

# 基于语料库的科技英语摘要短语 结构与功能分布研究

陈春丽, 张颜颜, 许炳梁

昆明理工大学外国语言文化学院, 云南 昆明

收稿日期: 2024年8月13日; 录用日期: 2024年9月20日; 发布日期: 2024年9月30日

## 摘要

近年来, 基于语料库的具有学科特色的学术短语列表构建备受关注。该研究对比了COCA学术语料库与自建的科技英语摘要语料库中的学术短语。目的是分析国内科技英语摘要中的学术短语与国际学术语料库在结构和功能上的差异, 进而探寻国内英语学习者在学术短语运用方面存在的问题。研究结果表明: 1) 在结构分类方面, 动词类短语在国内科技英语摘要中的使用频率比COCA学术语料库的高, 而功能词类短语的使用频率则呈现相反趋势; 2) 在功能分类方面, 两个语料库均以指示功能类短语为主。这些发现不仅有助于深入理解国内科技英语摘要的语言特点, 也为未来的学术写作与教学提供了有益参考。

## 关键词

语料库, 科技英语摘要, 学术短语列表, 结构与功能分类

# A Corpus-Based Study on the Distributions of EST Abstract Phrases Categorized by Structure and Function

Chunli Chen, Yanyan Zhang, Bingliang Xu

Faculty of Foreign Languages and Cultures, Kunming University of Science and Technology, Kunming Yunnan

Received: Aug. 13<sup>th</sup>, 2024; accepted: Sep. 20<sup>th</sup>, 2024; published: Sep. 30<sup>th</sup>, 2024

## Abstract

There has been a growing interest in compiling corpus-based academic formulas lists with disciplinary

characteristics in recent years. A contrastive analysis of the academic phrases, which are sourced from the academic corpus of COCA and the abstract corpus of EST papers published in China, is conducted in this paper. The distributions of academic phrases categorized by structure and function are studied to recognize the challenges faced by Chinese EFL learners. The analysis indicates that: 1) as for structural classification, verbal phrases are used more frequently in the EST abstract corpus than in the COCA academic corpus, while the frequency of function-based phrases represents the reverse tendency; 2) concerning functional classification, the application of referential expressions phrases are significant in both corpora. These findings serve as a potential reference to the study on linguistic features of EST abstracts and the teaching of academic writing in future.

## Keywords

Corpus, EST Abstracts, Academic Formulas List, Structural and Functional Classification

Copyright © 2024 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

## 1. 引言

学术短语(academic phrase), 作为学术文体中特有的高频词组, 其结构介于词汇与子句之间, 固化程度则超越自由组合, 但不一定是惯用语[1]。在语言学领域, 学术短语是程式语的一种类型。所谓程式语, 是指由连续或非连续的语言单位构成的序列[2]。

近二十年来, 基于语料库的研究在学术英语(English for Academic Purpose, 简称 EAP)领域取得了显著进展, 为学习者和教育工作者构建学术短语列表方面尤甚。这些研究不仅探讨了构建此类列表的方法论问题, 还提出了多种适用于教学的学术短语列表类型[3]-[9]。但是这些学术短语研究在特定学科、语域和语境情境下的应用尚未得到充分挖掘[10] [11]。

在国际上, Simpson-Vlach 和 Ellis 从功能角度对学术短语进行分类, 并提出了一种新的研究方法, 即通过计算每个程式语的教学价值(Formula Teaching Worth, 简称 FTW)分数来衡量短语的教学价值[5]。他们通过多元回归分析为所有程式语生成 FTW 得分, 且最终的学术短语列表(Academic Formulas List, 简称 AFL)理论上只包含适用于教学的程式语。Cunningham 从结构和功能上对学术短语进行分类, 并从数学期刊论文引言中提取了 180 个学术短语, 这有助于理解和提升数学学术写作[12]。Lu 在此基础上构建了社会科学期刊论文引言的学术短语列表, 以指导学术写作[11]。然而, 这些研究主要集中于学术论文的引言部分, 对其他部分如摘要、方法、结果和讨论等的关注相对较少。

相较于国际研究, 国内基于语料库的学术英语研究主要聚焦于揭示英语学习者与本族语者在学术语块使用上的差异。此类研究多采用语料库对比法, 将英语学习者语料库设为研究对象, 并以本族语者语料库为参照基准, 探讨学术语块的习得[13]、使用特征[14]、学科间使用差异[15]、语用认知及话语立场[16] [17]、功能分类等多个方面的问题[18]。然而, 在构建学术词块或短语列表的研究领域, 成果相对匮乏[19], 仅有少数如针对农科类学术短语的列表构建被提及[20], 而针对科技英语领域高引用论文的学术短语词表构建则鲜有探索。尽管有研究者如杨成等人[21]和潘潘[16]深入探讨了中外高水平期刊 ESP 论文英文摘要中四词词块的结构和功能, 揭示了中外期刊在词块使用上的显著差异, 但这些研究尚未与美国当代英语语料库(Corpus of Contemporary American English, 简称 COCA)进行对比。鉴于此, 本研究通过与 COCA 这一广泛认可的大规模英语语料库进行对比, 以期提供更加全面且具有代表性的英语学术论文

摘要短语使用情况。除此之外,现有研究多集中于连续性学术短语中的三元序列和四元序列分析[14][16][17][22],同时对二元序列、三元序列、四元序列和五元序列的探讨则极为罕见。关于学术短语结构与功能分类标准,学术界亦存在多样性,如徐昉[14]采用了 Cortes [23]的学术短语结构与功能分类标准,潘璠[16]采用了 Biber 等[24]的结构框架和 Hyland [4]的功能框架。然而,同时参照如 Biber [25]的结构分类标准和 Simpson-Vlach & Ellis [5]功能分类标准的研究尚显不足。

综上,近二十年来,基于语料库的研究在构建学术短语列表和探讨其教学价值方面取得了显著成果。国际上,学者们不仅提出了多种学术短语分类和构建方法,还开发了用于衡量短语教学价值的 FTW 评分系统。然而,这些研究主要集中在学术论文的引言部分,对其他部分的关注较少,且特定学科、语域和语境情境下的应用尚未得到充分挖掘。相比之下,国内研究主要聚焦于英语学习者与本族语者在学术语块使用上的差异,但关于常用学术短语列表构建的研究相对较少。此外,尽管一些研究探讨了学术短语的结构和功能,但分类标准存在差异,且对连续性学术短语中不同序列长度的研究并不均衡。

鉴于此,本研究通过与广泛认可的大规模英语语料库(如 COCA)进行对比,旨在提供更加全面且具有代表性的英语学术论文摘要短语使用情况,以期揭示出学术短语在不同语境和学科中的使用差异,为学术英语教学提供更加具体的指导。本研究还采用统一的学术短语结构与功能分类标准,以确保研究结果的准确性和可比性。通过整合不同学者的分类标准,本研究期望提出一套更加全面和实用的分类体系,为学术短语的研究和教学提供有力支持。

## 2. 研究设计

### 2.1. 研究问题

根据上述文献综述,本文旨在研究并回答以下问题:在科技英语摘要语域,国内英语学习者的学术短语习得与本族语者在结构类型及分布上有何差异?在科技英语摘要语域,国内英语学习者的学术短语习得与本族语者在功能范畴及分布上有何差异?国内大学 EAP 教师应该将哪类学术短语作为授课重点?

### 2.2. 语料库构建

本研究共采用了两组语料库:自建的科技英语摘要语料库和作为参考的 COCA 语料库。本研究从国内三所理工类院校学报(自然科学版)中收集了矿冶工程领域 2014-2021 年发表的高被引论文。经过数据清洗和人工筛查,共纳入 556 篇论文的摘要部分,形成科技英语摘要语料库,库容为 124,147 词。在创建该语料库前,本研究团队根据期刊的影响因子进行了筛选,并由两位学科专家确认了论文的代表性。采用目标抽样技术,从每种选定期刊的 2014 至 2021 年间抽取了 6 期样本,每期选取一篇研究文章的摘要,并以 TXT 格式保存。为保证数据的准确性,所有文件均经过了手动校对,纠正了数字转录错误和拼写错误,并去除了与学术短语无关的信息,例如“Abstract:”标识及出版年份等。

“在实际研究中,视研究的需要,常常依据某种条件从一个大的语料库(如通用语料库)中抽取若干个符合条件的文本构建一个子语料库,以达到对比的目的”[26]。因此,本研究还使用了于 2020 年 3 月更新的最新版的美国当代英语语料库,选取了 COCA 语料样本(Sample)库中的学术期刊子语料库,经过数据清洗和人工筛查,建成 COCA 学术语料库,库容为 1,351,064 词。

### 2.3. 研究方法

根据自建语料库的库容,并参照前人的研究[11][25][27],本研究聚焦于连续性的学术短语,从自建

语料库中提取二元序列、三元序列、四元序列和五元序列。学术短语列表的构建、审查和比较过程分为四个阶段：短语结构提取、短语结构筛选、结构和功能审查与短语结构比较。

1) 短语结构提取。本研究采用了基于连续词序列直接提取的方法，并运用本研究团队开发的程序进行学术短语的提取。程序包括语料库数据清洗、语料库基本特征统计和学术短语提取。程序在 Python 3.8.6 环境下编写，使用了 NLTK 自然语言处理模块，借鉴了蕾蕾基础程序的编写逻辑[28]。将清洗过的语料提取短语后，得到依据频次降序排列的多词序列短语。

2) 短语结构筛选。鉴于所提取的学术短语数量庞大，且低频短语在教学应用中价值有限，本研究根据学术短语所出现的频次，选取了每个语料库中多元序列中的前 200 个短语进行进一步研究，共计 1600 个短语。同时，考虑到相互信息值(MI score)对短语结构的意义和教学价值的影响大于频次[5] [14]，本研究计算了这 1600 个短语的 MI 值，并按照 MI 值降序排列，最终选取每个语料库中多元序列的前 100 个短语进行分类研究，共选取了 800 个短语。

本研究参照了卫乃兴[29]的 MI 值计算方法，公式如下：

$$MI(a, b) = \log_2 \frac{P(a, b)}{P(a)P(b)} \quad (1)$$

在语料库分析中，词形对 a 和 b 的共现概率表示为  $P(a, b)$ ，而  $P(a)$  与  $P(b)$  分别代表词形 a 和词形 b 独立出现的概率。当两个词形之间的结合力较强时，即意味着它们经常一起出现，此时  $P(a, b)$  显著高于  $P(a)$  和  $P(b)$  的乘积，即  $P(a) \cdot P(b)$ 。这种情形下，搭配强度指标  $MI(a, b)$  将趋于正值，表明词形 a 与词形 b 之间存在强烈的搭配倾向。反之， $MI(a, b)$  趋于负值，词形 a 与 b 搭配强度越低。假设语料库的总词容为  $Q$ ，其中词形 a 的观测频次表示为  $F(a)$ ，词形 b 的观测频次表示为  $F(b)$ ，而  $F(a, b)$  则指代词形 a 与词形 b 在语料库中共同出现的频次，则

$$P(a) = \frac{F(a)}{Q} \quad (2)$$

且

$$P(b) = \frac{F(b)}{Q} \quad (3)$$

那么，两个词形共现的概率为

$$P(a, b) = \frac{F(a, b)}{Q} \quad (4)$$

将(式 4)带入(式 1)，得

$$MI(a, b) = \log_2 \frac{Q \cdot F(a, b)}{F(a) \cdot F(b)} \quad (5)$$

3. 结构分类和功能分类。本研究对所分析的 800 个学术短语进行了详细的结构与功能分类。在结构分类方面，参照 Gray & Biber [25] 的分类标准，将含有一个或多个动词、助动词或情态动词的短语归类为动词结构(例如“was used to”)。包含一个或多个名词、形容词或副词的短语归类为非动词类实词结构(例如“of this review”)。仅由功能词(如介词、连词、代词、限定词等)构成的短语则被归类为功能词结构(例如“on the other hand”)。

在功能分类上，本研究参照 Simpson-Vlach & Ellis [5] 的分类标准，将学术短语的功能划分为指示、

立场、语篇三大类。与此同时，在每一主要类别下进一步进行了次级分类，将指示功能细分为六个子类别：属性说明、识别与聚焦、对比与比较、指示与定位、模糊标记及其他。立场功能被划分为模糊限制语、认知立场、强制性与指示、能力与可能性、评价、意图与预测这六个子类别。而语篇功能则分为元语篇与语篇指示、主题引入与聚焦、主题阐述、语篇标记这四个子类别。

4. 短语结构对比。本研究采用了对比分析方法，这是语料库语言学中常用的研究方法和分析视角[26]。所选取的基准语料库为美国杨伯翰大学语言学教授 Mark Davies 于 2008 年创立的美国当代英语语料库 (COCA)。本研究系统对比自建科技英语摘要库与 COCA 学术语料库，分析短语结构和功能频率的分布情况。通过深入分析，本研究旨在识别并解析英语学习者在短语使用上的差异和挑战，并提出具体的改进建议。

### 3. 结果与讨论

#### 3.1. 语料基本特征分析

本研究对两个语料库进行了基本指标分析，包括形符数(Token)、类符数(Type)、类/形符比(Type-Token Ratio, 即 TTR)、平均词长以及平均句长，以构建对语料库特性的基础了解[30]。分析结果详见表 1 (如下表)。表 1 数据显示，两个语料库的平均词长基本一致，平均句长相差较小，但类/形符比差异显著，主要是两个库容差异悬殊造成的。在研究词汇复杂性时，需确保两个语料库的库容基本一致，这会对研究结果产生显著影响[30]。然而，对于学术短语的研究，体裁一致比库容更为关键，即两个语料库均应为学术文章。许家金[31]指出，语料库的代表性比大小更为重要，即可在统计上代表真实话语的类型。

Table 1. Statistical features of the two corpora

表 1. 语料库基本特征统计

语料库	形符数	类符数	类/形符比	平均词长	平均句长
科技英语摘要语料库	124,147	9136	7.36%	4.95	26.82
COCA 学术语料库	1,351,064	52,337	3.87%	4.58	23.79

#### 3.2. 结构分类结果

参照 Gray & Biber [25] 的分类标准，表 2 展示了 COCA 学术语料库中学术短语的结构分类示例，表 3 统计了两个语料库中学术短语的结构分类结果。观察表 3 发现，在宏观层面，两个语料库都是非动词类实词结构占比最大(57%, n=228)，动词结构次之，功能词结构占比最少。这一发现与 Biber [25] 的研究结论相一致，但已显示出差异性。Biber [25] 指出，在学术写作中，功能词结构和动词结构的使用较为频繁，而非动词类实词结构的使用频率较低；相比之下，口语交流则更多依赖动词结构。在微观层面，科技英语摘要库中动词结构使用频率比 COCA 学术语料库高出 4%，这一差异可能与英语和汉语的语言特性有关。连淑能[32]提出，英语倾向于使用名词，呈现静态叙述，而汉语多用动词，呈现动态。因此，英语学习者可能在学术英语写作中受到母语的负迁移影响，导致动词结构使用频率较高，使得学术写作产生口语化倾向。

表 4 和表 5 的统计结果显示，从二元序列至五元序列，非动词类实词结构的使用频率逐渐上升，而功能词结构的使用频率则逐渐下降。这表明随着短语长度的增加，非动词类实词的使用更为频繁，而功能词的使用则相对减少。在动词结构方面，COCA 学术语料库的动词结构随着短语长度的增加，其使用频率有所下降。相比之下，科技英语库摘要语料库在二元、三元与四元序列中动词结构的使用较为均衡，

**Table 2.** Examples of structural classification phrase in the COCA academic corpus**表 2.** COCA 学术语料库学术短语的结构分类示例

短语长度	动词结构	非动词类实词结构	功能词结构
二元序列	have been, should be	levels of, nature of	duo to, those who
三元序列	there is no, they do not	in physical education, the present study	as well as, with respect to
四元序列	was found to be	the purpose of this	at the same time
五元序列	are less likely to report	of the rule of recognition	as a matter of fact

**Table 3.** Structural classification proportions of academic phrases in the two corpora**表 3.** 两个语料库学术短语的结构分类占比

语料库	动词结构	非动词类实词结构	功能词结构	总计
科技英语摘要语料库	134 (33.5%)	228 (57%)	38 (9.5%)	400
COCA 学术语料库	118 (29.5%)	228 (57%)	54 (13.5%)	400

**Table 4.** Structural classification distribution of academic phrases in the COCA academic corpus**表 4.** COCA 学术语料库学术短语的结构分类分布

结构分类	二元序列	三元序列	四元序列	五元序列	总计
动词结构	48	31	27	12	118 (29.5%)
非动词类实词结构	30	54	60	84	228 (57%)
功能词结构	22	15	13	4	54 (13.5%)
总计	100	100	100	100	400

**Table 5.** Structural classification distribution of academic phrases in the EST abstract corpus**表 5.** 科技英语摘要语料库学术短语的结构分类分布

结构分类	二元序列	三元序列	四元序列	五元序列	总计
动词结构	37	37	38	22	134 (33.5%)
非动词类实词结构	46	50	55	77	228 (57%)
功能词结构	17	13	7	1	38 (9.5%)
总计	100	100	100	100	400

但在五元序列中使用频率降低。在三大类结构中,本研究结果与陈艳君[20]的研究结论相一致,即在学术写作中,动词结构与非动词类实词结构都占据了主导地位,而功能词结构的使用相对较少。这种趋势可能与词序长度有关,即词序越长,动词及其他实词的使用概率越高。与之不同的是,本研究通过与 COCA 学术语料库的对比分析,发现随着短语长度的增加,英语学习者在功能词方面存在使用不足的问题,在动词方面存在过度使用的倾向,这可能与英语学习者的语言习得背景和写作习惯有关。

### 3.3. 功能分类结果

依据 Simpson-Vlach & Ellis [5]提出的功能分类框架,本研究对筛选后的短语进行了一级与二级功能分类,并提供了具体的分类实例(详见表 6)。基于归类结果,本研究进一步分析了 COCA 学术库与科技英

语摘要库中学术短语一级功能的分布情况(详见表 7)。

**Table 6.** Examples of functional classification of the two corpora  
**表 6.** 两个语料库学术短语的功能分类示例

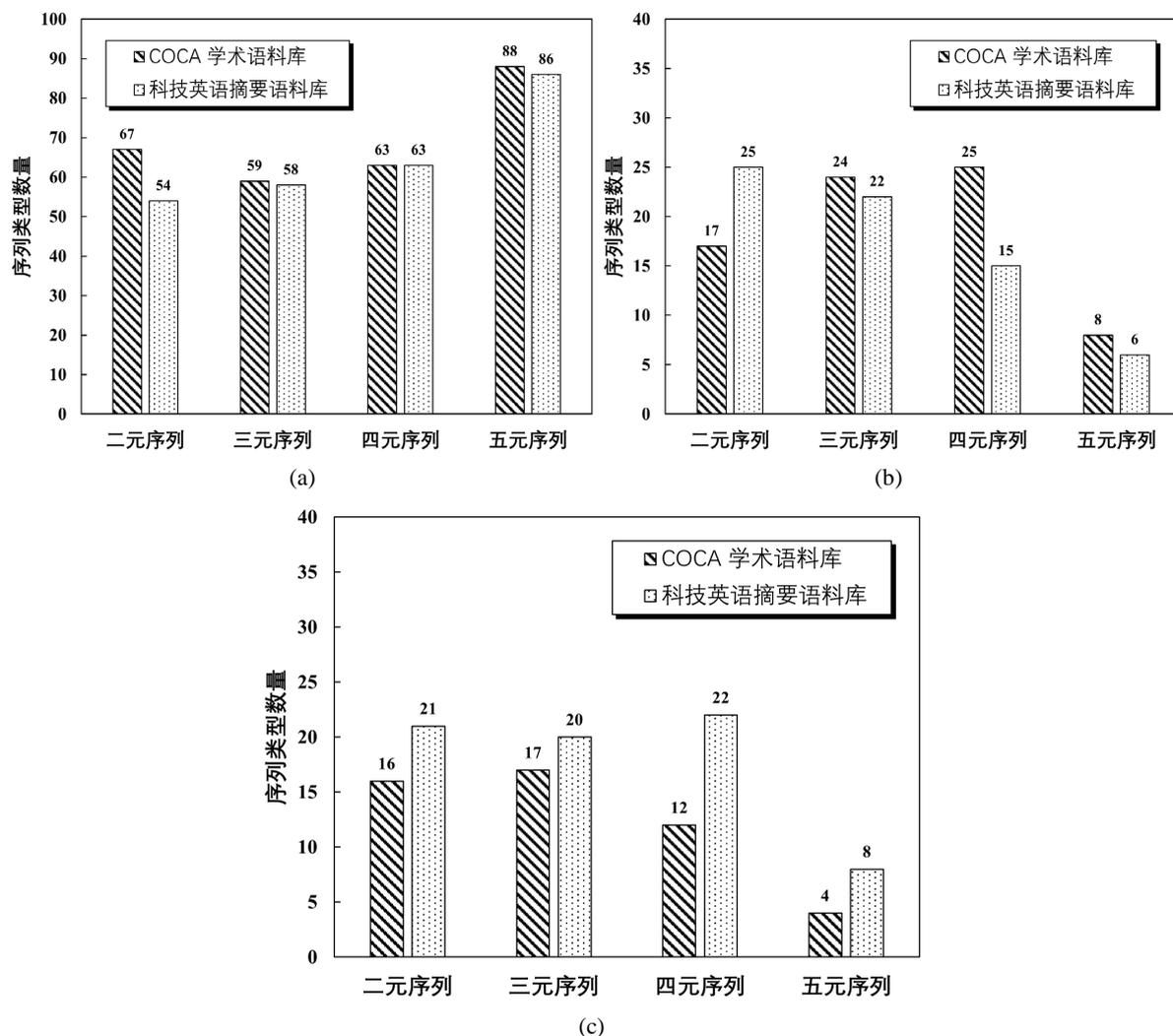
一级分类	二级分类	科技英语摘要语料库	COCA 学术语料库
指示功能	属性说明	on this basis, a new method for, is an effective method for	the presence of, on the basis, the amount of, of two kinds of teachers
	识别与聚焦	such as, there are, has been, to be, is one of the most	there will be, such as, it is, that is not, this is not the case
	对比与比较	less than, compared with, on the other hand, the relationship between	more than, on the one hand, on the other hand, the relationship between
	指示与定位	in China, iron and steel industry,	at the university, the state of Florida
	模糊标记	无	<i>et al.</i>
立场功能	模糊限制语	the main, is the main, are the main	should be, may not be, is likely to be
	认知立场	according to	according to, was found to be, have been shown to
	强制性与指示	无	must be, should not be, to ensure that
	能力与可能性	can be, (was/were) carried out, be used as, was obtained by	can be, be able to, it is likely that, it is possible
	评价	is beneficial to, which is consistent with, has little effect on	it is true that, it is important, (of) the positive turn
	意图与预测	to study, in order to explore, to solve the problem(s)	to make, to participate in, in order to understand, I would like to
语篇功能	元语篇与语篇指示	(in) this paper, is found to be, (in) this study, it is shown that	of this study was, are presented in table, found to be
	主题引入与聚焦	in recent years	for example, the first, the second
	主题阐述	due to, experimental results, are as follows, the results show	due to, as a result, the fact that, in order to, the effects of
	语篇标记	in addition, as well as, there is no obvious, at the same time	rather than, in other words, at the same time

**Table 7.** Functional classification proportions of academic phrases in the two corpora  
**表 7.** 两个语料库学术短语的功能分类占比

语料库	指示功能	立场功能	语篇功能	总计
科技英语摘要语料库	261 (65.3%)	68 (17%)	71 (17.8%)	400
COCA 学术语料库	277 (69.3%)	74 (18.5%)	49 (12.3%)	400

观察表 7, 我们发现科技英语摘要库中指示功能和立场功能方面的学术短语占比略低于 COCA 学术库; 相反, 科技英语摘要库中语篇功能的学术短语所占比例略高于 COCA 学术库。图 1 分析的结果显示, 在指示功能方面, 两个语料库的五元序列最为常见, 且随着学术短语长度的增加, 其指示功能趋于显著(详见图 1(a))。在立场功能方面, 两个语料库都是五元序列出现频率最低, 且科技英语摘要库从二元序列至五元序列呈现递减趋势(详见图 1(b))。在语篇功能方面, 两个语料库均显示出递减趋势(详见图 1(c))。各大分类下的子类分布情况将如何呈现? 英语学习者与本族语者的学术短语应用在功能分类上有何具体

差异? 接下来本文将基于功能分类的标准, 结合具体的二元、三元、四元及五元学术短语实例, 对上述问题进行深入讨论。



**Figure 1.** Comparison of the functional distribution of academic phrases in the two corpora ((a) referential expressions; (b) stance expressions; (c) discourse organizing expressions)

**图 1.** 两个语料库中按功能分类的短语数量对比((a) 指示功能; (b) 立场功能; (c) 语篇功能)

1. 指示功能类。根据 Biber *et al.* [3]的定义, 指示功能类短语用于指向具体或抽象实体, 或者指明上下文中的内容, 旨在明确它们的身份或强调其特定的重要属性。如表 7 所示, 两个语料库中指示功能类短语的占比接近三分之二。本研究采用 Simpson-Vlach & Ellis [5]提出的二次分类框架, 进一步细分了学术短语的指示功能为六个子类别: 属性说明、识别与聚焦、对比与比较、指示与定位、模糊标记与其他。

观察表 8 发现, COCA 学术库中属性说明类短语和指示与定位类短语所占比重相同(22%,  $n = 61$ ); 科技英语摘要库中属性说明类短语所占比重最大(42.9%,  $n = 112$ ), 约是 COCA 学术库的两倍, 指示与定位类短语次之(10.7%,  $n = 28$ )。在 Simpson-Vlach & Ellis [5]的分析中, 属性说明类的指示短语亦是占比最大。

**Table 8.** Statistics for referential expressions phrases in the two corpora  
**表 8.** 两个语料库中指示功能短语分类统计

指示功能分类	科技英语摘要库	COCA 学术语料库
属性说明	112 (42.9%)	61 (22%)
识别与聚焦	18 (6.9%)	52 (18.8%)
对比与比较	20 (7.7%)	16 (5.8%)
指示与定位	28 (10.7%)	61 (22%)
模糊标记	0 (0%)	1 (0.1%)
其他	83 (31.8%)	86 (31%)
总计	261	277

此外, 识别与聚焦类短语, (如 *such as the, there are, it is not, I do not know, this is not to say*) 包括典型的说明性短语以及带有系动词、助词、情态动词结构的句子或从句的词干, 用以举例或指示对象[5], 本研究中 COCA 学术库的识别与聚焦类短语占比约为科技英语摘要库的三倍, 以下是分别从科技英语摘要库和 COCA 学术库中选取的文本 1 和文本 2, 可以更加直观的对比英语学习者与本族语者识别与聚焦类短语(下划线加粗标出)的使用差异。

文本 1 (科技英语摘要库):

To solve the non-uniform gas-particle mixing problems **such as** the falling raw concentrate remain, shorter reactor shaft lifetime and less oxygen utilization at the high loading flash smelting process with a CJD burner, a mathematical model of a CJD burner with distribution air was built up... The results show that the process air is concentrated inside the circular cone under the nozzle, the particle distribution is effectively controlled by the momentum ratio...

文本 2 (COCA 学术库):

Putting a senior lawyer with technology experience in charge of a technology project doesn't guarantee success to mediate disputes... **It is** the legal staff facing off against the IT staff. Unfortunately, these two sides tend to speak different languages and perceive work issues from different points of view. And **that's not** even mentioning the systems integrators and the software vendors often have different views on how best to implement a system... At the same time, **there will be** people who will resist using a new system for understand...

根据 Simpson-Vlach & Ellis [5] 19~20 的定义, 属性说明类短语由三部分组成: 抽象概念类短语、可测量的名词短语及前置指代和后置指代类短语。在 20 世纪 80 年代, 杨惠中[33]对科技英语术语进行了研究, 并提出了区分科技术语的三个标准, 即语域(register)、文类(genre)和主题(topic), 以及信息流或概念结构。这些标准强调了术语在不同语域、文类和主题中的局部性和特异性。图 2(a)的结果表明, 两个语料库中属性说明类短语的趋势线分布基本一致, 数量上的差异可能是两个语料库的主题差异导致。COCA 学术库包含各个领域的学术期刊, 而科技英语摘要库则专门收录矿冶领域学术期刊的摘要。

此外, 本研究在科技英语摘要库中并未发现模糊标记类短语, 这可能是因为模糊标记类短语与摘要文本的简洁性和精准性的要求相悖, 导致此类功能短语的缺位。相比之下, COCA 学术库由于全文收录而非只收录摘要, 因此提取了一个模糊标记类短语。Simpson-Vlach & Ellis [5]也只在学术书面语篇中发现了一个此类学术短语 and so on。

2. 立场功能类。立场功能类的短语用于解释后续命题, 传达认知状态, 并对行为或事件展示说话者

的态度[3]。依据 Simpson-Vlach & Ellis [5]的二次分类框架,本研究将两个语料库中的立场功能类学术短语细分为六个子类:模糊限制语、认知立场、强制性与指示、能力与可能性、评价、意图与预测(见表9)。

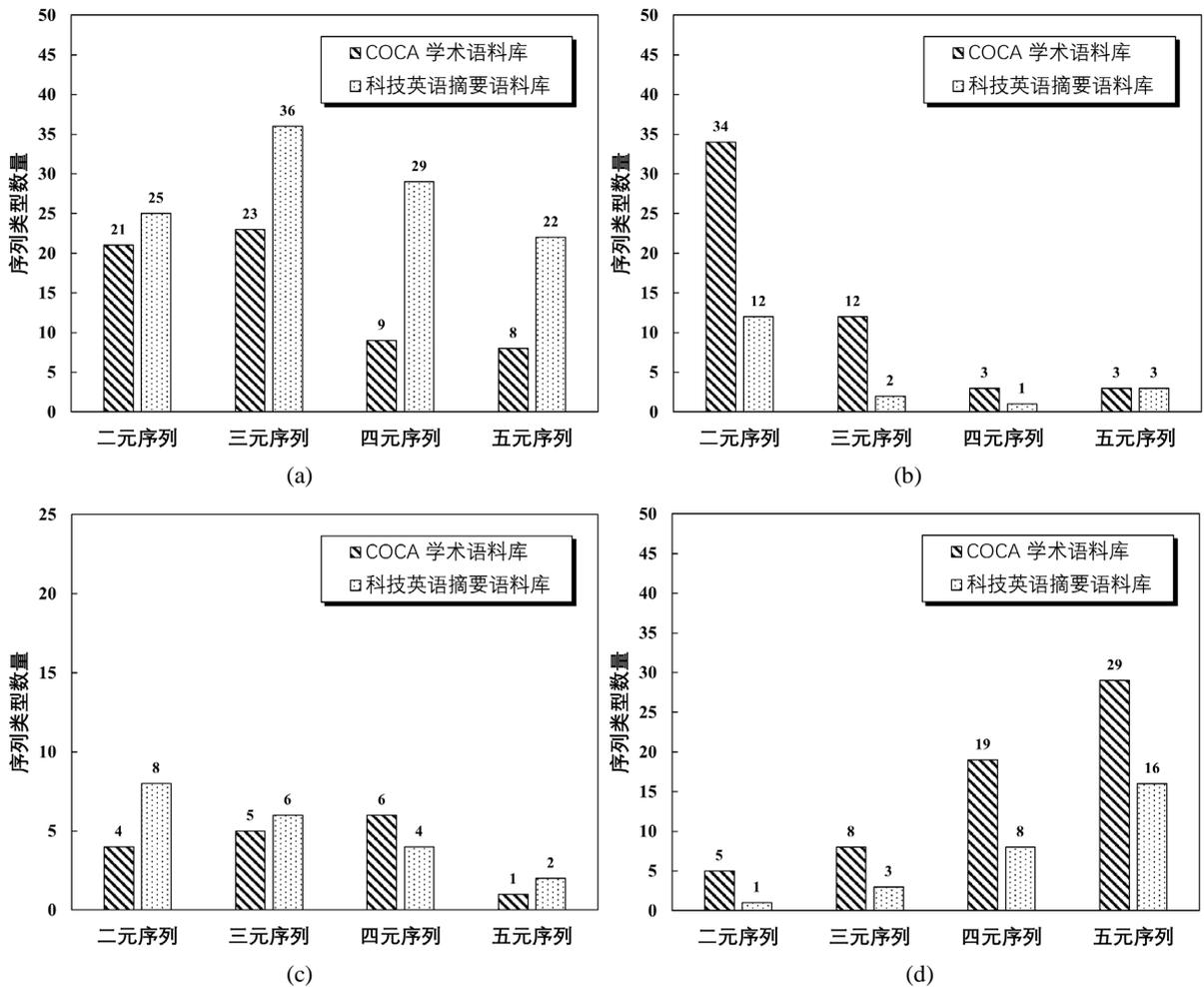


Figure 2. Comparison of the number of referential expressions phrases in the two corpora ((a) specification of attributes; (b) identification and focus; (c) contrast and comparison; (d) deictics and locatives)

图2. 指示功能短语数量对比图((a) 属性说明; (b) 识别与聚焦; (c) 对比与比较; (d) 指示与定位)

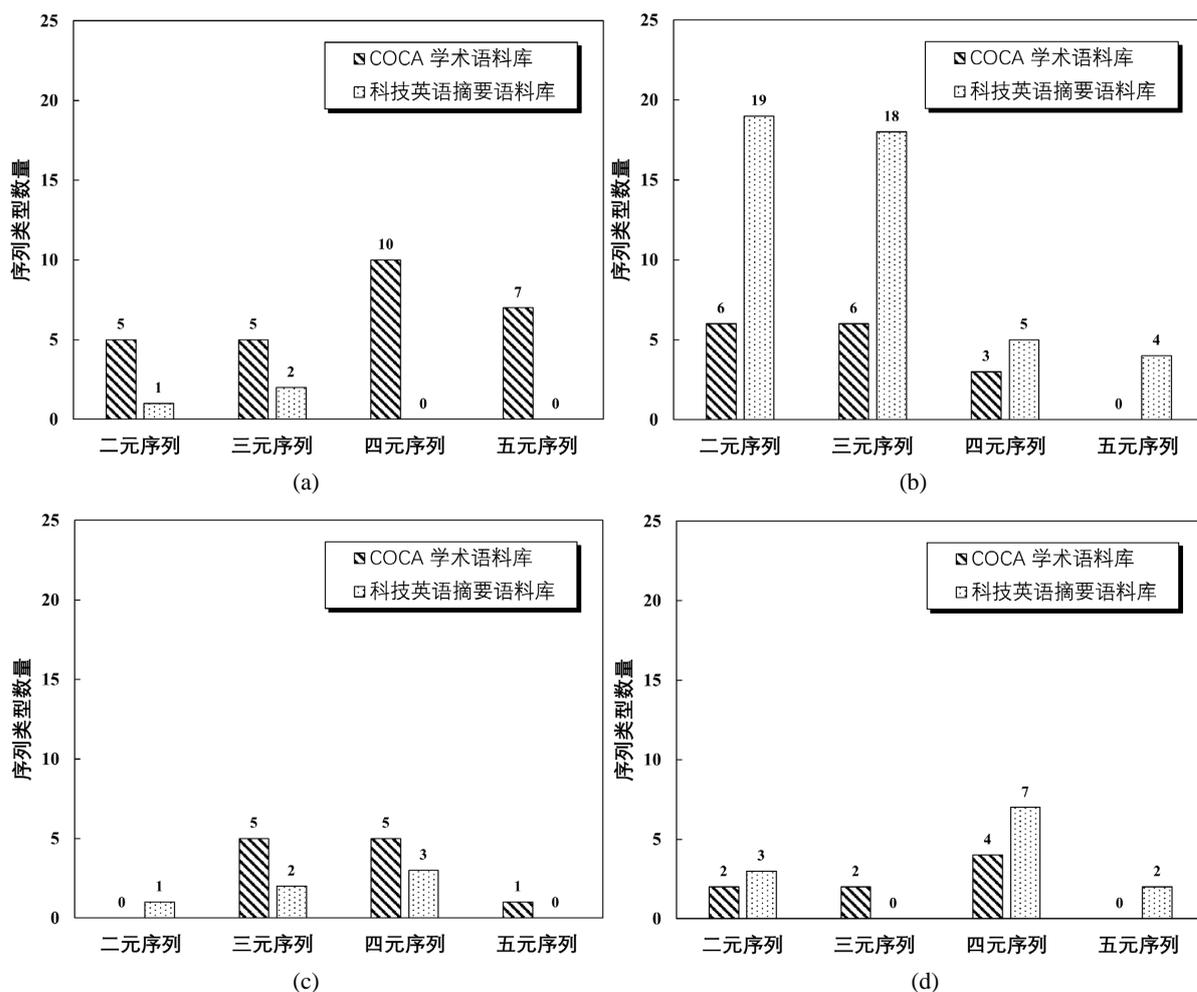
Table 9. Statistics for stance expressions phrases in the two corpora

表9. 两个语料库中立场功能短语分类统计

立场功能分类	科技英语摘要库	COCA 学术语料库
模糊限制语	3 (4.4%)	27 (36.5%)
认知立场	1 (1.5%)	4 (5.4%)
强制性与指示	0 (0%)	9 (12.2%)
能力与可能性	46 (67.6%)	15 (20.3%)
评价	6 (8.8%)	11 (14.9%)
意图与预测	12 (17.6%)	8 (10.8%)
总计	68	74

统计结果显示, COCA 学术库中模糊限制语占比最大(36.5%,  $n = 27$ ), 能力与可能性类短语次之(20.3%,  $n = 15$ )。科技英语摘要库中能力与可能性类短语占比最大(67.6%,  $n = 46$ ), 超过 COCA 学术库中模糊限制语和能力与可能性类短语占比之和, 其次是意图与预测类短语(17.6%,  $n = 12$ )。

在学术写作中, 模糊限制语是作者表达判断、推进主张的重要手段, 它介入文本并与读者展开对话, 增强了作者的交互作用, 使得观点得以表述, 同时也可预见潜在的反对意见[34]。图 3(a)的分析显示, COCA 学术库中的模糊限制语的使用频率是科技英语摘要库的十倍, 这表明国内科技英语学习者在论述时与读者互动不足, 表述观点和态度时很少采用间接和缓和的语气, 需要加强此类短语的习得与应用。这与杨成[21]的研究结论相符, 学习者缺少人际互动。与之不同的是, 杨成的结论是基于学习者多用被动表达和固定句型得出的, 而本研究是基于学习者极少使用模糊限制语得出的。



**Figure 3.** Comparison of the number of stance expressions phrases in the two corpora ((a) hedges; (b) ability and possibility; (c) evaluation; (d) intention)

**图 3.** 立场功能短语数量对比图((a) 模糊限制语; (b) 能力与可能性; (c) 评价; (d) 意图与预测)

能力与可能性短语则引出了潜在的或实际的行为或提议[5]。图 3(b)的对比结果表明, 科技英语摘要库的能力与可能性类短语是 COCA 学术库的三倍, 这可能与国内科技英语学习者重视描述试验结果和价值有关, 而国际上则更倾向于描述试验过程。其他同类研究并未提及学习者对于该类短语存在过度使用

现象。

3. 语篇功能类。语篇功能类短语在文本中起到组织和连接作用,有助于建立语句之间的逻辑关系,指示信息结构,管理话题流程。这些短语可以用于引入新的话题、转换话题、总结前述内容、表明因果或对比关系等,从而促进读者对文本的理解和解读。正确使用语篇功能类短语是实现学术写作有效沟通的关键。根据 Simpson-Vlach 与 Ellis [5] 的分类标准,该大类可以进一步划分为四个子类别:元语篇与语篇指示、主题引入与聚焦、主题阐述、以及语篇标记。表 10 中的统计结果显示,COCA 学术库中语篇标记类短语占比最大(36.7%, n = 18),主题阐述类次之(30.6%, n = 15),主题引入与聚焦类占比最小(8.2%, n = 4)。科技英语摘要库中元语篇与语篇指示类短语占比最大(46.5%, n = 33),主题阐述类次之(36.6%, n = 26),主题引入与聚焦类占比最小(1.4%, n = 1)。

表 10 中的统计结果表明,主题引入与聚焦类短语在两个语料库中的占比均为最低。而科技英语摘要库的元语篇与语篇指示类短语占比接近 COCA 学术库的两倍,可能存在过度使用的问题。

**Table 10.** Statistics for discourse organizing expressions phrases in the two corpora

**表 10.** 两个语料库中语篇功能短语分类统计

语篇功能分类	科技英语摘要库	COCA 学术语料库
元语篇与语篇指示	33 (46.5%)	12 (24.5%)
主题引入与聚焦	1 (1.4%)	4 (8.2%)
主题阐述	26 (36.6%)	15 (30.6%)
语篇标记	11 (14.9%)	18 (36.7%)
总计	71	49

#### 4. 结论和教学启示

本研究深入对比分析了国内科技英语摘要库和 COCA 学术语料库中的学术短语使用情况,并据此构建了一份详尽的科技英语摘要学术短语列表。这一研究旨在揭示两种语料库之间的语言使用差异,并为学术英语写作教学提供指导。研究发现,本研究与前人的研究存在一些共性。首先,两个语料库中动词结构的使用比例均居中,这与 Biber [25] 研究结果一致;其次,属性说明类指示短语的使用频率最高,与 Simpson-Vlach & Ellis [5] 的发现相吻合;最后,英语学习者在学术写作中存在缺乏人际互动的问题,这与杨成[21]的研究结论相符。此外,也与前人的研究存在一些差异,通过与 COCA 语料库的详细对比,我们发现如下特征:

**结构分类特征:** 在国内科技英语高引学术论文摘要中,动词类及非动词类实词学术短语占了绝大多数,而功能词类短语的出现频率相对较低。相较于本族语者,国内科技英语学习者在英语写作中更倾向于使用动词结构。这一现象反映了中英两种语言在语法结构上的根本差异,即英语倾向于使用名词,而汉语则更多使用动词。国内科技英语学习者在输出英语时出现的母语负迁移现象提示我们,在学术英语写作教学中需加强语言思维转换的训练。

**功能分类特征:** 在功能分类方面,指示功能类短语在两个语料库中均占了绝大多数,尤其是科技英语摘要库中的属性说明类短语数量庞大,这可能与科技期刊摘要文本的特征密切相关。立场功能类短语的使用占比接近五分之一,两个语料库在模糊限制语和能力与可能性短语方面的使用存在显著差异,这可能与国内科技英语学习者倾向于使用确定语气表达观点,以及更多关注试验的结果和价值的习惯有关。语篇功能类短语中,元语篇与语篇指示类短语存在过度使用的问题,因此在学术英语写作教学中,应强

调元语篇与语篇指示类短语的重要性, 帮助学生学会如何清晰地组织文章结构, 使读者能够容易地跟随作者的思路。

在学术英语写作教学中, 教师应该鼓励学生积累和应用功能词、识别与聚焦类短语、模糊限制语和语篇标记类短语, 以增强学术写作的准确性和逻辑性, 加强人际互动。此外, 学生应适度减少能力与可能性类短语、元语篇与语篇指示类短语的使用, 以避免语言的冗余和过度使用。本研究的语料选取局限于矿冶类高引用学术期刊, 仅探讨了连续性短语的结构与功能分布, 未来的研究可以将视野扩展至不同学科领域, 以及非连续性短语的使用情况。

## 基金项目

昆明理工大学 2022 年课程思政教改项目一般课题“融育于教、以用促学”大学英语混合式思政课程建设与实践的阶段性研究成果。

## 参考文献

- [1] 刘宇红, 殷铭. 学术词组中词汇语法连续体的双向互动研究[J]. 语料库语言学, 2021, 8(2): 29-40+165-166.
- [2] Wray, A. (2002) *Formulaic Language and the Lexicon*. Cambridge University Press.  
<https://doi.org/10.1017/cbo9780511519772>
- [3] Biber, D. (2004) If You Look at ...: Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics*, **25**, 371-405. <https://doi.org/10.1093/applin/25.3.371>
- [4] Hyland, K. (2008) As Can Be Seen: Lexical Bundles and Disciplinary Variation. *English for Specific Purposes*, **27**, 4-21. <https://doi.org/10.1016/j.esp.2007.06.001>
- [5] Simpson-Vlach, R. and Ellis, N.C. (2010) An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics*, **31**, 487-512. <https://doi.org/10.1093/applin/amp058>
- [6] Ädel, A. and Römer, U. (2012) Research on Advanced Student Writing across Disciplines and Levels. *International Journal of Corpus Linguistics*, **17**, 3-34. <https://doi.org/10.1075/ijcl.17.1.01ade>
- [7] Martinez, R. and Schmitt, N. (2012) A Phrasal Expressions List. *Applied Linguistics*, **33**, 299-320.  
<https://doi.org/10.1093/applin/ams010>
- [8] Ackermann, K. and Chen, Y. (2013) Developing the Academic Collocation List (ACL)—A Corpus-Driven and Expert-Judged Approach. *Journal of English for Academic Purposes*, **12**, 235-247. <https://doi.org/10.1016/j.jeap.2013.08.002>
- [9] Morley, J. (2015) *The Academic Phrase Bank: An Academic Writing Resource for Students and Researchers*. The University of Manchester Press.
- [10] Römer, U. (2010) Establishing the Phraseological Profile of a Text Type. *English Text Construction*, **3**, 95-119.  
<https://doi.org/10.1075/etc.3.1.06rom>
- [11] Lu, X., Yoon, J. and Kisselev, O. (2018) A Phrase-Frame List for Social Science Research Article Introductions. *Journal of English for Academic Purposes*, **36**, 76-85. <https://doi.org/10.1016/j.jeap.2018.09.004>
- [12] Cunningham, K.J. (2017) A Phraseological Exploration of Recent Mathematics Research Articles through Key Phrase Frames. *Journal of English for Academic Purposes*, **25**, 71-83. <https://doi.org/10.1016/j.jeap.2016.11.005>
- [13] 屈典宁, 邓军. 基于语料库的语块习得模式研究[J]. 外语界, 2010(1): 47-53.
- [14] 徐昉. 中国学习者英语学术词块的使用及发展特征研究[J]. 中国外语, 2012(4): 51-56.
- [15] 鲁莉, 王敏. 英语学术论文中词块使用的学科间差异研究[J]. 西安外国语大学学报, 2015(3): 36-40.
- [16] 潘璠. 语料库驱动的英语本族语和中国作者期刊论文词块结构和功能对比研究[J]. 外语与外语教学, 2016(4): 115-123.
- [17] 李梦骁, 刘永兵. 中国学习者英语学术论文结论语步的词块特征研究[J]. 外语教学, 2017, 38(1): 34-38.
- [18] 李晓红. 学术英语写作中短语框架的功能特征——基于中英博士论文引言语料库的对比研究[J]. 外语电化教学, 2021(1): 98-104+16.
- [19] 段士平. 学术英语语块发展研究动态路径: 复杂系统建模[J]. 外语学刊, 2019(1): 87-91.
- [20] 陈艳君. 基于语料库的农科类学术英语短语框架列表构建研究[J]. 语料库语言学, 2021, 8(2): 16-28+165.

- [21] 杨成, Henderson, F., 朱战炜. 基于可比语料库的中外期刊 ESP 论文摘要词块研究[J]. 西南科技大学学报(哲学社会科学版), 2023(4): 57-64.
- [22] 周惠, 刘永兵. 中国英语学习者学术语块使用的语篇评价意义研究——以学术立场语块的语义韵为例[J]. 外语界, 2015(6): 35-44.
- [23] Cortes, V. (2004) Lexical Bundles in Published and Student Disciplinary Writing: Examples from History and Biology. *English for Specific Purposes*, **23**, 397-423. <https://doi.org/10.1016/j.esp.2003.12.001>
- [24] Biber, D., et al. (1999) Longman Grammar of Spoken and Written English. Pearson Education Limited.
- [25] Gray, B. and Biber, D. (2013) Lexical Frames in Academic Prose and Conversation. *International Journal of Corpus Linguistics*, **18**, 109-136. <https://doi.org/10.1075/ijcl.18.1.08gra>
- [26] 许家金, 梁茂成. 创建子语料库, 促成对比研究[J]. 当代外语研究, 2011(10): 6-9+60.
- [27] Tan, Y. and Römer, U. (2022) Using Phrase-Frames to Trace the Language Development of L1 Chinese Learners of English. *System*, **108**, Article 102844. <https://doi.org/10.1016/j.system.2022.102844>
- [28] 雷蕾. 基于 Python 的语料库数据处理[M]. 北京: 科学出版社, 2020.
- [29] 卫乃兴. 基于语料库和语料库驱动的词语搭配研究[J]. 当代语言学, 2002, 4(2): 101-114.
- [30] Lu, X. (2012) The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern Language Journal*, **96**, 190-208. <https://doi.org/10.1111/j.1540-4781.2011.01232.1.x>
- [31] 许家金. 语料库语言学的理论解析[J]. 外语教学, 2003, 24(6): 6-9.
- [32] 连淑能. 英汉对比研究增订本[M]. 北京: 高等教育出版社, 2010.
- [33] Huizhong, Y. (1986) A New Technique for Identifying Scientific/Technical Terms and Describing Science Texts: (An Interim Report). *Literary and Linguistic Computing*, **1**, 93-93. <https://doi.org/10.1093/lc/1.2.93>
- [34] Hyland, K. (1998) Hedging in Scientific Research Articles. John Benjamins Publishing Company. <https://doi.org/10.1075/pbns.54>