中美学者学术英语写作中词汇难度特征比较研究
——以计算语言学领域论文为例

谢永慧　刘洋　杨尔弘✉　杨麟儿
北京语言大学　北京　100083

摘要

学术英语写作在国际学术交流中的作用日益凸显，然而对于英语非母语者，学术英语写作是困难的，为此本文对计算语言领域中美学者学术英语写作中词汇难度特征做比较研究。自构建1132篇中美论文全文语料库，统计语料中484个词汇难度特征值。经过特征筛选与因子分析的降维处理得到表现较好的五个维度。最后计算中美学者论文的维度分从而比较差异，发现美国学者的论文相较中国学者的论文中词汇单位更具常用性、二元词串更具稳固性、三元词串更具稳固性、虚词更具复杂性、词类更具关联性。主要原因在于统计特征值时借助的外部资源库与美国学者的论文更贴近，且中国学者没有完全掌握该领域学术写作的习惯。因此，中国学者可充分利用英语本族语者构建的资源库，从而产出更为地道与流利的学术英语论文。

关键词： 学术英语写作；词汇难度；中美学者；计算语言学

A Comparative Study of the Features of Lexical Sophistication in Academic English Writing by Chinese and American Scholars: A Case Study of Computational Linguistics

XIE Yonghui,　LIU Yang, YANG Erhong ✉,　YANG Liner
Beijing Language and Culture University
Beijing 100083, China

Abstract

Academic English writing plays an increasingly prominent role in international academic exchanges, and for non-native English speakers, academic English writing is difficult. Therefore, here is a comparative study of lexical sophistication features in academic English papers by Chinese and American scholars in the field of computational linguistics.A corpus of 1132 Chinese and American papers was constructed and 484 lexical sophistication eigenvalues were counted. After feature selection and factor analysis, five dimensions with better performance were obtained. Finally, we calculated the dimensionality scores of the papers of Chinese and American scholars to compare the differences, and found that compared with Chinese scholars, American scholars' papers have more common lexical units, more stable bigrams, more stable trigrams, more complex function words, and more relevant parts of speech. The main reason is that the external resource databases used for eigenvalue statistics is closer to the papers of American scholars, and Chinese scholars do not fully grasp the habit of academic writing in this field. Therefore, Chinese scholars can make full use of the resource databases constructed by native English speakers to produce more authentic and fluent academic English papers.

第二十届中国计算语言学大会论文集，第454页-第466页，呼和浩特，中国，2021年8月13日至15日。

Keywords: academic English writing , lexical sophistication , Chinese and American scholars , computational linguistics

1 引言

作为学术研究和成果交流的国际通用语，英语正在成为一种重要的学术技能(Hyland, 2009)。随着国际学术交流的不断深入，各学科领域越来越多的研究人员需要具备学术英语写作能力。然而，对于英语非母语者来说，达到母语者的水平是困难的，甚至被认为是不可能的(Haixiao and Clifford, 2011)，在复杂的学术英语写作环境中，他们经常处于劣势地位(Uzuner, 2008; Flowerdew, 2009; Huang, 2010)。

中国学者也不例外，虽然影响学术英语写作的因素有很多，如：写作经验、学术素养、体裁知识、语言问题(Zhao, 2017)，但中国学者经常发现自己的研究论文因语言问题而被科学期刊拒绝或需要重新提交(Yonggang, 2006)。这与中英语言、文化等各方面存在差异性，会产生负迁移有关。诚然，学术英语论文中语法错误是可以避免的，却不能保证被接收，需要更进一步强化作者的语体意识，提升论文语言上的流利性与地道性。为了弥补这样的差距性，有必要对母语者与非母语者的学术英语写作中的语言表达展开充分的研究。

2 相关研究

学术英语写作研究作为写作研究领域的一个重要分支，始自二十世纪八十年代(Xu, 2015)，集中于ESP（English for specific purposes）和EAP（English for academic purposes）两个领域(姜亚军and 赵刚, 2006)。考察的对象主要是学术语篇与相关写作实践活动，由此分化出学术英语写作研究的两大视域：系统功能语言学视域、学术语言能力理论视域。系统功能语言学视域下集中对各种形式的学术文本做研究，主要包括期刊论文、学位论文、书评、课题申报书和教材等。

具体而言，系统功能视域下多采用对比手段研究不同群体、学科、体裁的论文文本中语言特征、篇章结构等的差异和影响因素。不同群体包括母语者与非母语者的学生群体，母语者与非母语者的研究工作者，不同英语水平的群体(Chen and Baker, 2010; 王琴, 2020)；学科之间的对比研究涉及生物、历史、应用语言学、医学等学科(赵怡琳, 2018; 王彤, 2019)；研究的书面语体裁又包括完整的学位论文、期刊和会议论文，以及论文的一部分，如摘要、引言、结论、致谢等，鉴于语料采集问题，以摘要的研究为最多(孟凡玲, 2018; Lu and Deng, 2019; 余龙幸, 2019)。这些研究从不同的角度对学术英语写作提供了启示与见解，也体现出学术英语论文的语体变异性会受写作群体、文本领域和体裁等的影响。

因此，本文限定写作群体为学者，文本领域为计算语言学，体裁为学术论文全文，从系统功能语言学理论出发着重对词汇难度（Lexical sophistication）的相关特征在中美学术英语写作中的使用做比较研究。词汇难度通常指高级词汇和难词的产出(Laufer and Nation, 1995)，其不仅是英语母语者和非母语者写作质量的重要影响因素(Bachman et al., 1996; Crossley and McNamara, 2012)，也针对不同的语料已被证明是学术英语写作的一个重要指标。Guo et al. (2013)对新托福考试基于学术环境产出的两类写作（独立写作和综合写作）中词汇难度特征作比较，发现实词熟悉度和实词词频是综合写作文本的重要预测指标，每个单词的平均音节和名词上位关系是独立写作文本的重要预测指标。Crossley et al. (2019)探究了四门学科学生学术英语论文的差异，发现物理和机械工程专业的论文比生物和工业工程专业包含更多的复杂虚词；工业工程和物理专业的论文比机械工程和生物专业更简洁，包含更多高频常见的单词和短语；机械工程专业的论文相比较其他三个专业的论文实词词频更低。而未有将词汇难度引入对计算语言学写作的研究。

词汇难度不仅可依据词频进行测量，还可通过词义丰富度、词分布、词的熟悉度等多样化可操作的指标来测量，相应的词汇难度自动分析工具也随之开发，例如Vocabulary profile(Yoon et al., 2012),Text Inspector(Bax, 2012),CTAP(Chen and Meurers, 2016)和TAALES(Kyle et al.,

项目基金：国家语委“十三五”科研规划2020 年度重点（科研中心）项目（ZDI135-131），国家语委信息化项目（ZDI135-105），北京语言大学语言资源高精尖创新中心项目(TYZ19005)

2018)等。在文本分析中，词汇难度的微观特征过多，可能造成信息的重叠，也不利于把握语言特征反映文本的整体特征。依据不同的学科或理论框架可对词汇难度做不同的分类，从计算机的角度可依据语言特征的共现性对其降维和聚类，Biber (1988)创建的多维度分析法便是这样的分析研究框架。起初Biebr利用因子分析法对口语和书面语语域中67种语言特征的“共现”模式归类分析，归纳出具有区分作用的五个维度。随着多维分析法的广泛应用，其成为变异研究的重要研究范式，包括历时变异研究、跨语言变异研究和学术话语变异研究等。此外，Kim et al. (2018)及Crossley et al. (2019)通过因子分析法对词汇难度特征降维处理并分析，验证了从计算角度对词汇难度特征进行降维的可行性，因此本文将多维分析法扩展应用于中美学术论文的对比研究中。具体的研究问题如下：

1）基于中美学者计算语言学英语论文语料能否将微观的词汇难度特征降维成宏观的特征集，每个特征集能否将中美学者的论文区分开来?

2）降维后的特征集反映了中美学者词汇难度上的哪些差异，背后的原因有哪些，能给中国学者带来怎样的启示?

3 实验过程

本文总体的实验过程如下：

1）构建中美学者计算语言学领域学术英语论文语料库，并使用词汇难度自动分析工具TAALES提取语料中的语言特征。

2）剔除没有正态分布和具有多重线性的语言特征；依据语料和语言特征训练随机森林分类器，选取分类准确率的重要性排名中排名较高的语言特征。

3）对选取的语言特征做因子分析，为每个维度做解释性的命名。

4）计算两类文本在每个维度上的维度分，从而做出比较，并依据方差分析与SVM分类准确率进行验证。

3.1 语料库的构建

学术论文是自然语言处理(NLP)研究中越来越重要的文本领域。虽没有可直接应用于机器处理的中美论文语料，但多样的学术论文语料库也在不断构建，为我们获取中美学者论文提供了途径。本文计算语言学领域的论文选自S2ORC(Lo et al., 2019)，该语料库中包括多领域的论文全文及丰富的元数据信息，例如：作者姓名、标题、发表年份、来源库等，并以JSON格式存储。但S2ORC并没有提供作者的国籍或所属机构信息，故需要进行筛选。通过Python编程，依据元数据信息提取对应的中美学者近十一年的计算语言学领域的论文全文的过程如下：

1）设定元数据中的年份为2010年至2020年，论文来源库为ACL Anthology[0]，筛掉不符合条件的元数据。ACL Anthology是计算语言学领域论文的数字档案，它包括权威的《计算语言学》杂志（Computational Linguistics journal）的论文，以及许多相关会议及研讨会的文献，如ACL, EACL, NAACL, ANLP等，选取ACL Anthology的论文可作为计算语言学领域的代表。

2）构建中、美的姓和名列表匹配元数据中每篇论文作者的姓名，据此确定每位作者的身份，只有当一篇论文中每位作者的姓名均为中国或者美国的普遍人名时，才提取该论文的全文[1]。其中，中国学者的姓氏取自《汉姓罗马字标注》中中国大陆的509个姓氏，名由人工拼读来判断，可进行汉语拼读即认定符合要求；美国学者的名取自美国人口普查局1990年美国人口普查[2]数据中最常用的900个男性的名字和3000个女性的名字（分别占男女人口的88.59%和88.614%），姓氏取自2010年美国人口普查数据中最常用的4999个姓氏（占美国人口的57.75%）。这一筛选方法虽然不能保证作者国籍准确无误，但能够有效确定作者的母语背景[3]。

[0]https://www.aclweb.org/anthology/

[1]此方法借鉴于Lu and Deng (2019)和Wood et al. (2001)对英语本族语者和英语非本族语者的区分方法。

[2]https://www.namecensus.com/

[3]为了评估筛选后论文母语背景的准确性，从中美论文中各随机抽取100篇检验其准确率。以美国学者论文为例，首先通过网络检索作者国籍确定其母语背景，对于未能检索到国籍信息的作者，利用ANN(ACL Anthology Network)(Radev et al., 2013)中提供的作者单位信息来判定，工作单位在美国则判定该作者拥有美国学者同等学术写作能力，否则，再通过网络检索其学士、硕士、博士所属机构，若本硕博三个阶段中有两个及以上阶段在美国就读，则也判定其拥有美国学者同等学术写作能力，否则判定为非美国学者。结果发现，美国学者论文中，有二十篇论文的作者不是美国国籍、没有获得美国国家的学位且不在美国工作，故美国学者论文的筛选准确率为80%，但其

经过筛选得到美国学者论文566篇[4]，中国学者论文2305篇，为了保证数据的平衡性，最终选定中美学者的论文均为566篇，语料总规模为1132篇，时间范围是2010年到2020年，语料库规模如表1所示。

类别	篇章数	总词数	平均词数
美国学者论文语料库	566	1929902	3409.7
中国学者论文语料库	566	1962351	3467.1
总计	1132	3892253	3438.4

Table 1: 语料库规模

3.2 语言特征提取

词汇难度自动分析工具TAALES已被广泛应用于写作评估(Bestgen, 2017)、文本可读性(Crossley et al., 2017)等研究中，具有研究的适用性。本文使用最稳定的版本TAALES 2.2，该版本共包含13类484个指标：词频（Word Frequency）、词分布（Word Range）、心理语言学特性（Psycholinguistic Norms）、习得/接触年龄（Age of Acquisition/Exposure）、学术语言（Academic Language）、语境多样性（Contextual Distinctiveness）、词识别特性（Word Recognition Norms）、语义网络（Semantic Network）、Ngram频率、Ngram分布、Ngram关联强度（Ngram Association Strength）、词家族信息（Word Neighbor Information）、字母频率（Character Bigram Frequency）。这些词汇难度的特征也已被证明与写作质量或阅读能力相关。本文将中美论文共1132篇语料输入TAALES工具，提取得到了484个词汇难度相关特征。

3.3 语言特征筛选

在确定了一组初步的特征之后，还需其他基于因子分析的统计因素的考虑(Egbert and Staples, 2019)，包括正态分布与多重共线性。正态分布是统计分析的前提与基础。本文对484个特征的1132篇文本绘制直方图以检查其正态性，任何非正态分布的特征都不再考虑，筛掉54个特征，剩余430个特征。对具有多重共线性特征的筛选，可避免特征重复，因为在极端情况下对相同的变量进行了两次测量，将会对因子分析的各个方面产生影响，包括数据的可解释性。本文通过相关系数来检验变量之间的多重共线性，设定相关系数为.900[5]，任何两个变量超过.900，依据其方差膨胀因子（VIF），剔除方差膨胀因子较大的特征，共排除280个特征，剩余150个特征。

为选择更好的具有区分中美论文的语言特征进行因子分析，本文使用随机森林模型对剩余的150个特征做进一步筛选。随机森林有如下优点：在目前的算法中具有较高的准确率；能够评估变量对分类的重要性；在运行过程中能够产生泛化误差的内部无偏估计——袋外误差估计等。将150个语言特征频率输入随机森林进行训练，得到的袋外分数[6]为.789，选取在150个特征中分类准确率的重要性排名较高的前49个特征重新进行训练，得到袋外分数为.806。故最终确定使用排名较高的49个特征。

3.4 因子分析

本文基于SPSS.23软件对49个语言特征做探索性因子分析，以将高度相关的特征变量聚集成组。首先，需要确定变量的可解释性，依据KMO抽样适当性和巴特利特球形检验进行测量。KMO值为.817，表明这些变量之间的偏相关系数高，进行因子分析的效果好；巴特利特检验显著性p值小于.05，表明变量高度相关，足够为因子分析提供合理基础。

采用主成分分析方法提取因子，如图1所示，碎石图在6因子处有明显拐点。综合比较5、6、7因子方案来选取最优方案，对三个方案中方差解释的特征值累积、各因子的特征

中只有两篇的作者不是英语母语者，故英语母语者的筛选准确率为98%；中国学者的判定方法相同，其中有九篇论文的作者不是中国国籍、没有获得中国国家的学位且不在中国工作，语料筛选的准确率均为91%，综合来看，本文的语料筛选方法具有可行度。

[4]由于对一篇论文中每位作者的姓名均进行了限定，使得美国学者与它国学者跨国合作的许多论文被筛掉，最终美国学者的论文数量有限。

[5]文本在处理多重共线性时相关系数设置较高，是因为后续还进行了严格的筛选，所以在这一步中可尽量保全更多的特征。

[6]袋外分数是袋外估计准确率得分，即对袋外样本正确预测的比例，可以反应模型的泛化能力。

值和共享方差的百分比、各因子包含的特征数量作比较，最终选定最为合适的6因子方案，即将49个语言特征划分为六个维度。如表2所示，采用6因子方案，方差解释初始特征值累积为61.012%，说明这6个因子可解释全部语料的61.012%[7]。

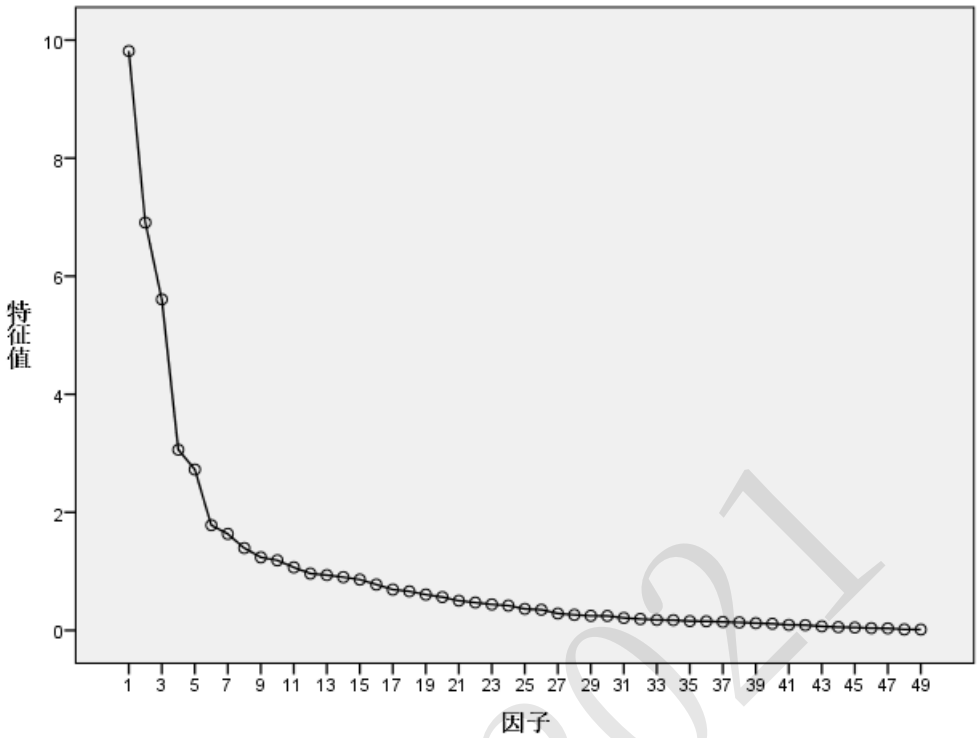

Figure 1: 碎石图

因子	总计	方差百分比	方差累积百分比
1	9.814	20.028	20.028
2	6.908	14.098	34.126
3	5.607	11.443	45.569
4	3.060	6.245	51.814
5	2.725	5.561	57.376
6	1.782	3.636	61.012

Table 2: 总方差解释中的初始特征值

采用凯撒正态化斜交法进行因子旋转，以简化结构，构建了潜在因子，增加因子的解释能力。为保证量表的构念效度，排除因子负荷绝对值小于.35的变量(雷蕾, 2016)。为了保证因子得分的实验独立性，每个语言特征都被包含在其具有最高负荷值的因子中。共排除4个语言特征，得到45个有效的语言特征。

4 文本维度分析

以上通过因子分析识别出了六个维度高频共现的语言特征集，且每个特征都有相应的负荷值，负荷值越高代表其与该维度的相关性越强，对该维度越重要。每个维度既可以包含一组特征集，也可以包括两组互补的特征集，负荷值为正的特征归入“积极”特征组，负荷值为负的特

[7]按吴明隆 (2010)，因素分析时，由于以少数的因素构念来解释所有观察变量的总变异量，加上行为及社会科学领域的测量不如自然科学领域精确，因而萃取后保留的因素联合解释变异量若能达到60%以上，表示萃取后保留的因素相当理想，如果萃取后的因素能联合解释所有变量50%以上的变异量，则萃取的因素也可以接受。

征归入“消极”特征组，当一组特征频繁共现在文本中时，另一组特征则极少在文本中出现，反之亦然(比伯et al., 2012)。然而，上述界定的六个维度能否将中美学术论文区分开来，还需要计算两类文本的“维度分”来加以验证，即每一维度中正负荷值语言特征的标准分之和减去负负荷值语言特征的标准分之和，并且每一个特征的标准分都经过了不同负荷值的定量加权，维度分的统计结果如图2所示。

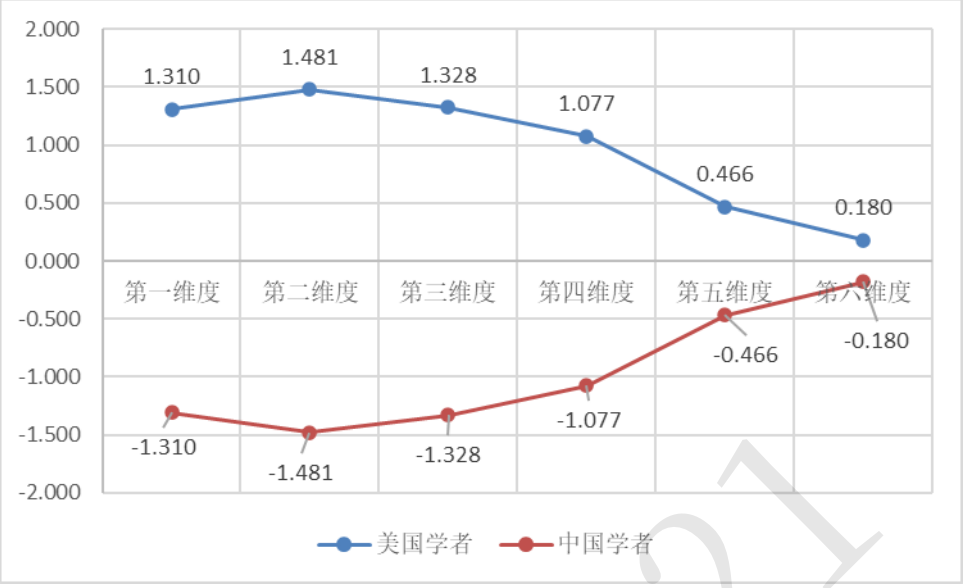

Figure 2: 中美学者论文维度分

此外，本文运用方差分析与SVM模型对以上结果进行检验。为了确保维度分数差异的显著性，对每个维度上所有文本得分的均值进行单因素方差分析，本文选择Welch T 检验比较差异，因为Welch T 检验是对T检验的改进，使其能够在两组数据样本容量和方差不一致的情况下检验两组数据均值之间的差异，且效果优于T检验(Zimmerman, 2004)。为了验证以上研究所发现结果的准确性，通过SVM分类模型计算每个维度对中美学术论文分类的准确率与F值。即在带有语言特征标注的训练数据集上训练生成SVM分类器，并将此分类器运用在测试数据集中做测试，统计其准确率与F值，训练集和测试集的文本比例为 7 ： 3 。

维度	维度分的p值	SVM分类准确率	SVM分类F值
第一维度	.000	.762	.755
第二维度	.000	.674	.673
第三维度	.000	.703	.681
第四维度	.000	.676	.680
第五维度	.000	.653	.642
第六维度	.011		

Table 3: 各维度方差分析与SVM分类结果

由表3的统计结果可知，第六维度中美学者论文维度分的p值为.011，大于.05，表明该维度中两类文本的差异不具有显著性，其他维度均表现出显著性差异。从SVM模型分类角度看，剩余的五个维度对中美学术论文的分类准确率和F值尚可接受。基于两种方法的检验证明了有五个维度的维度分具有合理性。每个维度的特征之所以聚类在一起，反映出特征间在某些方面具有共性，以下具体从词汇固有属性、语用功能等方面进行分析，并为每个维度做解释性的命名，以最好的概括维度中的特征。

4.1 第一维度：词汇单位常用性

表4为第一个维度下所有具有显著负荷值的语言特征，均为积极组特征。具体包括基

语言特征	特征解释	负荷值
COCA_news_Range_Log_AW	COCA新闻中词分布对数	.884
COCA_academic_Range_AW	COCA学术中词分布	.849
COCA_news_bi_prop_70k	COCA新闻中70k最高频的二元词串占比	.849
COCA_fiction_Frequency_CW	COCA小说中实词词频	.752
COCA_academic_Frequency_CW	COCA学术中实词词频	.745
COCA_academic_tri_prop_70k	COCA学术中10k最高频的三元词串占比	.745
COCA_magazine_tri_prop_10k	COCA杂志中70k最高频的三元词串占比	.738
Log_Freq_HAL_CW	HAL新闻中实词词频对数	.712
poly_verb	WordNet中动词多义性	.446

Table 4: 第一维度的语言特征及其负荷值

于COCA（Corpus of Contemporary American English）[8]新闻、学术、小说和杂志四个子语料库统计的词频、词分布及其对数、高频二元和三元词串[9]占比，基于HAL（Hyperspace Analogue to Language）语料库(Lund and Burgess, 1996)的实词词频对数。从短语学视角看，独立使用的个体单词和多词单位（multiword units）可操作化定义为词汇单位（lexical units），人们理解、处理和使用多词单位的方式与使用单个词汇项的方式相似(Hoey, 2005; Evert, 2008)，故可将词和词串统称为词汇单位。词汇单位的使用高频且分布广，说明其具有常用性的特征，故将第一维度概括为词汇单位常用性。中美学者论文在该维度的加权得分分别为-1.310、1.310（$p<.05$），准确率为.762，F值为.755，表明中国学者论文中生僻的词汇单位较多，美国学者论文中词汇单位更具有常用性。

以往基于阅读与写作的研究，已证明频率低、分布在较少文本中的词与较难阅读的文本、高质量的作文正相关(McNamara et al., 2015; Kyle et al., 2018)，而高频、广分布的多词单位与高质量作文正相关(Crossley et al., 2012)。本文研究结果有所不同，高频、广分布的词和多词单位均与美国学者的论文更相关，低频、分布较少文本的词和多词单位均与中国学者的论文更相关。

4.2 第二维度：二元词串稳固性

语言特征	特征解释	负荷值
COCA_news_bi_MI^2	COCA新闻中二元词串组合强度（MI^2）	.934
BNC_Spoken_bi_Normed_Freq_Log	BNC口语中二元词串频率对数	.835
COCA_academic_bi_T	COCA学术中二元词串组合强度（T）	.832
COCA_Magazine_bi_Frequency_Log	COCA杂志中二元词串频率对数	.830
COCA_News_bi_Range_Log	COCA新闻中二元词串分布对数	.820
COCA_Academic_bi_Range_Log	COCA学术中二元词串分布对数	.698
COCA_Magazine_bi_Frequency	COCA杂志中二元词串频率	.690
COCA_academic_tri_DP	COCA学术中三元词串组合强度（DP）	.364

Table 5: 第二维度的语言特征及其负荷值

表5为第二个维度下所有具有显著负荷值的语言特征，均为积极组特征。包括基于COCA新闻、学术、杂志各子语料库和BNC(British National Corpus)[10]口语语料库统计的二元词串频率、频率对数、分布对数和组合强度，组合强度通过T分数、互信息平方、Delta P分数来测量。二元词串的使用频率高、分布广且组合能力强，说明该维度反映出二元词串的稳固性，故将第二维度概括为二元词串稳固性。中美学者论文在该维度的加权得分分别为-1.481、1.481（$p<.05$），准确率为.674，F值为.673，表明中国学者论文中使用的二元词串稳固性较弱，美国学者论文中二元词串更为稳固。

[8]https://www.english-corpora.org/coca/
[9]词串指连续的多个单词组成的多词单位
[10]https://www.english-corpora.org/bnc/

相比较单个单词的历时研究，最近的研究开始强调多词单位对于语言发展的重要性。Lu and Deng (2019)证明英语母语者和中国学习者博士学位论文摘要中不同结构和范畴的词串在使用上存在显著的差异性，余龙幸 (2019)证明中外材料学科论文摘要中不同功能和不同长度的词串使用具有差异性。本文的研究与以上观点具有一致性，即词串可作为区分计算语言学中美学者论文的因素。

4.3 第三维度：三元词串稳固性

语言特征	特征解释	负荷值
COCA_news_tri_2_MI	COCA新闻中三元词串组合强度（MI）	.815
COCA_spoken_tri_2_MI	COCA口语中三元词串组合强度（MI）	.807
COCA_news_tri_2_MI^2	COCA新闻中三元词串组合强度（MI^2）	.803
COCA_spoken_bi_MI	COCA口语中二元词串组合强度（MI）	.783
COCA_academic_bi_MI	COCA学术中二元词串组合强度（MI）	.674
COCA_news_tri_2_DP	COCA新闻中三元词串组合强度（DP）	.641
COCA_academic_tri_2_DP	COCA学术中三元词串组合强度（DP）	.640
COCA_magazine_bi_DP	COCA杂志中二元词串组合强度（DP）	.495
COCA_academic_tri_2_AC	COCA学术中三元词串组合强度（AC）	.483
COCA_fiction_bi_DP	COCA小说中二元词串组合强度（DP）	.433
Freq_N_PH	ELP中音韵邻近词词频	-.487

Table 6: 第三维度的语言特征及其负荷值

表6为第三个维度下所有具有显著负荷值的语言特征，多为积极组特征。包括基于COCA各子语料库统计的三元词串组合强度，具体通过互信息、互信息平方、Delta P分数、AC分数（Approximate collexeme strength）来测量。这些特征反映出了三元词串的稳固性。中美学者论文在该维度的加权得分分别为-1.328、1.328（$p<.05$），准确率为.703，F值为.681。表明中国学者论文中三元词串稳固性没有美国学者强，与二元词串的结果一样。Kyle et al. (2018)证明稳固性强的Trigram与L1和L2的高分作文正相关，本文扩展该结论到中美学术论文文本的预测上，即稳固性强的二元词串与美国学者的论文更相关。

4.4 第四维度：虚词的复杂性

语言特征	特征解释	负荷值
LD_Mean_RT_Zscore_FW	ELP中虚词判断时间（Z分数）	.810
McD_CD_FW	McD中虚词共现概率	.779
Freq_N_OG_FW	ELP中虚词表音邻近词词频对数	.730
McD_CD	McD中共现概率	.574
Freq_N_OG	ELP中表音邻近词词频对数	.558
BNC_Written_Freq_FW	BNC写作中虚词词频	-.584

Table 7: 第四维度的语言特征及其负荷值

表7为第四个维度下所有具有显著负荷值的语言特征，多为积极组特征。包括基于ELP（English Lexicon Project）(Balota et al., 2007)统计的虚词判断时间[11]Z分数、表音邻近词（Phonographic neighbors）[12]词频、虚词的表音邻近词词频对数。表音邻近词作为词家族的一类，其大小与特性已被证明有助于解释词语判断、快速命名任务中的差异(Adelman and Brown, 2007; Andrews, 1989)。词语判断实验又是测量词识别特性的主要方法之一。词语判断时间越长，则加工处理的时间越长，说明该词在认知上越复杂。特征还有基于McD（McDonald co-occurrence probability ）(McDonald and Shillcock, 2001)统计的虚词

[11]即参与者判断一个单词是否是英语中真实单词所花费的时间。
[12]即单词间只有一个字母和一个音位不同，例如：“stone”和“stove”。

共现概率与所有词的共现概率。词与他词的共现概率可体现出词语语境的丰富与否(Kyle et al., 2018)。综上，可将第四维度概括为虚词的复杂性，包括认知复杂与语境复杂。

中美学者论文在该维度的加权得分分别为-1.077 、1.077（p<.05），准确率为.676，F值为.680，表明中国学者论文中虚词的使用没有美国学者复杂，即美国学者的论文中虚词的判断时间更长，在认知上更复杂，语境使用上也更加丰富。

4.5 第五维度：词类关联性

语言特征	特征解释	负荷值
PLD_CW	ELP中实词音韵邻近词平均编辑距离	.624
BNC_Written_Freq_FW_Log	BNC书面语中虚词词频对数	.529
WN_Zscore_CW	ELP中实词命名时间（Z分数）	.516
eat_tokens_FW	EAT中虚词自由联想词型符数	.476
USF_AW	USF中词语自由联想词	-.611
poly_adj	WordNet中形容词平均义项数	-.455

Table 8: 第五维度的语言特征及其负荷值

表8为第五个维度下所有具有显著负荷值的语言特征，包括积极组和消极组特征。积极组特征有基于ELP数据库统计的实词音韵邻近词（Phonological neighbors）[13]平均编辑距离、实词命名时间[14]Z分数。平均编辑距离越远代表该词与周围词越不相似，关联性越弱，认知上越复杂(Andrews, 1989; Grainger, 1990)，也会增加实词的命名时间。积极组特征还包括基于BNC书面语子语料库统计的虚词词频对数，基于EAT词典（Edinburgh Associative Thesaurus ）(Kiss et al., 1973)统计的虚词自由联想词的型符数。EAT中的自由联想词指给定一个单词所得到的反映词的数量，其型符多说明联想词多。虚词词频对数越高代表越其简单，也越容易激发所关联的联想词。综上，积极特征组反映出了实词的弱关联性和虚词的强关联性。

消极特征组包括基于USF标准（University of South Florida norms）(Nelson et al., 2004)统计的词语自由联想词数量，基于WordNet统计的形容词平均义项数。形容词义项多说明其使用语境丰富，则与他词的关联性较强，USF中的自由联想词指产生给定单词的刺激词的数量，也代表其关联性较强。因此，消极特征组反映出词语的强关联性，尤其是形容词。

中美学者论文在该维度的加权得分分别为-0.466、0.466（p<.05），准确率为.653，F值为.642，表明中国学者的学术论文中词语整体的关联性较强，尤其是形容词，美国学者则实词关联性强，虚词关联性较弱。

5 讨论

本研究结果显示，中美学者的学术论文在词汇难度方面存在差异性，但并非484个特征对测量二者的差异均起作用，经过筛选，最终得到40个包含于五个维度的有效特征。

原有的484个特征中，基于AWL（Academic Word List）和AFL（Academic Formulas List）提取的学术语言特征，基于TASA（Touchstone Applied Science Associates）语料库和MRC数据库(Coltheart, 1981)等提取的心理语言学词汇特征，和字母的二元组合特征均没有出现，说明中美学者论文在这些特征上的差异不凸显。其中，学术语言特征与学术文本的联系最为密切，英语二语学习者也承认，学术词汇有助于他们在学术写作中选择更好的词汇(Choo et al., 2017)，但许多研究者怀疑学术词汇表在学术写作中的重要性，认为学科专用词汇（也称为专业术语）更有益(Durrant, 2016; Paribakht and Webb, 2016)。本文的结果支持了怀疑学术词汇重要性的观点。

在有效的五个维度中，前三位维度均与词汇单位的频率与分布相关，后两个维度均与词类的认知功能相关。中国学者的五个维度分均低于美国学者，可从内外部因素进行分析。

从外部因素看，受参考资源库的影响。TAALES词汇难度指标的统计值多基于各种外部语料库和数据库得到，例如有COCA语料库、BNC语料库、ELP词典、EAT数据库等，这些

[13]即单词间只有一个音位发音不同，例如“geese”、“lease”和“cease”。
[14]词语的命名时间指参与者开始大声朗读一个单词所花费的时间。

资源库主要来源于英语母语者语言使用的语料或母语者参加词汇实验的数据。因此，美国学者论文中词汇使用会更贴近于这些资源库中的语料情况。毋庸置疑，使用英语本族语者的资源库进行特征统计是合理的。所以，这一发现为中国学者的词汇进步带来了启示，即除了对本国教材和语料的学习，可多关注英语本族语者构建的数据资源，以加强语言表达的地道性。此外，本文已帮助中国学者筛选出了相对有效的资源。前三个维度在测量词汇单位的频率、分布和组合强度时，TAALES中采用了COCA语料库、BNC语料库、SUBTLEXus字母语料库(Brysbaert and New, 2009)、Thorndike–Lorge语料库(Thorndike and Lorge, 1944)等，但最终具有区分性的指标多数来源于COCA语料库，BNC语料库也有部分指标，所以这两个语料库可作为重点学习与参考的材料。后两个维度测量词类的认知功能时，使用ELP词典、McD共现概率表、EAT和USF数据库提取的特征均有涉及，所以这些资源都可加以学习和应用。

与学术语言能力理论视域下采纳主位视角所展开的研究不同，系统功能语言学视域下对文本分析的研究范式无法深入探究出论文创作者的内部因素，只能得出以下推测性的分析，即中国学者的学术英语语体意识较为薄弱。受汉语负迁移的影响，中国学者对学术英语词汇系统认识不够全面，没有完全掌握计算语言学领域学术英语写作的习惯。从第一、二、三维度可知，中国学者对高频、分布广的词汇单位使用不够充分，还需加强对高频、常用词的学习，加强对二元词串和三元词串的组合能力。因此，可重点学习COCA和BNC语料库及各子语料库词频表中的高频、广分布的单词和多词单位。这一差异性也说明在保证一定学术风格的前提下，学术英语写作不需要使用过多生僻的低频词汇来传递作者的研究信息，而需要增强与读者的对话意识，提高语义的适切度，以适当降低阅读的难度。从第四和第五维度可知，在虚词和实词的使用上，中国学者对认知复杂、语境复杂的虚词使用并不充分，对关联性弱的实词使用不够，有待加强。因此，可具体学习ELP中判断时间较长、邻近词词频高的虚词，ELP中命名时间较长、邻近词较少的实词，McD中共现概率较高的虚词。

6 结语

通过本文的研究，解决了提出的两个问题。得出以下结论：1）基于中美学者计算语言学领域的学术英语论文语料能将45个表现较好的词汇难度特征降维成六个维度；其中有五个维度共40个特征对中美学者的论文区分作用显著，分别为：词汇单位常用性、二元词串稳固性、三元词串稳固性、虚词复杂性、词类关联性。2）具体而言，美国学者的论文中使用的词汇单位更具常用性和稳固性，虚词更具复杂性，实词关联性较弱（即词语的词频率高、分布广，二元词串和三元词串的频率高、分布广且组合能力强，虚词的判断时间长、共现概率高且邻近词词频高，实词命名时间长且邻近词少）。中国学者反之。主要原因在于TAALES采用的外部参考库与美国学者的论文更贴近，加之中国学者受汉语负迁移的影响，学术英语语体意识较为薄弱。因此，中国计算语言学学者可充分利用英语本族语者构建的资源库产出更为地道的表达，更好地展开学术英语写作。

本文的创新之处在于：1）构建了计算语言学领域的中美学者学术英语论文全文语料库，可直接应用于计算机使用，以往这样的语料库资源是短缺的；2）针对特定领域的论文进行了超越简单词统计的词汇难度特征的研究，得出的结论更具针对性，区别于以往较为宽泛的研究。本文的贡献在于：1）对Biber的多维分析进行了扩展性应用，有利于丰富其理论体系的建设；2）为中美学者身份识别研究提供了有效的五个维度共40个词汇难度特征；3）为计算语言学领域的中国学者与写作教师鉴别出对学术英语写作有用的词汇资源库。

在将来的研究中，我们将继续探索英语母语者与非母语者学术英语写作中区别性的语言特征，从词的视角扩展到短语、句式、篇章等层面展开全面而深入的研究。随之，将区别性的特征自动化集成于学术论文润色平台中，辅助更多的非母语者写出更为地道与流利的学术论文，助推科研论文的接收与发表，从而有助于提高国际间语言科技成果的交流，促进计算语言学学科的蓬勃发展。

参考文献

James S Adelman and Gordon DA Brown. 2007. Phonographic neighbors, not orthographic neighbors, determine word naming latencies. *Psychonomic Bulletin & Review*, 14(3):455–459.

Sally Andrews. 1989. Frequency and neighborhood effects on lexical access: Activation or search? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(5):802.

Lyle F Bachman, Adrian S Palmer, et al. 1996. *Language testing in practice: Designing and developing useful language tests*, volume 1. Oxford University Press.

David A Balota, Melvin J Yap, Keith A Hutchison, Michael J Cortese, Brett Kessler, Bjorn Loftis, James H Neely, Douglas L Nelson, Greg B Simpson, and Rebecca Treiman. 2007. The english lexicon project. *Behavior research methods*, 39(3):445–459.

S Bax. 2012. Text inspector. *Online text analysis tool*.

Yves Bestgen. 2017. Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence. *System*, 69:65–78.

Douglas Biber. 1988. *Variation across speech and writing*. Cambridge University Press.

Marc Brysbaert and Boris New. 2009. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41(4):977–990.

Yu-Hua Chen and Paul Baker. 2010. Lexical bundles in l1 and l2 academic writing. *Language learning & technology*, 14(2):30–49.

Xiaobin Chen and Detmar Meurers. 2016. Ctap: A web-based tool supporting automatic complexity analysis. In *Proceedings of the workshop on computational linguistics for linguistic complexity (CL4LC)*, pages 113–119.

Lee Bee Choo, Debbita Tan Ai Lin, Manjet Kaur Mehar Singh, and Malini Ganapathy. 2017. The significance of the academic word list among esl tertiary students in a malaysian public university. *3L: Language, Linguistics, Literature®*, 23(4).

Max Coltheart. 1981. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

Scott A Crossley and Danielle S McNamara. 2012. Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, 35(2):115–135.

Scott Crossley, Zhiqiang Cai, and Danielle S McNamara. 2012. Syntagmatic, paradigmatic, and automatic n-gram approaches to assessing essay quality. In *Twenty-Fifth International FLAIRS Conference*.

Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. 2017. Sentiment analysis and social cognition engine (seance): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior research methods*, 49(3):803–821.

Scott A Crossley, Kristopher Kyle, and Ute Römer. 2019. Examining lexical and cohesion differences in discipline-specific writing using multi-dimensional analysis. *Multi-dimensional Analysis: Research Methods and Current Issues*, page 189.

Philip Durrant. 2016. To what extent is the academic vocabulary list relevant to university student writing? *English for Specific Purposes*, 43:49–61.

Jesse Egbert and Shelley Staples. 2019. Doing multi-dimensional analysis in spss, sas, and r. *Multi-dimensional analysis: Research methods and current issues*, pages 125–144.

Stefan Evert. 2008. Corpora and collocations. *Corpus linguistics. An international handbook*, 2:1212–1248.

John Flowerdew. 2009. Goffman's stigma and eal writers: The author responds to casanave. *Journal of English for Academic Purposes*, 8(1):69–72.

Jonathan Grainger. 1990. Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of memory and language*, 29(2):228–244.

Liang Guo, Scott A Crossley, and Danielle S McNamara. 2013. Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, 18(3):218–238.

WANG Haixiao and HILL Clifford. 2011. A paradigm shift for english language teaching in asia: From imposition to accommodation. *Journal of Asia TEFL*, 8(4).

Michael Hoey. 2005. Lexical priming: a new theory of words and language. *London: Routledge*.

Ju Chuan Huang. 2010. Publishing and learning writing for publication in english: Perspectives of nnes phd students in science. *Journal of English for Academic Purposes*, 9(1):33–44.

Ken Hyland. 2009. *Academic discourse: English in a global context*. A&C Black.

Minkyung Kim, Scott A Crossley, and Kristopher Kyle. 2018. Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *The Modern Language Journal*, 102(1):120–141.

George R Kiss, Christine Armstrong, Robert Milroy, and James Piper. 1973. An associative thesaurus of english and its computer analysis. *The computer and literary studies*, pages 153–165.

Kristopher Kyle, Scott Crossley, and Cynthia Berger. 2018. The tool for the automatic analysis of lexical sophistication (taales): version 2.0. *Behavior research methods*, 50(3):1030–1046.

Batia Laufer and Paul Nation. 1995. Vocabulary size and use: Lexical richness in l2 written production. *Applied linguistics*, 16(3):307–322.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S Weld. 2019. S2orc: The semantic scholar open research corpus. *arXiv preprint arXiv:1911.02782*.

Xiaofei Lu and Jinlei Deng. 2019. With the rapid development: A contrastive analysis of lexical bundles in dissertation abstracts by chinese and l1 english doctoral students. *Journal of English for Academic Purposes*, 39:21–36.

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior research methods, instruments, & computers*, 28(2):203–208.

Scott A McDonald and Richard C Shillcock. 2001. Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, 44(3):295–322.

Danielle S McNamara, Scott A Crossley, Rod D Roscoe, Laura K Allen, and Jianmin Dai. 2015. A hierarchical classification approach to automated essay scoring. *Assessing Writing*, 23:35–59.

Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. 2004. The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407.

T Sima Paribakht and Stuart Webb. 2016. The relationship between academic vocabulary coverage and scores on a standardized english proficiency test. *Journal of English for Academic Purposes*, 21:121–132.

Dragomir R Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. 2013. The acl anthology network corpus. *Language Resources and Evaluation*, 47(4):919–944.

Edward Lee Thorndike and Irving Lorge. 1944. The teacher's word book of 30,000 words. *Bureau of Publications, Teachers Co*.

Sedef Uzuner. 2008. Multilingual scholars' participation in core/global academic communities: A literature review. *Journal of English for Academic Purposes*, 7(4):250–263.

Alistair Wood, J Flowerdew, and M Peacock. 2001. International scientific english: The language of research scientists around the world. *Research perspectives on English for academic purposes*, 71:83.

Fang Xu. 2015. A review of academic english writing research. *Foreign language teaching and research*, 1:94–105.

Yu Li Liang Yonggang. 2006. A study of the writing model of english scientific papers. *Foreign Language Education*, 1.

Su-Youn Yoon, Suma Bhat, and Klaus Zechner. 2012. Vocabulary profile as a measure of vocabulary sophistication. In *Proceedings of the seventh workshop on building educational applications using NLP*, pages 180–189.

Jun Zhao. 2017. Native speaker advantage in academic writing? conjunctive realizations in eap writing by four groups of writers. *Ampersand*, 4:47–57.

第二十届中国计算语言学大会论文集，第454页-第466页，呼和浩特，中国，2021年8月13日至15日。

Donald W Zimmerman. 2004. A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, 57(1):173–181.

余龙幸. 2019. 语料库驱动的中外材料学科论文摘要词块的对比研究. Master's thesis, 东南大学.

吴明隆. 2010. 问卷统计分析实务:*SPSS*操作与应用. 问卷统计分析实务:SPSS操作与应用.

姜亚军and 赵刚. 2006. 学术语篇的语言学研究: 流派分野和方法整合. 外语研究, 6:1–5.

孟凡玲. 2018. 中美硕士论文致谢的对比研究. Ph.D. thesis, 北京理工大学外国语学院.

比伯, 康拉德, 瑞潘, 刘颖, and 胡海涛. 2012. 语料库语言学. 语料库语言学.

王彤. 2019. 基于语料库的中外学者经济类期刊论文中词块的比较研究. Ph.D. thesis, 北京第二外国语学院英语学院.

王琴. 2020. 中外学者英语学术论文讨论部分语步对比分析. Ph.D. thesis, 长春理工大学.

赵怡琳. 2018. 中国英语专业硕士与英语本族语专家在英语学术写作中介入标记语使用情况的对比研究. Ph.D. thesis, 西安外国语大学.

雷蕾. 2016. 应用语言学研究设计与统计. 应用语言学研究设计与统计.