大学英语作文自动评分中分级词表的有效性研究*

葛诗利

（北京语言大学语言信息处理研究所，北京 100083）

E-mail: geshili@blcu.edu.cn

摘要：自动作文评分中核心的问题是机器可用的、高信度的评分特征的选取。大学英语作文自动评分中一个基本的特征是词汇分布，而词汇分布的衡量需要一个对大学英语作文描述清晰、刻画准确的分级词表，也就是一个效度较高的词表。实验证明目前已有的词表难以达到这个要求，但通过改进或者调整现有词表可以得到对不同作文质量区分效度较高的词表。

关键词：自动作文评分，分级词表，效度，大学英语写作

The Validity of Word List in Automated Essay Scoring for College Students

GE Shi-li

(Foreign Language Department, Guangdong University of Finance, Guangzhou, Guangdong 510520, China)

E-mail: geshili@blcu.edu.cn

Abstract: The key in automated essay scoring (AES) is the selection of machine-usable and high-reliable features for scoring. Lexical distribution (LD) is a basic feature in AES for non-English major college students in our country. The scaling of LD needs a word list with high validity, which can only be achieved from the detailed study of the research subject—college English writing. Experiment proves that the wide available word list does not possess the validity, but with the adaptation, a new word list with a high validity for college English writing can be achieved.

Key words: Automated Essay Scoring (AES), Word List, Validity, College English Writing

1. 文献综述

自动作文评分是计算机技术在语言测试方面的最新应用，也是语言技术发展的必然趋势。自动作文评分中核心的问题是机器可用的、高信度的评分特征的选取。目前自然语言处理中准确率最高的、也是最基本的研究就是词汇分析。词汇分析一般包括词长分布、词汇分布和词汇丰富性等。

1.1 词汇分布和分级词表

词汇分布是指一篇作文中的词汇占某一分级词表每个级别词汇的比例。其中分级词表是一个衡量的标准，它相当于一把尺子或者一杆秤，其效度的高低直接关系到最后得到的词汇分布数据的有效性。

当今外语写作研究中使用最多的是 Laufer & Nation（1995: 307-22）的 3 个分级词表，本研究中称为“通用词表”。该词表根据词频的高低，将单词分为 3 个等级：1 级词汇、2 级词汇和 3

* 本论文的工作得到国家自然科学基金课题（编号 60272055、60572159）的资助。

级词汇。每一级别词汇有一个词表，即 Baseword list 1 到 3。其中所含词族和单词类符数量如表 1 所示。3 个词表以外的单词称为低频词汇（less frequent words），全部归入第 4 级。

表 1：Laufer & Nation（1995: 307-22）所用的 3 个词表

	Baseword list 1	Baseword list 2	Baseword list 3	合计
词族数量	998	987	570	2555
类符数量	4119	3708	3107	10934

1.2 通用词表的效度分析

国内大多数词汇分布研究所采用的词表都是 Laufer & Nation（1995: 307-22）的 3 个级别的词汇表。究其原因，一个是 Laufer & Nation（1995: 307-22）的实验严格地证明了使用该词表所刻画的词频概貌在测量作文中词汇分布方面的信度和效度都很好；另一个是国内词汇分布研究很多是面向英语专业学生，或者说高水平英语学习者，如李志雪、李景泉（2005: 56-59）和文秋芳（2006: 9-13）。这些研究对象的英语水平或者说词汇水平已经赶上或接近 Laufer & Nation（1995: 307-22）的研究对象。因此，这个词表也具有相似的信度和效度。研究所得的结论也比较类似，虽然非熟练学生更多地依靠第 1 级的 1000 个最常用词汇，但低级词汇作为基本的产出性词汇对英语水平的区分度不大，高级别词汇才能真正反映写作者的英语水平。

然而刘东虹（2003：186）利用该词表所做的面向非英语专业的学生的研究却得到了不同的结论，即“写作能力强的学习者第一类词的使用量较多”。刘东虹（2003：184-85）的统计显示，“低分组...使用的第一类词较少，但使用复杂词较多”，但“t 检验结果显示两个写作组...第一类 1000 词...上的差异具有显著意义，而在运用复杂词方面没有显著差异”。这说明非英语专业学生，不管是写作能力强弱，都不能够真正掌握和使用复杂词汇。因而其能力的差别就集中体现在对最基本的 1000 词汇的运用上，高级别词汇无区别意义。

1.3 研究假设

以上利用同一个词表的词汇分布研究却得出了截然相反的结论。其原因可能就是其 1 级词表的 998 个词族、4119 个类符对我国大学英语学习者来说已经不仅仅是基本词汇了，可能只有较好或者最好的学习者才能真正掌握和灵活运用这些词汇，而很多中等或者较差的学习者只能使用其中的一部分。据此，我们提出以下研究假设：

1）通用词表各个级别的词汇在大学英语各分数档作文中的分布类似，不具有显著性差异。

2）依据通用词表提取大学英语各分数档作文的词汇分布特征，并利用此特征进行作文的自动评分，难以获得较好的评分效果。

对于研究假设 1，四个词汇级别中，越多级别的词汇分布具有显著性差异，词表的效度就越高；对于同一级别的词汇，五个分数档中，两两之间存在差异的分数档数量越多，词表的效度也就越高。对于研究假设 2，较好的评分效果能够证明词汇分布特征对大学英语写作不同的分数档的作文具有较好的区分度，也间接地证明了词表的效度。

2. 研究设计

2.1 作文语料来源、加工和处理

本研究的语料主要来自中国学习者英语语料库（CLEC）的 st3 子库（选取 1050 篇），部分取自关兴华、陈建辉（2005）编著的《大学生英语作文》。St3 子库中大部分是大学英语四级考试作文，存在着大量的计算机不能够自动处理的拼写错误。由于这些错误很多并不妨碍阅卷人对文章的理解。因此，为了不影响计算机处理，对这些错误进行人工改正。错误的人工改正主要是以 MS Word 2003 的拼写检查器为标准，能够发现错误并且所提供的改正中有正确选项的，进行改正，否则不作改动。

由于做过拼写改正，并且为了增强作文分级的准确性，1050 篇作文全部重新评分。评分标准按照四级考试标准执行。鉴于研究的探索性质，只给出五个档次分，而不上下浮动。即只有 2、5、8、11 和 14 五个分数档。评分是由 3 位高校英语教师在 7 天内完成。首先由两位教师分别评分，每天的工作量在 150 篇左右，评分结果相同的，作为该作文最后得分。评分结果不同的，交由第三位教师复评。如果第三位教师评分与前面某一位评分相同，该分数作为最后得分。如果三次评分跨 3 个分数档，则中间档作为最后得分。还有一种极少出现的情况就是前两个评分相差两档，而第三次评分高于最高评分或者低于最低评分，即三次评分跨 4 个分数档，这样就保留到评分结束后由三位教师讨论给出最后评分。

1050 篇作文初评有 357 篇分数不同（分数差异 2 档的 54 篇，无 2 档以上差异），进入复评，最终确定每篇作文的得分。初评中两位评分员之间（inter-rater）的相关系数为.761。评分完成后从前 3 天评分的作文中随机选取 30 篇作文再次由两位评分员评分，计算评分员内部（intra-rater）的相关系数分别为.712 和.784，并且以上相关系数均在.01 水平（双侧）上具有显著性。这说明两位评分员的作文评分已经比较准确，再加上第三位评分员最后的核定，作文的最后得分应该足够客观和准确，可用于本研究中自动评分模型的训练和测试。作文的分数分布见表 2。

表 2：1050 篇作文的分数分布

	2 分	5 分	8 分	11 分	14 分	总计
数量	67	443	329	158	53	1050

5 个分数档的作文中，2 分和 14 分的明显偏少。故由前两名评分教师一起从《大学生英语作文》中顺序挑选 97 篇文章加入 14 分档。但由于 2 分作文难以获取，故只选择其中的 60 篇。5、8、和 11 分档各选择 150 篇。所有 660 篇作文构成此次研究的语料。其中每个分数档的 2/3 的 440 篇作为训练集，1/3 的 220 篇作为测试集。

2.2 数据统计与分析

1）编写程序，把所有 660 篇作文拆文成词。缩写全部拆分，还原。作为属格的’s 记为一个类符。

2）统计所有词汇，以词表和大型语料库为辅助人工区分正误。不是常用的由连字符构成的复合词全部拆分。

3）以 Laufer & Nation（1995: 307-22）的 3 个词表为标准把词汇分为 3 级，正确的表外词汇为第 4 级，错词为第 5 级。其中所有的数字识别归并为一个类符，所有首字母大写的专名识别归并为一个类符，以及作为属格的’s 全部计入第 1 级词汇。

4）由于第 4 级词汇没有词族词表，全部统计之后，人工设计该词表，总计 783 个词族，951 个类符。错误词数量 166 个，由于数量较少，在此忽略不计。

5）统计全部660篇语料的词汇分布百分比。

6）如果数据满足方差齐性检验，运行SPSS软件包的单因素方差分析（ANOVA），否则运行非参数检验的多个样本比较的秩和检验（Kruskal-Wallis法，也称H法）分析判断各分数档作文之间在除错误词汇的4个词汇级别上有无显著差异。

7）以训练集440篇语料的词汇分布百分比输入SPSS，进行多元线性回归，求得回归系数，得到回归方程。然后编写程序，把测试集中220篇语料的词汇分布百分比代入方程，求得预测分数。最后，比较预测分数与原始分数，求得评分准确率。

2.3 自动评分性能评价标准

在本研究中，自动评分效果是通过以下参数的计算来进行比较：各分数档评分的精确率和召回率，总体准确率和误判率。分别定义如下：

各分数档精确率 = 本为X档作文并且被评为X档作文的数量 / 所有被评为X档作文的数量 * 100

各分数档召回率 = 本为X档作文并且被评为X档作文的数量 / 本为X档作文的总数 * 100

总体准确率 = 测试集自动评分准确的作文数量 / 测试集作文总数 * 100

总体误判率 = 自动评分成绩与原分数档相差一个以上分数档作文数量 / 测试集作文总数 * 100

各分数档精确率越高，说明作文被评为该分数档的可信度越高。例如，某一分数档X准确率为50%，作文A被评为X档，那么，作文A确实属于X档的可能性为50%。

各分数档召回率越高，对总体准确率的贡献越大。

总体准确率越高，说明自动评分模型越准确，模型愈可付诸实用。

总体误判率越低，说明自动评分模型不可接受的误差越小，模型付诸实用的可接受性也就增强了。

3. 结果与讨论

根据Laufer & Nation（1995: 307-22），文章的词汇分布的计算是以词族为计算单位。采用该方法的研究还有李志雪、李景泉（2005: 56-59）。当然，也有以类符（type）为计算单位，如文秋芳（2006）和刘东虹（2003: 180-87）。以及以形符为计算单位，如倪岚（2000: 38-41）。3种计算方法各有特点，本文采用Laufer & Nation（1995）的方法。

3.1 词汇分布数据分析

表3是以通用词表为标准，4个词汇级别在5个分数档共660篇作文语料中的词汇分布的平均数和标准差统计数据及显著性检验。该表显示，1、2级词汇在作文中的分布存在差异，进一步的LSD两两比较表明，14分档作文与其它分数档作文在这两个级别词汇的使用上有显著性差异（P<.05），而其他各分数档之间没有差异。观察该表可以看出，14分档的作文非常明显地少用1级词汇，差距达到5%左右，而其他分数档的作文没有很大差异。这表明在大学英语写作中，除了极少数写作水平非常高的学习者，绝大多数学习者的写作都严重依赖Laufer & Nation（1995: 307-22）所划分的1级词汇。对比其他分数档，14分档的作文在2级词汇的使用上明显增加，

而其他分数档对2级词汇的使用都较少。

表3：通用词表的词汇分布

		1级词汇(平均数/标准差)	2级词汇（平均数/标准差）	3级词汇（平均数/标准差）	4级词汇（平均数/标准差）
2分档		87.22/4.63	6.56/3.20	2.39/2.21	3.83/2.38
5分档		86.92/4.30	6.95/3.10	2.64/2.29	3.49/2.13
8分档		86.15/4.64	6.41/2.66	3.75/2.43	3.69/1.95
11分档		85.77/4.59	6.83/3.26	3.70/2.77	3.70/2.17
14分档		81.15/5.45	7.55/3.05	5.62/3.11	5.68/2.92
Levene	P	.144	.562	.010	.000
ANOVA	F	35.066	4.327		
	P	.000	.002		
Kruskal	H			110.91	61.23
-Wallis	P			.000	.000

除了1、2级词汇的分布之外，其他词汇分布均不满足Levene方差齐性检验的要求，只能使用非参数检验的多个样本比较的秩和检验分析判断各分数档作文之间在这些词汇级别上有无显著差异。尽管秩和检验证明这些词汇分布在不同分数档作文间具有显著性差异（P=.000<.01），但由于该检验的局限性，并不能够准确判断到底哪些分数档之间存在差异。较准确的差异判断需要采用线性回归来观察每一特征在回归方程中的作用。

也就是说，目前能够判定的只是14分档作文在1级和2级词汇的使用上与其它分数档作文存在显著性差异，而其它所有分数档之间要么不存在差异，要么难以判断。由于共有5个分数档，对于每个词汇级别，有10种两两比较的关系，4个词汇级别共有40种比较关系。其中只有2个级别的词汇在14分档作文和其它4个分数档作文的分布，共8种关系中，存在差异。只占关系总数的1/5，所以这个结果基本证明了研究假设1。

3.2 通用词表词汇分布特征的自动评分效果

计算得到训练集的4个级别词汇分布百分比，输入SPSS，以每篇作文的分数作为因变量，各级别词汇的百分比作为自变量，进行多元线性回归，求得回归系数，得到回归方程。回归方程的方差分析显示，方程具有显著性。观察各级别词汇的回归系数可以发现，3级词汇未进入方程，即3级词汇不具有任何区分度。从偏相关系数看，4级词汇影响最大，其次是1级词汇，最小是2级词汇。Laufer & Nation（1995: 307-22）的研究结论，“文章真正的词汇水平应取决于其他更高级别的词”指的是1级词汇之外的2、3级词汇，但是这里的统计数据说明2、3级词汇很少甚至不起作用。

由此可以进一步验证研究假设1，即虽然在1、2级词汇的分布上14分档作文与其它分数档有着显著性差异，但是这种差异作用甚微。从线性回归统计方法来说，4级词汇在各分数档作文中的分布差异性比较大。但是由于3、4级词汇在所有作文中的数量都很少，其分布特征能否在自动评分中起到作用尚需验证。

根据训练集得到了词汇分布到分数的回归方程，统计测试集中220篇语料的词汇分布百分比，代入方程，求得预测分数。最后，比较预测分数与人工评分，评分总体准确率是25.45%。

因为测试语料由2分作文20篇和其它分数档作文各50篇构成，如果所有作文都评为篇数最多的一个分数档，就构成了评分总体准确率的底线（baseline），也就是50/220=22.73%。词汇分布的线性回归方法超过底线不足3%。这说明通用词表所统计的大学英语四级考试作文的词汇分布对作文得分预测力非常弱。该方法对各分数档作文评分的精确率和召回率以及总体准确率和误判率如表4所示。

表4：通用词表词汇分布的自动评分性能

	2分档	5分档	8分档	11分档	14分档
精确率	-	17.24	30.56	21.13	25.00
召回率	.00	10.00	66.00	30.00	6.00
总体准确率	25.45				
总体误判率	24.55				

词汇分布的线性回归评分方法总体准确率太低，总体误判率太高，不具有使用价值。研究问题2得到了肯定的答案，它说明依据通用词表的词汇分布统计数据对大学英语写作的作文质量几乎没有预测力，即通用词表在大学英语写作评分上效度极低。如何改进词表，各分数段作文才能在各级别词汇上有一个更清晰的、具有显著性差异的描述，换言之，如何改进词表，词汇分布才能更好地、比较清晰地区分各分数段作文就成为一个重要的研究方向。

4. 词表改进方法与评测

只有能够把不同分数档作文之间词汇分布的差别表现出来的词表才适合用来描述大学英语写作能力的差异，这样的词表可以通过以下方法获取：

1）由大规模大学英语写作语料库按词频分布统计得到。这需要一个规模大、代表性强的语料库，但目前尚难以得到。

2）由大型通用语料库的词频统计得到。但需要透彻研究我国大学英语写作，制定的词汇分级能够精确反映写作质量。

3）可利用现有词表，通过研究大学英语写作进行调整，即适当减少1级词汇的数量，调整2级词汇数量，增加3级词汇数量。

本研究采用第3种方法，即基于Laufer & Nation（1995）的词表，结合所要研究的作文语料来获得改进的词表。具体步骤如下：

1）统计训练集440篇作文中各分数档作文的词汇频率。

2）编程提取11分和14分档作文中出现的而其他分数档作文中出现次数不超过某一阈值的词汇。

3）这些词汇如果出现在Laufer & Nation（1995）的1级词表中，就把该词族提升到2级词表，如果出现在2级词表中，就提升到3级词表。

这里的阈值是一个经验数值，本研究中取15。最后得到3个改进后的词表，各级别词表中词汇情况见表5。其中1级词汇缩减过半，2、3级有所增加。如果这个词表能够较清晰地描述大学英语四级考试作文不同分数档的词汇分布，将说明在大学英语写作中，大多数学习者能够掌握而熟练运用的基本词汇比国外的英语学习者，或者国内英语专业学习者的基本词汇要少得多。

表 5：改进后词表

	TBaseword list 1	TBaseword list 2	TBaseword list 3	合计
词族数量	424	1140	991	2555
类符数量	1357	4681	4896	10934

采用与 3.1 和 3.2 节相同的研究步骤，发现高分作文较少使用 1 级词汇，更多地使用高级词汇，2 级词汇对作文质量的区分度似乎不大。但由于大部分数据方差齐性不满足，未能进行有效的两两比较，差异的显著性需要在回归分析中进一步验证。

根据训练集的词汇分布数据，得到回归方程。1 级词汇没能进入方程，就是说，1 级词汇对作文成绩的影响力微乎其微。换言之，1 级词汇在各分数档作文间的分布不存在显著性差异。从统计的偏相关系数看，3 级词汇对分数的影响最大，4 级次之，2 级最小。也就是说，3、4 级词汇在各分数档作文间的分布差异比较显著。这更接近于 Laufer & Nation（1995）的研究结果，也说明改进后的词表对不同写作质量的大学英语作文区分效度更大了。

统计测试集作文的词汇分布百分比，代入方程，求得预测分数。比较预测分数与人工评分，评分总体准确率达到 32.27%。比使用通用词表的方法提高了 6%以上。这表明改进词表的效度有了较大的改善，据此获得的词汇分布特征对作文得分的预测力增加较大。

本文采用的改进词表法相对比较简单，也取得了较好的实验效果。但这种方法有一定的局限性，由于数据只是从数百篇作文中统计得出，针对同一批作文的自动评分，采用此方法调整词表，可能取得较好的评分效果，大范围推广使用的效果却难以保证。但这项研究至少证明，在我国大学英语作文的自动评分中，只有利用有针对性的、高效度的分级词汇表获取词汇分布特征，才能取得一个较好的评分效果。这个词表可以通过本文提出的前两种方法统计获得，经广泛验证后推广使用。

参 考 文 献

[1] Laufer, B. & P. Nation. (1995). *Vocabulary Size and Use: Lexical Richness in L2 Written Production* [J]. Applied Linguistics, (3).

[2] 关兴华，陈建辉. 大学生英语作文[M]. 吉林大学出版社. 长春：吉林大学出版社，2004.

[3] 李志雪，李景泉. 中国高水平英语学习者产出性词汇使用情况研究—基于对中美大学生英语作文范文的对比分析[J]. 山东外语教学，2005，(5).

[4] 刘东虹. 词汇量在英语写作中的作用[J]. 现代外语，2003，26 (2).

[5] 倪岚. 对英语专业二年级学生写作词汇的研究[J]. 国外外语教学，2000，(2)：38-41.

[6] 文秋芳. 英语专业学生使用口语-笔语词汇的差异[J]. 外语与外语教学，2006，(7)：9-13.