92.47

在 CET 词表与 iWriteBaby 语料库共有词中，前一百个词的累计频次为 5,018,391，占 iWriteBaby 语料库词汇总数的 62.37%。通过观察图 3 可知，与 CET 词表和 CETC 前一百个共有词(图 1)相似，其中包含了大量如冠词、代词、介词、连词和助动词等在内的功能词以及部分实义词。而与 CETC 不同，iWriteBaby 语料库中的语料来源于大学生的写作文本，因此就 CET 词表与 iWriteBaby 语料库词表的前一百个共有词来说，其中的实义词相较于 CETC 除了同样具有核心词汇的特点外还有更明显的主题性，如 college、phone、book、internet 和 stress 等。

Figure 3. CET-iWriteBaby first 100 shared words
图 3. CET-iWriteBaby 前一百共有词

接下来着眼于 CET 词表和 iWriteBaby 语料库共有词中的低频词进行分析。同上，把 iWriteBaby 语料库词表中的词频进行每百万词标准化，提取其中每百万词频次 ≤10 的部分及相应的总数量，结果如图 4 所示。由于受 iWriteBaby 语料库总词量的影响，原词频 ≤ 4 的词汇在进行每百万词标准化之后的词频结果是 0。最终结果显示标准化词频 ≤10 的词汇总数有 3050 个，而这 3050 个词的累计标准化词频总量仅有 7782，占 iWriteBaby 语料库总量的不到 1%。类似地，这 3050 个低频词中有 1098 个六级词汇。

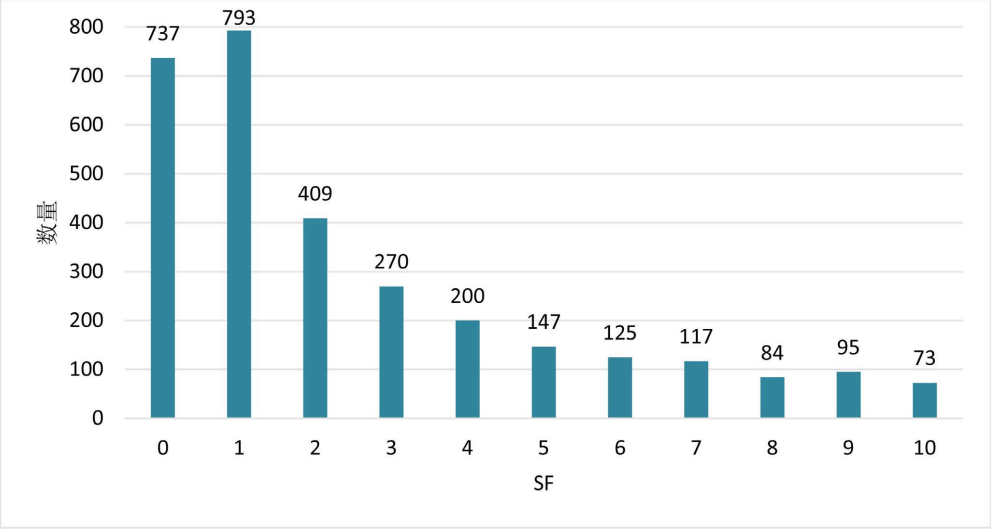

Figure 4. The number of shared words whose SF per million are less than 10 in CET-iWriteBaby
图 4. CET-iWriteBaby 共有词中每百万标准词频 ≤ 10 的词量

3.4. 总结与启示

以上三个部分是CET词表与三个参照对象——COCA5000、CETC和iWriteBaby语料库的对比结果。总体来看，CET词表中包含了各个参照对象中的绝大部分高频词，三组对比的前2000共有词目重合率约为75%。从图5可知，三个参照对象与CET词表的共有词中，前100词目的累计频次占据三参照对象的超一半词量，前2000共有词目的累计频次更甚。CET-CETC前2000词累计词频覆盖率相较于其他两参照对象更低是受卷面语料特点的影响。不同于其他语料，卷面文本在一定期间内需要具有差异性，以确保试卷时用时新，且由于CET词表是2016年考试大纲改版后才编制出的，本研究中CETC语料仅使用了2016~2024年间的大学英语四、六级试卷。受上述原因影响，最终成形的语料库中的文本话题多样，但重合度相对低，加之卷面词数有限，最终使得词频突出的词相较另外两库更少，多有中频词。

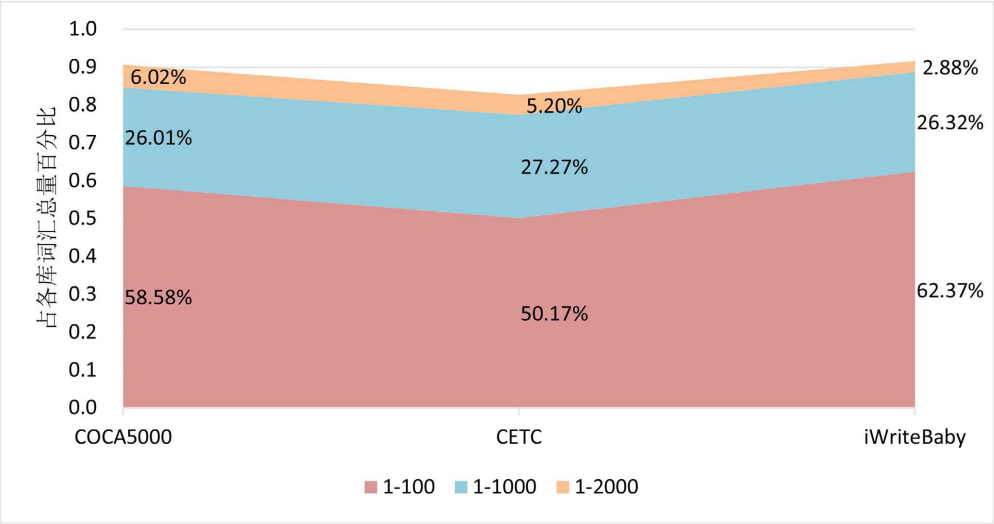

Figure 5. The cumulative frequency proportion of the first 2000 shared words of CET and the three reference objects
图 5. CET 与三参照对象前 2000 共有词目累计频次占比

此外，也有部分词汇尽管在各个参照对象中都有出现，但使用频次却极低，这些词加之在三个参照对象中都从未出现过的词汇，它们被收录进 CET 词表中的合理性值得思考。

在考虑这个情况之前，还需要厘清一个问题，即在 CET 词表中，中心词后单列的派生词的价值和科学性。根据大学英语四、六级考试大纲所言，“派生词原则上不单列(特别常用的除外)……如果形式上是派生词，而实际上已不被看成派生词，则单独列出。”一方面，“特别常用”的标准是什么，同时这些词在各类语料中的实际使用情况也不得而知。另一方面，“形式上是派生词而实际上不被看作派生词”指的是派生词已然与中心词没有了相似的词义，或是两者词义相差较大，如 remark 和 remarkable，proceed 和 proceedings 等。鉴于此，加之许多形式派生词的使用频率已然比 CET 词表中的中心词还要高，那么为何不将这些词单列出来也需要思量。

一直以来，在词表编制的选词单位这一问题上就存在争议，其中主要聚焦于是以词目(lemma)还是以词族(word family)为选词单位更为合理。词族由一个基本词及其所有派生和屈折形式组成，以词族为单位编制词表的底层逻辑就是英语学习者一旦知道基本词，二语学习者几乎毫不费劲地就可以理解该词族中其他派生形式词汇的意思，不必单独学习就可以理解(Bauer & Nation, 1993) [5]。基于此，许多词表选择词族来编制词表，以提供集成性词表。然而，有学者对以词族为单位编制词表表示质疑。一方面，属于同一词族的词汇意义并不总是相近的，如“reactionary”和“reactivation”两词同属一个词族，但两者核心意义却大相径庭(Gardner & Davies, 2014) [6]。此外，根据 Gardner (2007)和 Nippold & Sun (2008)等学者的观点，对于大多数学龄儿童和学习二语的成年人来说，派生词有关的知识更为复杂，它的接受要比屈折词有关的知识接受更晚。且以词族编制词表供二语学习者使用不利于为他们的词汇学习提供最直接的帮助[7] [8]。因此在近些年，研究词表的学者在编制词表时认为词目的使用应当优于词族。

回到本研究中来看，CET 词表形式上更偏向以词族为单位，但它并没有呈现中心词的所有派生词，而是只列举了所谓“特别常用”和“形式是派生但实际不被看作派生词”的派生词。一方面，就实际使用情况来说，CET 词表的派生词中并非全都“特别常用”，反而有的派生词使用频次不低却并未列入词表派生词中。另一方面，当一个词形式上是派生词，但实际上已经不被看作派生词，那么为何还要将其放在形式上的中心词下，这不免会误导词表使用者。

当加上所有列出的派生词再计算 CET 词表对三个参照对象的覆盖率，可以得到如表 7 的数据。把 CET 词表中心词和 2605 个派生词一同合并之后，词目总数达到 8008 词，CET 词表对三个参照对象的覆盖率理所当然地有所提升。但是与上述覆盖率数据相反，CET 词表派生词的词目覆盖率远远高于对词目累计总频次的覆盖率，这恰恰说明派生词的使用频率总体上要更低，累计词目覆盖量十分有限。不过实际观察对比词表会发现也有部分派生词的使用频次相较于有些中心词还要高，这也呼应了上述“选词单位”的争论[9]，以词目为单位编制词表，而非将高频派生词词目置于更低频中心词词目下，或许能更直观地让学习者了解高频词汇，减轻负担。

Table 7. The lemmas coverage of CET derivatives on the three reference objects
表 7. CET 派生词对三个参照对象的词目覆盖状况

				CET 派生词	覆盖率(%)
COCA5000	词目总数	4380	已覆盖词目数	855	19.52
	词目累计总频次	830,134,384	已覆盖词目总频次	39,415,997	4.75
CETC	词目总数	10,872	已覆盖词目数	1715	15.77
	词目累计总频次	382,778	已覆盖词目总频次	25,272	6.60
iWriteBaby	词目总数	17,936	已覆盖词目数	2194	12.23
	词目累计总频次	8,046,643	已覆盖词目总频次	404,846	5.03

注：COCA5000 行的覆盖率数据是基于 COCA 前五千词目的累计频次计算而得，不是 COCA 整库数据。

表 8 是 CET 词表派生词在 COCA5000 词表各频段的数量。CET 词表派生词只有 282 个在 COCA 前两千词的行列中，同时有 1750 个派生词并未出现在 COCA5000 里，而由于 COCA5000 本身是一个高频词表，其中词目最低的每百万标准词频约为 12，因而当与 COCA5000 对比时，CET 词表中与其不重合的词汇便可以看作低频词，其中中心词有 2259 个，派生词有 1750 个。

Table 8. The number of CET derivatives in each frequency band of COCA5000
表 8. CET 派生词在 COCA5000 各频段数目

COCA 频段	COCA 词目排名	CET 派生词在各频段总数
1	1~1000	83
2	1001~2000	199
3	2001~3000	248
4	3001~4000	236
5	4001~4380	89

同样地，与 CETC 和 iWriteBaby 语料库对比，CET 词表派生词中每百万标准词频 ≤ 10 的词数在两库中分别有 896 个和 1684 个，而其中位于两库词表前两千词行列的词数却远远低于低频词，分别只有 332 个和 303 个(见图 6)。

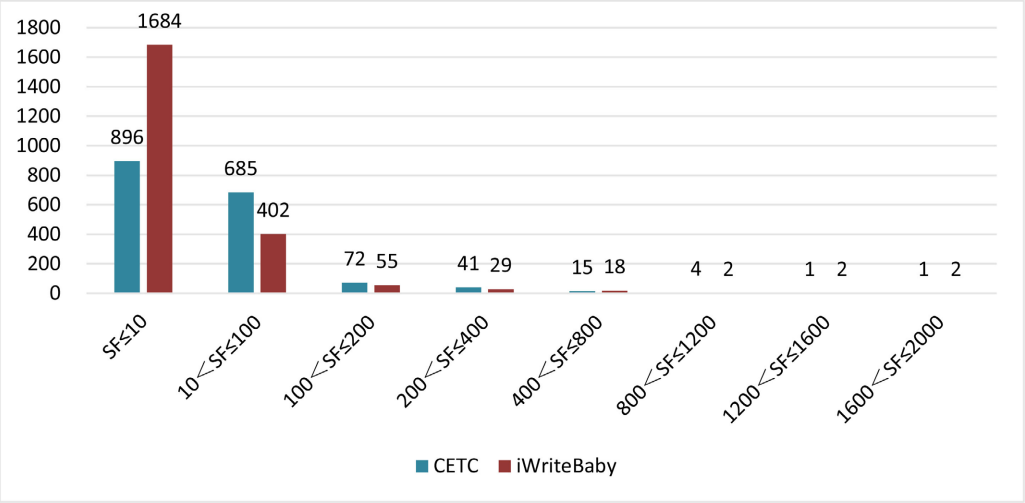

Figure 6. The number of shared words between CET derivatives and CETC and iWriteBaby corpus in each frequency band
图 6. CETC 与 iWriteBaby 语料库同 CET 派生词共有词各频段数量

当交叉分析提取上述所有对比组产出的低频词后，整理出表 9 内容。以“CET_ONLY(COCA)”和“CET_CETC”为例，它们分别表示“以 COCA 为参照 CET 词表的独有词”和“CET 词表与 CETC 的共有词”，第二行的其他内容也作同样解释。第一行的“=0”和“≤10”则分别表示“从未出现”和“每百万标准词频十及十以下”。由于 COCA5000 只有频次前 5000 的词目，而非原表，因此以 COCA 为参照的低频词数据只有“CET_ONLY(COCA)”，即 CET 词表中未出现在 COCA5000 中的词目，没有“CET_COCA5000”，即两表共有词中的低频词。但 COCA5000 词表中每百万频次最低是 12，因此它可以用于进一步印证 CET 词表在另外两库中的共有低频词并非高频。

Table 9. The total number of unused and low-frequency words in CET vocabulary list comparing the three reference objects
表 9. 对比三参照对象 CET 词表中未使用词及低频词总数

	F = 0			SF ≤ 10	
	CET_ONLY(COCA)	CET_ONLY(CETC)	CET_ONLY(iWriteBaby)	CET_CETC	CET_iWriteBaby
中心词	2259	975	261	1568	3050
派生词	1750	891	412	896	1684

在表 9 的基础上进行分析后发现，CET 词表中在 COCA5000、CETC 和 iWriteBaby 语料库都没有出现的词目数有 437 个，包括中心词 158 个和派生词 279 个。而 CET 词表在 CETC 和 iWriteBaby 语料库共有词中每百万标准化词频 ≤ 10 的词目数有 1497 个，其中有中心词 888 个和派生词 609 个。这些词汇不管在大学英语四、六级考试的卷面上、大学生词汇实际使用中还是本族语语料库中使用频率都十分有限，这有悖于 CET 词表的“定量原则”，应该考虑进行更新。

4. 结果与讨论

综上所述，鉴于大学英语四、六级考试的重要性，CET 词表对中国大学生具有重要意义。本研究通过三类语料库考查当前最新版 CET 词表的科学性与实用性，一方面在理论意义上，可以丰富国内在词表领域的研究成果，另一方面也具有实际应用价值，为进一步优化 CET 词表的编制提供启迪，以帮助提升国内大学生英语学习者的词汇学习效率和词汇水平。

在本研究中，通过分别参照 COCA5000、CETC 和 iWriteBaby 语料库对 CET 词表进行研究发现，尽管受三个参照对象词目总量差异的影响，CET 词表对它们的词目覆盖率差异较大，但其对词目累计频次的覆盖却并非如此，除了 CET 词表对 CETC 的词目累计频次覆盖率较之另外两个参照对象稍低，这一情况可能是受卷面语料特点的影响。词目覆盖率低但词目累计频次覆盖率高恰恰说明三个对比组中的大部分共有词在三个参照对象中都有高频使用，这些词汇主要包括部分常用功能词和通用核心词汇，与以往研究结果相吻合。

而在分析参照对象独有词时发现其中存在一部分非低频词，它们与 CET 词表中的 5403 个中心词并不重合，而是出现在中心词后的派生词中。由此引出关于 CET 词表编制的问题。CET 词表虽然形式上是以词族为单位，但它却并不像其他以词族为单位的词表，将派生词全都陈列出来，而是仅列出“使用频次极高”或“形式上派生但实际上不被看作派生”的词汇，这样的做法或许是考虑到减轻学习者的词汇学习负担。可一方面，许多派生词使用频率并不高甚至在本研究中的三个参照对象中都没有出现过，而有一部分派生词词频实际上非常可观但却并未列入表中。另一方面既然某些形式派生的词实际上已经不被看作派生词，将其以一种附属形式列在中心词之下可能会对学习者词汇学习产生误导。加之经对比后发现 CET 词表中心词也存在一部分在其他语料库中都从未出现过的独有词以及共有词中还有一部分每百万标准化词频 ≤ 10 的低频词，且还有部分中心词的实际使用频次远远不如意义相似的派生词，因而 CET 词表需要基于以上进行优化。

具体优化内容可以考虑基于词频，将已然不被看作派生词的非低频词和高频派生词同样单列作词表词目，然后将同时出现在三组对比中的独有词删去并把同样出现在三组共有词中的低频词进行精简化。此外，考虑到 CET 词表本质上是一个等级考试词表，该词表的编制在做到尽量精简的同时还要体现词汇的“等级性”。因此在优化 CET 词表时为了体现四级与六级的差异同时也为了满足部分学习者学习高级词汇的需求，词表需要对更高级词汇进行科学地选择与保留。而具体如何确定词量与选定词汇则需要进一步研究。

参考文献

[1] 许家金. iWriteBaby 中国学习者英语语料库的创建[J]. 语料库语言学, 2019(1): 105-109+117.

[2] Stoeckel, T. (2019) An Examination of the New General Service List. *Vocabulary Learning and Instruction*, **8**, 53-61. https://doi.org/10.7820/vli.v08.1.stoeckel

[3] Carter, R. (1987) Is There a Core Vocabulary? Some Implications for Language Teaching. *Applied Linguistics*, **8**, 178-193. https://doi.org/10.1093/applin/8.2.178

[4] 戴曼纯. 论第二语言词汇习得研究[J]. 外语教学与研究, 2000(2): 138-144.

[5] Bauer, L. and Nation, P. (1993) Word families. *International Journal of Lexicography*, **6**, 253-279. https://doi.org/10.1093/ijl/6.4.253

[6] Gardner, D. and Davies, M. (2014) A New Academic Vocabulary List. *Applied Linguistics*, **35**, 305-327. https://doi.org/10.1093/applin/amt015

[7] Gardner, D. (2007) Validating the Construct of Word in Applied Corpus-Based Vocabulary Research: A Critical Survey. *Applied Linguistics*, **28**, 241-265. https://doi.org/10.1093/applin/amm010

[8] Nippold, M.A. and Sun, L. (2008) Knowledge of Morphologically Complex Words: A Developmental Study of Older Children and Young Adolescents. *Language Speech and Hearing Services in Schools*, **39**, 365. https://doi.org/10.1044/0161-1461(2008/034)

[9] 汪晓琪, 刘建达. 英语词表类型和编制研究: 关键问题与趋势展望[J]. 外语界, 2024(2): 90-96.