基于语料库的英语学术词块跨学科考察

陈 欣
(南华大学 外国语学院,湖南 衡阳 421001)

[摘 要] 文章基于自建的涵盖三大学科、库容为 2'774'166 词的英语学术论文语料库,调查分析了三类功能词块(即研究导向词块、文本导向词块、参与者导向词块)在不同学科期刊论文中的分布规律。研究结果表明,学术语篇词块中研究导向词块比例最高、文本导向词块其次、参与者导向词块最低;词块跨学科分布差异显著,主要体现在工科学术语篇中研究导向词块占总词块数量一半以上,人文社科中文本导向词块所占比例高于研究导向词块,此外,参与者导向词块在语言学中的比例最高,经管学科次之,工程类学科最低。学术词块跨学科分布规律与各学科研究内容、方法、范式及汇报方式密切相关。对英语学术词块使用特征进行基于语料库的跨学科考察有助于提升学习者英语学术写作能力,并能为 EAP 与 ESP 教学提供参考资源。

[关键词] 语料库; 跨学科; 学术词块
[中图分类号] H31 [文献标识码] A [文章编号] 1673 - 0755(2015)04 - 0135 - 06

词块是介于词与句子间的语言单位,是扩展化搭配(extended collocation),是多词组合的形式与意义结合体[1]。随着 AntConc、WordSmith 等语料库检索工具的问世,从大规模语料库中对词块进行提取、甄别与统计的技术操作已极为便捷。语料库技术的不断发展以及人们对词块习得重要性认识的逐步深入促进了近年来国内外词块研究的持续升温。

对词块使用特征进行跨体裁、跨语域的调查与分析已成为语料库语言学研究的一大热点。美国语料库语言学家 Biber 对课堂教学话语、日常会话话语及教材文本中的词块进行了对比分析,结果表明,课堂教学话语词块数量最多,分别为后面两种语体词块的两倍与四倍。这源于课堂教学话语既具有口头话语的特征,又不乏书面正式用语的要素。此外,不同词块在话语中承载着不同的语用功能,在不同语域中的分布情况也迥异,如教学话语中语篇组织词块与立场词块的比例远远高于教材文本[2]。与此类似,Cortes [3]、Scott & Tribble[4] 以及梁茂成等[5] 也发现了不同语域文本词块的系统差异,学术论文中典型的词块在二语学习者书面语中却很不常见。同时,学术词块使用特征与规律业已在学术英语研究中引起广泛关注。不过相关研究存在语料代表性较差、语料库库容较小、囿于关注某些或某类词块的使用情况、词块的识别与提取方式缺乏多方验证等一系列问题。基于此,本文力图对英语学术词块进行基于语料库的跨学科考察。

一 英语学术论文语料库及词块提取

本研究所使用的语料基于自建的英语学术论文语料库(English Academic Article Corpus, 以下简称为 EAAC)。语料涵盖以下学科门类:语言学(Linguistics, L)、经济、商务与管理(Economics, Business and Management, EBM)、工程、能源与技术(Engineering, Energy and Technology, EET)。选择上述三大学科除了它们涵盖几乎所有高校重要学科门类、符合研究的实用性外,更重要是的,它们在研究内容、研究范式与论证模式等诸多方面存在较大差异,将其作为考察语料有助于揭示词块使用的跨学科多样性。

所有文本均取自于全球著名的学术期刊出版社 Elsevier Science 电子数据库,笔者对其所涵盖的上述三大学科影响因子排名位居前列的国际性学术期刊进行调查,每学科选择 6 本代表性期刊,对 2012—2015 年间刊载的研究性论文进行随机取样,每期刊遴选学术论文 20 篇,合计 360 篇。所有论文都删除标题、摘要、关键词、致谢、参考文献、附录及

[收稿日期] 2015 - 04 - 06
[基金项目] 湖南省哲学社会科学基金项目"基于语料库的英语学术词块使用特征研究"(编号:14YBA330)、"基于 MD/MF 的英文学术期刊论文语体特征的跨学科考察"(编号:14YBA331)资助
[作者简介] 陈欣(1983 -),女,湖南郴州人,南华大学外国语学院讲师。

图表等冗余信息,仅保留无杂质的正文文本,建成英语学术论文语料库(EAAC),其库容(即型符数)达2'774'166 词。表 1 为英语学术论文语料库及三个子库的基本情况:

表 1 英语学术论文语料库(EAAC)基本情况

学科及子库简称	期刊名	影响因子(IF)	库容(型符数)
语言学(L)	Journal of Memory and Language	2.647	1,024,952
	Journal of Second Language Writing	1.846	
	Journal of Neurolinguistics	1.596	
	Journal of Phonetics	1.365	
	English for Specific Purposes	0.953	
	Language & Communication	0.852	
经济、商务与管理(EBM)	Journal of Financial Economics	3.769	1,061,094
	Evolution and Human Behavior	2.866	
	Journal of Accounting and Economics	2.833	
	Social Science & Medicine	2.558	
	Health & Place	2.435	
	Journal of Retailing	1.193	
工程、能源与技术(EET)	International Journal of Plasticity	5.971	688,120
	Renewable & Sustainable Energy Reviews	5.510	
	Journal of Power Sources	5.211	
	Bioresource Technology	5.039	
	Solar Energy Materials and Solar Cells	5.030	
	International Journal of Hydrogen Energy	2.930	
总计	每刊 20 篇,共 360 篇学术论文		2,774,166

由于四词词块比五词词块更为常见,同时比三词词块更能明晰地呈现结构与功能,因此,四词词块是本文的调查对象。四词词块的提取与统计使用语料库工具 Antconc3.2.4,首先将 EAAC 语料库导入 Antconc 软件,进入 Cluster 选项卡,设定 N-gram 的长度为 4 词,最低频数设定为 20,并且值域标准为至少在 5 个不同文本中出现,通过词块生成及人工判断、删除非语言单位及不符合语法的单位,生成 4 词词块总类符数(type)为 262 个,4 词词块总频次(frequency)为 16327,四词词块总型符数(token)为 65308,占 EAAC 语料库总型符数的 2.4%。笔者进而对四词词块按照出现频次的高低排序,得到英语学术四词词块表。表 2 为英语学术四词词块表中最高频的 20 个词块及其频次:

表 2 英语学术论文语料库中最高频的 20 个四词词块

词块	频次	词块	频次
on the other hand	297	in the form of	178
at the same time	251	the nature of the	164
in the context of	239	as well as the	152
in the present study	231	the end of the	149
in terms of the	225	as a result of	146
on the basis of	223	with respect to the	135
the level of the	210	at the end of	128
the extent to which	193	are more likely to	119
at the level of	185	can be used to	113
as a function of	179	to the fact that	108

从表 2 可见,词块有别于传统语言学上的成语(idiom)、搭配(collocation)及词汇语法关系(lexico-grammatical association),词块体现着多个单词之间在某一语域中存在高频率、高分布的共现关系。这些四词词块有的是结构完整的固定表达(fixed expression),如 on the other hand、at the same time、in the present study 等。然而在更多情况下,词块并不具备结构完整性(如 in the context of, in terms of the 等),但它们语义透明、形式规则,是拓展的搭配(extended collocation)。此外,不难看出,学术词块中常见结构类型为介词短语 + of 片段(如 in the context of, in terms of the, on the basis of, at the level of 等)、名词短语 + of 短语片段(如 the level of the, the nature of the, the end of the)、其他介词短语片段(如 on the other hand, at the same time)、名词短语 + 其他后置修饰语片段(如 the extent to which),它们是构建学术语篇不可或缺的语言成分。

二 学术词块的分类

词块可依据多个标准进行分类,如依据结构成分,可以分成"n. + and + n."、"prep + n."、"v. + n. + prep."等结构类型;依据在文本中出现频次情况,词块可分为高、中、低频词块;依据词块含有词数,分为二、三、四、五词词块等;依据语法属性,词块可分为名词性、动词性、形容词性词块等[6]。限于篇幅,

本文对词块不做结构类型等方面的分类与统计。本研究旨在探究词块在学术文本中的话语功能,故词块在语篇中的功能特征是分类的依据。

本分析词块的理论框架主要参考 Hyland 对词块的功能类型范畴[7]。学术语篇是作者向读者汇报其研究的文本。不难看出,学术语篇涉及到 3 大要素,即研究、文本及参与者(含作者与读者)。因此,作为学术语篇构建的基石,学术词块或多或少凸显研究、文本或参与者中某一因素。故从话语功能维度出发,学术词块可分为研究导向词块(research-oriented chunk)、文本导向词块(text-oriented chunk)与参与者导向词块(participant-oriented chunk)[8]。

(一) 研究导向词块

研究导向词块的话语功能在于帮助作者对其在真实世界中的实验、研究活动与经历进行构建,包括:定位词块(location):表明时间/场所,如 at the same time, in the present study 等;程序词块(procedure),如 the operation of the, the use of the, the role of the 等;量化词块(quantification),如 one of the most, a wide range of, the magnitude of the 等;描写词块(description),如 the structure of the, the surface of the, the size of the, 等;主题词块(topic):彰显研究领域,如 as a second language, English for specific purpose 等。

(二)文本导向词块

文本导向词块则主要关乎文本的组织结构及其作为信息或论证的意义,文本导向词块细分为 4 类:过渡词块(transition):建立事物间增补或对比的逻辑关系,如 on the other hand, at the same time, in addition to the 等;结果词块(resultative):标示事物间推理、使役或因果关联,如 as a result of, it was found that, the results of the 等;建构词块(structuring):对话语片段进行组织或对文本信息位置进行指向,如 as shown in the, in the next section, in this chapter we 等;架构词块(framing):通过提出限定条件与前提对研究论证进行定位,如 the extent to which, with respect to the, with the exception of, on the basis of 等。

(三)参与者导向词块

参与者导向词块聚焦于文本的作者或读者,根据参与者的身份特征可将参与者导向词块分为两类:立场词块(stance):传达作者的态度与评价,如 is more likely to, may be due to, I would like to 等;介入词块(engagement):直接与读者形成交流,如 we can see that, as can be seen, it is important to 等。

三　学术词块的跨学科分布考察

依据学术词块功能分类标准,对研究导向词块、文本导向词块与参与者导向词块在 3 大学科中的使用分布情况进行统计(如表 3 所示)。在学术语篇中,文本导向词块使用最为高频(46.3%),研究导向词块次之(38.0%),参与者导向词块使用最为低频(15.7%)。跨学科学术语篇中词块使用的横向比较则表明,各类词块在不同学科学术语篇的分布比例具有显著性差异,学术词块使用呈现鲜明的跨学科异质性,这表明不同学科领域的研究者力图通过他们的语言选择彰显其所属学科的特殊性:

表 3　三类四词词块的跨学科分布情况(词块型符数及所占比)

学科	研究导向词块	文本导向词块	参与者导向词块	合计
语言学(L)	7424(30.8%)	11980(49.7%)	4700(19.5%)	24, 104(100%)
经济、商务与管理(EBM)	9380(37.2%)	11852(47.0%)	3984(15.8%)	25, 216(100%)
工程、能源与技术(EET)	8008(50.1%)	6428(40.2%)	1552(9.7%)	15, 988(100%)
总体	24812(38.0%)	30260(46.3%)	10236(15.7%)	65, 308(100%)

(一)研究导向词块的跨学科分布特征

在所调查的 3 大学科中,总体上而言,研究导向词块占词块总量的 38.0%,且在所有学科中的其比例均在 30% 以上。这表明无论是对于人文社会科学,还是自然科学,研究导向词块对学术语篇的构建都至关重要。学术论文呈现的是作者研究的过程与研究的发现,有关“研究”本身的描述语言不可或缺,研究导向词块正是能承载这一语言功能的词块类型。具体而言,研究导向词块的话语功能具有多元性,主要涵盖定位时间/场所,如例(1)汇报程序,如例(2)进行量化,如例(3)进行描述,如例(4)体现主题,如例(5)(注:词块用斜体标示,所有例子均提供语境片段及来源学科代码,下同)。

(1) They also engage in a number of metalinguistic processes, pointing to elements of language and texts, while *at the same time* discussing content. (L)

(2) *The use of a* diverse set of empirical networks not only provides our study with a real-world setting as

the diffusion environment… (EBM)

(3) Moreover, *the magnitude of the* tax effects, both overall (3.6 and 4.4 percent) and incremental (2.6 and 3.6 percent), are analogous to those of corporate tax shelters. (EBM)

(4) The market risk premium is *the difference between the* expected return on an investment and risk-free rate. (EET)

(5) The present study systematically examined one pre-service teacher's beliefs and practices *regarding written teacher feedback*. (L)

尽管研究导向词块在各学科学术论文数量大、比例高,但其跨学科分布差异较显著。通过对其在不同学科中的分布进行跨学科比较不难发现,其在工程、能源与技术学科中所占比例是最高的,达到50.1%,明显高于另外两个学科。研究导向词块在理工科学术论文中的高频使用充分凸显了自然科学领域以物理世界、实验室为中心的研究取向。很大一部分研究导向词块旨在对研究对象或研究背景进行描述,对模型、设备、材料、或者研究环境方面进行具体化,其典型结构为名词短语 + of 结构,如例(6)、(7)所示:

(6) As a consequence, *the size of* the microstructure is generally no longer negligible with respect to the specimen dimensions and size effects occur modifying the mechanical properties. (EET)

(7) Nevertheless, *the value of the* critical strain involving two stages for HP relationship depends on the mechanical test. (EET)

在工程、能源与技术学科论文中,研究导向词块的语言功能还旨在阐述研究的具体程序,汇报实验或研究的开展方式,如例(8)、(9)、(10)所示:

(8) The comparison result Pe is used to manage *the operation of the* fuel cell and affects the system response based on the energy storage media used. (EET)

(9) *The purpose of the* PTLs is to enable electronic, thermal and mass transport between PEMFC subcomponents. (EET)

(10) As nickel sheets are mechanically isotropic, the von Mises (V. M.) criterion *can be used to* compute the equivalent stress, equivalent strain and triaxiality. (EET)

(二)文本导向词块的跨学科分布特征

总体而言,文本导向词块在学术词块中的比重最大,占46.3%。三大学科的横向比较则表明:其在语言学学术论文词块中所占比例最高(49.7%),经管学科次之(47.0%),工程类学科最低(40.2%)。也就是说,无论是人文社科还是自然科学领域,文本导向词块都是极其重要的词块类型。这源于学术论文的书面语体属性,借助于文本导向词块进行信息组织与文本架构是必要的。

文本导向词块在语言学与经管学科学术词块中所占比重显著性高于工程、能源与技术学科,这一结果与人文社科研究实证性较弱,思辨、阐释与理论性较强的特点不无关联。研究者在语篇中进行知识重铸的方式往往是通过从伦理层面寻求读者的理解、认同与移情,而非认知层面的说明与分析[9]。尽管作者观点的形成亦是以对真实世界现象的考察为基础,但知识的建构方式通常为理论思辨与逻辑推导,而非实验实证式的探索与考证。因此,总体上而言,人文社科研究成果的汇报更力求论证过程详尽而充分,而文本导向词块正切合这一需要,因为文本导向词块在学术语篇中的话语功能主要为确定互文联系、提供理论依据、建立概念关联、信息位置指引以及指出适用范围。鉴于此,社会科学文本中大约50%的文本导向词块是用来通过确立逻辑关联,明确个案,指出不足作为论述建构框架:

(11) A firm with a specific economic objective can therefore choose a transactional form that offers an optimal tax outcome *in terms of the* timing, character, and source of related gains/losses. (EBM)

(12) This will allow us to classify individuals into segments *on the basis of* these characteristics. (EBM)

(13) However, these two measures diverged *with respect to the* role of duration as a predictor. (L)

此类词块通常由介词 + of 结构构成,力图使读者聚焦于某一具体的情形,或者具体说明某一假设成立的条件,或者阐释、比较与强调论证的某些方面。尽管在自然科学学术语篇中架构词块占文本导向词块的比例较大,但其在语言学与经管类学术论文中比重更高。因为人文社科期刊论文的潜在读者的知识背景更为多元,既有本专业或本学科的学者,还包括别的领域旨在寻求理论与成果应用的研究者或实践者。在这一情形下,语言学与经管类学科期刊论文作者对研究背景的定位更为谨慎,并力求更为详尽地阐释各个因素或变量之间的关系。在文本导向词块中另一高频词块类型为建构词块。它们旨在为研究论证提供立足与投射的支架,如:

(14) *In the present study*, several experiments

were carried out to select optimal operating conditions for the mineralization of simulated textile wastewater. (EET)

(15) *In this section we* summarize the implications of our research to the managers of both a downstream retailer and an upstream manufacturer. (EBM)

建构词块还能为文本中某些内容、信息、材料等进行位置指引,帮助读者明确作者的意图。工程类论文尤其重视此类词块的使用,这反映这一学科对图表以及数字类信息的依赖,需要通过这些信息来支撑他们的论证:

(16) *As shown in Table* 1, 164 episodes were observed in eight hours of recorded lessons. (L)

(17) However, for an identical value of triaxiality, the kind of mechanical test has no influence *as shown in Fig*. 9. (EET)

通过调查发现,结果词块也是文本导向词块中的重要类型,它们旨在介绍对研究过程和结果的阐释与解读:

(18) Indeed, *it was found that* in 9.3% of all observed Russian cases the interview moved inside after the initial interaction and after the enumerator asked the first few questions. (L)

(19) Overall, *these results suggest that* derivatives users avoid more tax than non-users (H1), and…… (EBM)

(三)参与者导向词块的跨学科分布特征

参与者导向词块承载的语用功能包括表达立场与介入两类。参与者导向词块的使用体现了学术语篇以作者和读者为中心的特征,彰显了学术文本中的互动[10]。所谓立场(stance),指的是作者在语篇中显性地表达认知与情感判断、做出评价分析等。而介入(engagement)指的是作者将读者视为参与者,使其介入于语篇之中,并与其开展交流。

在所有的参与者导向词块中,2/3 旨在表明作者的立场,它们的绝大多数出现在人文社科文本之中。在人文社科论文写作中,作者需要通过更为显性地表达立场、评价、判断来确立自己的论点,如:

(20) …while some uninformed investors may be forced to exit the financial markets because of poor performance, they *are likely to be* replaced, at least in part… (EBM)

(21) However, it is possible that the syllable-based generalization found in Experiment 2 was inadvertently facilitated by exposing participants to one-syllable CVC items… (L)

这些例子不但揭示了在人文社科文本中立场词块的广泛运用,而且也表明立场词块能在较大程度上传达作者对于自身所构拟的研究预设所持的保守态度,以便向读者申明其论点具有个体主观性。此类词块尤其以先行词 it 结构最为典型。它们能帮助作者免于提出错误的阐释,并提醒读者在应用成果时需慎重处理。

立场词块广见于人文社科学术语篇之中,但在自然科学学术语篇中,介入词块则更为常见。介入词块是作者构建学术语篇互动的有效策略,因为介入词块明示了读者身份在学术语篇中的出席,使其在阅读文献时产生"介入感",互动过程中引导他们对研究过程与结果的关注,提升其批判性认识。介入词块的结构特征常表现为含有表示义务语义的情态动词(如 should 等)或表示必要性与重要性的表语形容词(如 important 等)。

(22) *It should be noted* here that a higher LPSP means less system reliability to meet an external load. (EET)

(23) *As can be seen* in Table 1, the pooled dataset is relatively balanced in terms of the number of sentences and words from each type of speech. (L)

(24) When compared with the data on English speaking respondents *we can see that* the majority of English interviews occurred either across a threshold or outside for the entire duration of the interview. (L)

由此可见,作者在研究写作中体现了对话体维度,引导读者的行动与理解。这些词块旨在为读者定位,要求他们关注文本中的某些信息,引领他们形成对论证与阐释内容的理解。在理工科论文中这类词块更为常见,这反映了自然科学领域更关注研究的精确性,尤其要确保对研究流程与结果的正确理解。自然科学语篇中知识建构以线性思维与问题驱动为突出特征,论证与推理过程力求高度标准化与简明性。

四　结语

本文的研究主要有以下两点发现:

其一,三类功能词块在学术论文中呈现不均衡分布,总体而言,文本导向词块所占比例最高,研究导向词块次之,参与者导向词块最低。这源于学术论文的书面语体属性,必须进行信息组织与文本架构;其次,不同于其他语域,学术论文呈现的是作者研究的过程与研究的发现,有关"研究"本身的描述

语言不可或缺;最后,不同于对话语体,学术文本的互动性并不强。

其二,词块使用的功能维度具有显著的学科异质性,承载研究导向、文本导向与参与者导向三大功能的词块在不同学科期刊论文中分布迥异:在工程、能源与技术学科学术论文中,研究导向词块所占比例最高,文本导向词块其次,参与者导向词块最低;语言学与经管学科的三类词块比例排序相同,即文本导向词块比例最高,研究导向词块其次,参与者导向词块最低;尽管参与者导向词块在三大学科中均占最低的比例,但其在语言学中的比例明显最高,经管学科次之,工程类学科最低。学术词块的跨学科异质性与理工科注重实证实验研究,而人文社科注重理论思辨阐释的学科差异性不无关联。

对学术词块使用特征进行基于语料库的跨学科考察有助于推动学术用途英语(EAP)与专门用途英语(ESP)研究的发展。通过明确学术语篇词块使用规律与特征,我们对学术文本语体特征的认识更为全面与深入。对学术词块的跨学科调查与分析对专业英语教材与词典的编写具有一定的参考价值。此外,鉴于中国英语学习者的学术词块产出存在诸多问题,本文的研究能为词块教学提供参照范本,有助于提升学习者词块输出的准确性与流利度,并全面增强其英语学术写作能力。

[参考文献]

[1] Biber D, S Johansson, G Leech, S Conrad, EFinegan. Longman Grammar of Spoken and Written English [M]. Harlow: Pearson, 1999.

[2] Biber D. University Language: A Corpus-based Study of Spoken and Written Registers [M]. Amsterdam: Benjamin, 2006.

[3] Cortes V. Lexical bundles in published and student disciplinary writing: Examples from history and biology [J]. English for Specific Purposes, 2004, 23(4): 397-423.

[4] Scott M, Tribble C. Textual Patterns [M]. Amsterdam: Benjamin, 2006.

[5] 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.

[6] 马广惠. 词块的界定、分类与识别[J]. 解放军外国语学院学报,2011(1):1-4,127.

[7] Hyland K. As can be seen: Lexical bundles and disciplinary variation [J]. English for Specific Purposes, 2008, 27(1): 4-21.

[8] 肖庚生,陈 欣. 英语学术论文词块特征研究[J]. 吉林工商学院学报,2013(3):113-116.

[9] Hyland K. Disciplinary interactions: Metadiscourse in L2 postgraduate writing [J]. Journal of Second Language Writing, 2004, 13(2):133-151.

[10] Hyland K. Stance and engagement: A model of interaction in academic discourse [J]. Discourse Studies, 2005, 7(2):173-191.

A Corpus-based Interdisciplinary Investigation into English Academic Chunks

CHEN Xin

(*University of South China*, *Hengyang* 421001, *China*)

Abstract: Based on a self-built 2, 774, 166 word corpus of three-discipline English academic articles, this paper explores the disciplinary variations in the frequencies and preferred uses of chunks in terms of discourse function, i. e. research-oriented chunks, text-oriented chunks, and participant-oriented chunks. The findings indicate that research-oriented chunks occupy the highest proportion of all the chunks in academic articles, text-oriented chunks the next, participant-oriented chunks the lowest. It is also found that there are significant disciplinary differences in chunk distribution, which, to be specific, include the following aspects: firstly, research-oriented chunks take up more than half of all the chunks in the academic articles of engineering discipline, while in the disciplines of humanities and social sciences, the proportion of text-oriented chunks is higher than that of research-oriented chunks; furthermore, participant-oriented chunks account for the highest proportion in linguistics, the discipline of economics, business and management the next, engineering discipline the lowest. The characteristics of disciplinary chunk distribution are closely related to research objects, methods, paradigms and ways of presentation across disciplines. The corpus-based investigation of English academic chunks across disciplines can help enhance learners' English academic writing skills and provide reference resources for the teaching of EAP and ESP.

Key words: corpus; cross-discipline; academic chunk