

中国学者 SCI 论文中四词词簇结构特点研究

江苏师范大学 姜晓艳

提 要: 本研究检索自建中国学者 SCI 论文语料库中高频四词词簇,从类型数、总数方面分析其结构特点,且分别针对中国学者 SCI 四个分区与自建本族语学者 SCI 论文语料库进行对比研究。整体上中国学者使用最多的类型是介词、动词短语词簇;与本族语学者相比,中国学者 SCI 各分区高频词簇的差异集中在使用总数上,特别是在带 of 的名、介词短语,主、被动态动词短语以及状语从句方面差异较大。本研究可以为中国学者撰写英语学术论文提供重要的参考指导。

关键词: 中国学者; SCI 论文; 四词词簇; 结构特点

一、引言

近年来,中国学者将在 SCI 上发表论文作为其学术科研成果的一大重要表征,各高校、科研单位也将在 SCI 上发表论文作为统计科研人员研究成果的重要手段,在此背景下,英语学术论文特点的研究呈现出上升趋势,“学术英语已成为专门用途英语研究领域的热点”(孙凤兰,2015),而语言问题成为众多非英语国家学者在国际上发表论文的一大困难(徐昉,2014)。^①

词簇,又称“词块”“词串”“多词序列”“多词单位”等,是“文本中反复出现的最高频的词语序列”(Biber *et al.*, 1999: 989; Biber *et al.*, 2004),是一种具有词汇与语法特征的语言结构(杨惠中,2002),可作为整体存储和使用(Wray, 2000),从而无形中减轻了语言处理和输出的负担,使语言交际更加快捷、流利、有效(马广惠,2009)。学术文体中一些高频词簇能够提高语言表达的地道性,彰显作者在某一语言社团的“内行人”身份(王芙蓉、王宏俐,2015),因此,学术英语中词簇的研究已成为一大热点,而计算机技术及语料库语言学的发展为词簇的提取提供了便利以及大量的客观数据,所以本研究将以自建的英语学术论文语料库为基础,使用 WordSmith Tools (Scott, 2012) 检索软件,全面描述中国学者 SCI 论文中高频词簇的结构,发现中国学者词簇使用特征。

二、相关研究

在学术英语语篇中,词簇的研究一直是一大

热点,有诸多来自国外的研究成果,如: Biber *et al.* (1999)、Biber *et al.* (2004) 分别调查了会话与学术论文中的词簇、课堂教学与教科书中的高频词簇,发现词簇使用在四个方面存在显著特征; Biber (2009) 研究了口语和书面语中词簇结构的固定性,发现口语中的词簇比较固定,而书面语结构中功能词比较固定,填充的实词种类不一; Jalali & Moini (2014) 对发表的医药期刊中学术论文引言部分的四词词簇做了研究,总结其结构和功能特点;此外部分学者还研究了不同学科之间词簇使用差异(Cortes, 2004; Hyland, 2008b);二语学习者与本族语者之间词簇使用差异(Chen & Baker, 2010; Ädel & Erman, 2012);不同水平学习者词簇使用差异(Hyland, 2008a; Staples *et al.*, 2013; Huang, 2015),还有一些学者从本族语期刊论文中选取目标词簇,用于检测二语学习者对这些词簇的使用(Karabacak & Qin, 2013; Qin, 2014);部分学者研究了二语学者与本族语学者学术词簇使用差异,如 Pérez-Llantada (2014) 和 Pan *et al.* (2016) 分别对比了西班牙语学者、中国学者的期刊论文与本族语学者期刊论文中的四词词簇。

然而,相比国外研究,国内研究关于学术英语中词簇的研究数目、对象及方法都远不够全面深刻。主要涉及本族语学者学术论文中词簇特点(杨元媛,2013;王芙蓉、王宏俐,2015);中国学习者与本族语学习者学位论文中词簇对比研究(周惠、刘永兵,2015);中国学习者学位论文与本族语学者

^① SCI (Science Citation Index),即科学引文索引,是目前国际上公认的最具权威性的科技文献索引。

期刊论文中词簇对比研究(娄宝翠, 2010; 徐昉, 2012; 王丽、李清婷, 2014); 中国学者与本族语学者英语期刊论文词簇对比研究(郑红红, 2014; 孙凤兰, 2015)。这些研究发现中国学习者在学位论文写作方面过多依赖某些词簇, 同时没有或过少地使用本族语者常用的词块, 而且关于中国学者英语学术论文中词簇的研究所建立的语料库学科较少, 容量较小, 提取的高频词簇数目不足, 不能全面描述中国学者英语学术论文写作的特点。因此, 有必要建立较大规模的语料库对中国学者英语学术论文中的词簇结构进行研究, 以揭示中国学者英语学术论文的特点, 为中国学者撰写高层次的英语学术论文提供参考, 为中国学习者提高学术论文写作能力提供一定的指导。

三、研究设计

1. 语料库的建立

本研究自建了中国学者 SCI 论文语料库, 即 SCI-CH 语料库, 语料库的总容量约为 458 万字, 1 000 篇学术论文。所有论文全部来自中科院 SCI 分区表中四个分区, 选取每个分区中影响因子较高的国际期刊, 发表时间为 2014—2015 年, 覆盖的学科包括数学、物理、化学、生物、工程技术、农林科学、环境科学、地学这八个学科, 且每个分区中每个学科选取的论文数目一致。所有论文的第一作者均为汉语姓名且所属单位为中国大陆的高等院校或研究所, 论文中的作者信息、图表、致谢、附录等其他信息全部删除, 只保留正文部分。具体数据见表 1。在表 1 中, 一区的字符数相对二、三、四区较少, 这是由于国内学者在一区较高影响因子的期刊中发表论文数量有限, 但即便如此, 也可以完全满足建立语料库对于字符数的要求。^①

表 1. SCI-CH 语料库统计信息

	分 区				总计
	一区	二区	三区	四区	
学 科 数目	8	8	8	8	8
每个学 科选取 篇数	15	40	30	40	125
文本数	120	320	240	320	1 000
字符数	835 976	1 483 290	1 046 489	1 217 495	4 583 250

① 以 2015 年中科院 SCI 索引分区表为参照。

2. 词簇的提取

研究表明四词词簇结构往往包含三词词簇, 频数远高于五词词簇, 结构和功能更具变化性 (Cortes, 2004; Hyland, 2008a; 徐昉, 2012; 杨元媛, 2013), 所以本文将中国学者 SCI 论文中的四词词簇作为研究对象, 从 SCI-CH 语料库中提取了前 100 个四词词簇, 其中频率最低的出现了 106 次, 至少覆盖了 50 个文本, 符合高频词簇的提取惯例, 此外还从四个分区中各提取出频数最大的 50 个四词词簇, 这些词簇至少覆盖了 5 个文本, 用于对比各分区之间的结构分布。

3. 研究问题

本研究主要回答以下问题:

(1) 在 SCI-CH 语料库中, 四词词簇总体上呈现出什么结构特点?

(2) 在 SCI-CH 语料库各分区中, 四词词簇分布呈现出什么结构特点, 与本族语学者 SCI 论文相比是否存在差异?

四、分析与讨论

Biber *et al.* (2004) 对词簇的结构类型做了总结, 分为三大类: 包含动词短语的词簇、包含小句的词簇、包含名词和介词短语的词簇。由于其所统计的词簇包括了口语词簇和书面语词簇, 不完全适合本研究, 所以本研究在其基础上对词簇结构重新归类, 把中国学者论文中的词簇分为四大类和 11 个小类, 即包含名词短语的词簇(带 of 的名词短语、其他名词短语)、包含介词短语的词簇(带 of 的介词短语、其他介词短语、比较性表达方式)、包含动词短语的词簇(含 Be 动词短语、主动态动词短语、被动态动词短语)和包含从句的词簇(状语从句片段、That-从句片段、Wh-从句片段), 并从类型数 (Types) 和总数 (Tokens) 两个方面统计了这些词簇, 使用卡方检验 (Chi-square test) 和对数似然检验 (Log-likelihood test) 来判断它们之间是否存在显著性差异。

1. SCI-CH 语料库中四词词簇整体结构特点

如图 1 所示, 在 SCI-CH 语料库中, 高频四词词簇类型数和总数的分布一致, 类型数多的词簇, 总数所占比例也大, 其中包含介词短语的词簇无论在类型方面还是总数方面所占比例最大, 分别达到 35.00% 和 39.43%, 其次是包含动词短语的词簇, 而包含名词短语的词簇所占比例最小。

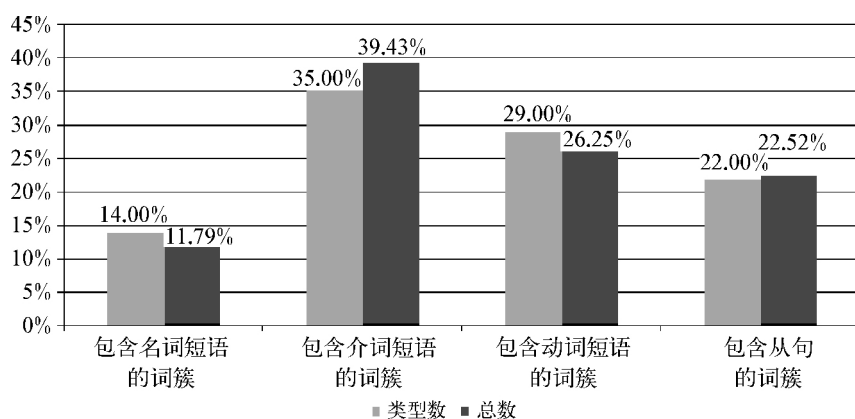


图 1. SCI-CH 语料库中四词词簇整体结构类型

此外,包含从句的词簇类型数和总数所占比例相差不大,而包含介词短语的词簇总数所占百分比超过了类型数所占百分比,说明中国学者使用包含介词短语的词簇不仅类型多,使用也更频繁。

包含名词短语和介词短语的词簇可以看作是短语类词簇,包含动词短语和从句的词簇看作是小句类词簇 (Biber *et al.*, 2004; Pan *et al.*, 2016), 图 2 显示中国学者在 SCI 论文中,类型数方面,短语类词簇所占比例少于小句类词簇,但在总数方面,短语类词簇所占比例却多于小句类词簇,这说明中国学者在论文撰写中,倾向于使用短语类词簇,但所掌握类型相对较少。

表 2 是 SCI-CH 语料库中所有结构类型的详细分布,可以看出在所有子类中,被动态动词短语的类型数和总数最多,这说明中国学者在撰写论文时十分依赖被动结构。对此,我们可以理解为在学术论文撰写中,国内学者较少使用第一人称复数 We,已有研究证实学术论文中被动与 we

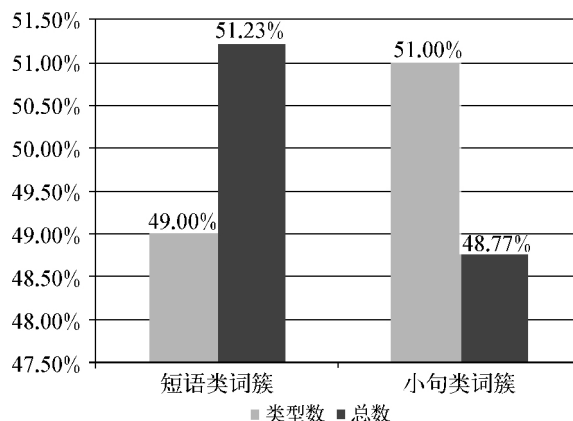


图 2. SCI-CH 语料库中短语类词簇与小句类词簇

呈负相关 (Millar *et al.*, 2013), 中国学者更多的将主语设定为所研究的物,努力使文章显得更客观。此外,其他三个大类中带 of-结构的名词短语、带 of-的介词短语、That-从句片段的使用数目较多、特征明显,对于上述的四个子类,本研究将详细分析其结构框架。

表 2. SCI-CH 语料库中四词词簇整体结构类型

结构类型	子 类	示 例	类型数(Types)		总数(Tokens)	
包含名词短语的词簇	带 of-结构的名词短语	the surface of the	10	14	1 552	2 267
	其他名词短语	an important role in	4		715	
包含介词短语的词簇	带 of-的介词短语	in the presence of	14	35	2 965	7 580
	其他介词短语	on the other hand	14		2 996	
	比较性表达方式	as well as the	7		1 619	
包含动词短语的词簇	含 Be 动词短语	is one of the	7	29	1 210	5 046
	主动态动词短语	plays an important role	2		299	
	被动态动词短语	are shown in fig	20		3 537	
包 含 从 句 的 词簇	状语从句片段	as shown in fig	5	22	1 838	4 329
	That-从句片段	we found that the	16		2 341	
	Wh-从句片段	which is consistent with	1		150	
总 计			100		19 222	

在对带 of 的名词短语详细观察统计后发现, the +名词+of the 这种结构最多,一共有 7 个类型,出现了 908 次,其类型数在名词短语的词簇中占 50%,在所有高频词簇中占 7%,如表 3 所示,填充名词为 surface, end, results, increase, formation, basis, effect; 在带 of 的介词短语中, in the +名词+of 的类型数最多,达到了 7 次,如表 3 所示,主要填充的名词为 presence, case, range, absence, regulation, terms, formation; 在被动态动词短语中, 被动+介词结构的词簇出现了 14 次,占被动态动词短语的 70%,在所有词簇中占 14%,总数也同样较多,是中国学者英语学术论文中的典型结构,如表

3 所示,主要被动式为 shown (4) , used(3) , listed, found, expressed, written, seen, based, organized, 其中 shown 和 used 占了一半; 在 That-从句片段中, The results/studies +动词+ that 结构比较典型,填充动词主要为 show (4) , indicate (2) , suggest, 其中 show 的频数最高。上述统计结果表明国内学者撰写学术论文的特点是采用被动结构,客观论述所做工作的方法、数据以及结论; 除此之外,无论是被动结构还是 That-从句片段中,使用最多的动词都是 show,中国学者使用 show 的频率非常高,既显示了中国学者对 show 的偏爱,也从侧面表明中国学者缺乏与 show 功能相似的动词,因而用词单一。

表 3. SCI-CH 语料库中最常用结构框架

词簇结构	具 体 词 簇	类型数	总数
the + 名词 + of the	the surface of the(188) ; the end of the(125) ; the results of the(124) ; the increase of the(123) ; the formation of the(121) ; the basis of the(116) ; the effect of the(111)	7	908
在所有高频词簇中所占百分比		7.00%	4.72%
in the +名词+of	in the presence of(488) ; in the case of(223) ; in the range of(197) ; in the absence of(179) ; in the regulation of(125) ; in terms of the(120) ; in the formation of(118)	7	1 450
在所有高频词簇中所占百分比		7.00%	7.54%
被动+介词结构	are shown in fig(379) ; is shown in fig(284) ; are listed in table(180) ; used in this study(164) ; can be found in(155) ; are shown in table(146) ; was used as the (141) ; can be expressed as(127) ; can be written as(127) ; can be seen from (122) ; was used as a(120) ; is based on the(113) ; results are shown in(110) ; is organized as follows(109)	14	2 277
在所有高频词簇中所占百分比		14%	11.85%
The results/ studies + 动词 + that	these results suggest that(178) ; the results showed that(174) ; results showed that the(132) ; these results indicate that(127) ; results show that the(122) ; studies have shown that(109) ; results indicate that the(107)	7	949
在所有高频词簇中所占百分比		7.00%	4.94%

2. 各分区结构特点

本研究除了对中国学者 SCI 论文整体结构特点进行分类统计外,还对 SCI-CH 语料库的四个分区,也就是 SCI 检索论文四个分区中中国学者使用的高频四词词簇进行了统计研究,具体统计数据如表 4 所示。表 4 列出了每个分区中高频词簇的分类,以

及各类词簇的类型数和总数,可以看出,各分区词簇结构分布与整体结构分布类似,使用最多的都是带 of 结构的介词短语和被动态动词短语。具体而言,一区和二区带 of 结构的介词短语类型数、总数最多,三区 and 四区被动态动词短语类型数、总数最多,使用最少的都是 Wh-从句片段和主动态动词短语。

表 4. SCI-CH 语料库中四个分区高频四词词簇结构分布

词 簇 结 构		类型数(Types)				总数(Tokens)			
		一区	二区	三区	四区	一区	二区	三区	四区
包含名词短语的词簇	带 of-结构的名词短语	6	3	5	5	234	210	244	339
	其他名词短语	4	3	1	2	131	213	80	151

续 表

词 簇 结 构		类型数(Types)				总数(Tokens)			
		一区	二区	三区	四区	一区	二区	三区	四区
包含介词短语的 词簇	带 of-结构的介词短语	11	9	9	8	524	867	549	596
	其他介词短语	8	9	8	8	420	765	601	656
	比较性表达方式	3	5	4	5	103	469	245	428
包含动词短语的 词簇	含 Be 动词的短语	3	3	3	3	131	235	155	225
	主动态动词短语	1	1	1	1	33	59	47	52
	被动态动词短语	6	8	12	9	233	680	665	738
包含从句的词簇	状语从句片段	4	2	3	2	282	509	411	370
	That-从句片段	4	6	4	7	163	372	182	458
	Wh-从句片段	0	1	0	0	0	67	0	0
总 数		50	50	50	50	2 254	4 446	3 179	4 013

除此之外,本文又从类型数和总数两方面,将四个分区的数据对比,通过卡方检验,发现各分区之间类型数方面没有显著差异,但总数方面差异显著,每两个分区之间都存在显著差异,如表 5 所示。由表 5,我们可以得出结论,中国学者在撰写 SCI 论文时,各区所使用的四词词簇类型相似,但每种类型词簇的使用总数却显著不同,因而我们认为中国学者 SCI 各区论文特点上不尽相同,这与传统思维中 SCI 论文各区之间的差异仅相关于科研成果相矛盾。

3. 各分区结构与本族语差异特点

为了对 SCI-CH 语料库中四个分区词簇结构的分布进行更进一步的分析,本研究从类型数和总数两方面,将四个分区的数据分别与自建的本族语学者 SCI 论文语料库(SCI-EN 语料库) 进行对比,探讨中国学者各分区与本族语学者之间

表 5. SCI-CH 语料库中四个分区结构之间卡方检验结果

卡方 检验 (Sig.)	类型数				总 数			
	一区	二区	三区	四区	一区	二区	三区	四区
一区	—	.738	.480	.801	—	—	—	—
二区	—	—	.822	.976	.000**	—	—	—
三区	—	—	—	.888	.000**	.000**	—	—
四区	—	—	—	—	.000**	.000**	.017*	—

注: * 表示 $p < 0.05$; ** 表示 $p < 0.01$

的差异。SCI-EN 语料库的建立方法同 SCI-CH 语料库,各分区比例与 SCI-CH 语料库相似,总字符数为 4 389 503。通过对数似然检验发现,与 SCI-EN 语料库相比,SCI-CH 中各区在类型数方面差异较小,但总数方面差异显著,具体统计数据如表 6 所示。

表 6. SCI-CH 语料库各分区与 SCI-EN 比较的似然对数结果

结 构		类型数(Types) LL 值				总数(Tokens) LL 值			
		一区	二区	三区	四区	一区	二区	三区	四区
包含名词短 语的词簇	带 of 的名词短语	2.99	-7.8**	-4.23*	-4.23*	-101.55**	-586.43**	-251.10**	-259.17**
	其他名词短语	0.94	0.27	0.44	0.00	+37.43**	+27.93**	2.42	+5.01*
包含介词短 语的词簇	带 of 的介词短语	0.06	0.07	0.07	0.29	-17.00**	-100.57**	-127.28**	-242.58**
	其他介词短语	0.21	0.55	0.21	0.21	+35.11**	+33.16**	+51.14**	+18.92**
	比较性表达方式	0.71	2.83	1.66	2.83	+4.61*	+263.46**	+85.72**	+252.17**
包含动词短 语的词簇	含 Be 动词的短语	0.05	0.05	0.05	0.05	0.97	-6.65**	-9.98**	2.97
	主动态动词短语	2.20	2.20	2.20	2.20	+129.27**	+166.13**	+157.08**	+154.57**
	被动态动词短语	0.38	0.00	1.10	0.08	-41.39**	0.39	+39.36**	+13.16**

续 表

结 构		类型数(Types) LL 值				总数(Tokens) LL 值			
		一区	二区	三区	四区	一区	二区	三区	四区
包含从句的 词簇	状语从句片段	+4.6 [*]	1.39	2.90	1.39	+440.32 ^{**}	+604.90 ^{**}	+594.96 ^{**}	+386.74 ^{**}
	That-从句片段	0.15	0.12	0.15	0.46	0.83	1.33	-16.03 ^{**}	+43.91 ^{**}
	Wh-从句片段	0.81	0.34	0.81	0.81	-31.32 ^{**}	+18.47 ^{**}	-42.91 ^{**}	-52.83 ^{**}

注: * 表示 $p<0.05$; ** 表示 $p<0.01$; +表示过多使用,-表示过少使用

表 6 列出了类型数和总数方面 SCI-CH 语料库各分区与 SCI-EN 语料库的似然对数结果。在类型数方面,中国学者一区论文与本族语学者论文的差异集中于状语从句片段,明显多用了状语从句片段;二、三、四区的中国学者与本族语学者相比明显少用了带 of 的名词短语;使用总数方面,大部分四词词簇结构存在着过度使用的现象,或多或少地说明中国学者在论文撰写中词句尚需精炼。

在包含名词短语和介词短语的词簇中,各分区中国学者都明显少用了带 of 的名、介词短语,而倾向于多使用其他类型的名、介词短语。有研究表明,名词短语词簇、介词短语词簇与论文的主要信息有着很强的相关性(Pan *et al.*, 2016),所以中国学者要提高带 of 的名词、介词词簇的使用量,同时减少使用比较性表达方面的词簇,从而更好地表达论文所提出的观点结论。在包含动词的词簇方面,主要的显著差异集中在主动态短语和被动态短语方面,其中主动态短语使用过多,被动态短语三区、四区使用过多,一区使用过少。许多中国学者认为被动语态的大量使用是学术英语的主要特征,但近些年来,学术英语中主动语态的使用比过去更加普遍,被动语态的使用有变少趋势(肖碧丹, 2015),英语国家的许多学者都认识到过多使用被动语态会使论文枯燥难懂,所以中国学者要注意主动结构与被动结构的使用,使得论文更加直接易懂又不失客观。在包含从句的词簇方面,各分区的状语从句片段都过多使用,但 that 从句片段、Wh-从句片段呈现出不同的使用趋势。整体而言,和本族语学者相比,各分区中国学者要适当增加带 of 的名词短语、带 of 的介词短语,减少比较性表达,注意主动、被动态动词短语的使用。

五、结论

本研究采用语料库研究方法,考察了中国学者在 SCI 检索论文中四词词簇的使用情况,并研究分析了中国学者在 SCI 各分区的学术论文中词簇使用情况。研究发现,整体上,中国学者学术论文中使用最多的词簇类型是介词短语的词簇,最

少的是名词短语词簇;各区论文与自建的 SCI-EN 语料库相比类型数方面差异较小,总数方面差异十分显著;明显少用了带 of 的名词以及带 of 的介词短语,多用了比较性表达、状语从句,需注意主动语态以及被动语态的使用。

基于本文的研究数据及结论,我们认为,中国学者在日常阅读文献时,应多注意本族语学者所撰写论文中带 of 的名词短语和带 of 的介词短语的使用;在撰写论文时,应减少比较性表达方式,注重使用带 of 的名词短语和带 of 的介词短语,且需合理安排主动、被动语态,使所撰写的文章直接易懂;最后,在修改润色论文时,可以借用语料库统计软件结合本文统计数据,进行指导性的修改,使论文撰写更加地道,减少论文撰写难度,这同样也是本文后续拟进行的研究工作。

参 考 文 献

Ädel, A. & B. Erman. 2012. "Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach". *English for Specific Purposes* 31(2) . pp81-92.

Biber, D. 2009. "A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing". *International Journal of Corpus Linguistics* 14(3) . pp275-311.

Biber, D., S. Johansson, G. Leech & E. Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Pearson Education Limited.

Biber, D., S. Conrad & V. Cortes. 2004. "If you look at ...: Lexical bundles in university teaching and textbooks". *Applied Linguistics* 25(3) . pp371-405.

Cortes, V. 2004. "Lexical bundles in published and student disciplinary writing: Examples from history and biology". *English for Specific Purposes* 23(4) . pp397-423.

Chen, Y. H. & P. Baker. 2010. "Lexical bundles in L1 and L2 academic writing". *Language Learning & Technology* 14 (2) . pp30-49.

Huang, K. 2015. "More does not mean better: Frequency

- and accuracy analysis of lexical bundles in Chinese EFL learners' essay writing". *System* 53. pp13-23.
- Hyland, K. 2008a. "Academic clusters: Text patterning in published and postgraduate writing". *International Journal of Applied Linguistics* 18(1) . pp41-62.
- Hyland, K. 2008b. "As can be seen: Lexical bundles and disciplinary variation". *English for Specific Purposes* 27 (1) . pp4-21.
- Jalali, Z. S. & M. R. Moini. 2014. "Structure of lexical bundles in introduction section of medical research articles". *Procedia-Social and Behavioral Sciences* 98. pp719-726.
- Karabacak, E. & J. Qin. 2013. "Comparison of lexical bundles used by Turkish, Chinese, and American university students". *Procedia-Social and Behavioral Sciences* 70. pp622-628.
- Millar, N., B. Budgell & K. Fuller. 2013. "'Use the active voice whenever possible': The impact of style guidelines in medical journals". *Applied Linguistics* 34(4) . pp393-414.
- Pan, F., R. Reppen & D. Biber. 2016. "Comparing patterns of L1 versus L2 English academic professionals: Lexical bundles in Telecommunications research journals". *Journal of English for Academic Purposes* 21. pp60-71.
- Pérez-Llantada, C. 2014. "Formulaic language in L1 and L2 expert academic writing: Convergent and divergent usage". *Journal of English for Academic Purposes* 14. pp84-94.
- Qin, J. 2014. "Use of formulaic bundles by non-native English graduate writers and published authors in applied linguistics". *System* 42. pp220-231.
- Scott, M. 2012. *WordSmith Tools* (version 6) . Stroud: Lexical Analysis Software.
- Staples, S., J. Egbert, D. Biber & A. McClair. 2013. "Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section". *Journal of English for Academic Purposes* 12(3) . pp214-225.
- Wray, A. 2000. "Formulaic sequence in second language teaching: Principle and practice". *Applied Linguistics* 21 (4) . pp463-489.
- 娄宝翠, 2010, 学习者硕士学位论文中的词串研究《当代外语研究》第 9 期: 27—33。
- 马广惠, 2009, 英语专业学生二语限时写作中的词块研究《外语教学与研究》第 1 期: 54—60。
- 孙凤兰, 2015, 中国学者科研论文中英语词串使用型式研究《外语教学》第 1 期: 69—74。
- 王芙蓉、王宏俐, 2015, 基于语料库的语言学和工科学术英语词块比较研究《外语界》第 2 期: 16—24。
- 王丽、李清婷, 2014, 二语学习者学位论文引言中词块的结构、功能以及语步特征《解放军外国语学院学报》第 4 期: 32—39。
- 肖碧丹, 2015, 科技英语被动语态使用量变化趋势《中国科技翻译》第 4 期: 5—8。
- 徐昉, 2012, 中国学习者英语学术词块的使用及发展特征研究《中国外语》第 4 期: 51—56。
- 徐昉, 2014, 非英语国家学者国际发表问题研究述评《外语界》第 1 期: 27—33。
- 杨惠中, 2002, 《语料库语言学导论》, 上海: 上海外语教育出版社。
- 杨元媛, 2013, 基于语料库的科技英语论文四词词簇特点研究《外语教学理论与实践》第 4 期: 45—52。
- 郑红红, 2014, 中国学者应用语言学英语论文中的词块研究《语料库语言学》第 1 期: 47—57。
- 周惠、刘永兵, 2015, 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, 《外语界》第 6 期: 35—44。

(通讯地址: 221116 江苏省徐州市铜山新区上海路 101 号外国语学院)
(文字编校: 徐江)

A study of the structural characteristics of the 4-word clusters in Chinese scholars' SCI academic papers

by JIANG Xiaoyan

Abstract: This study aims to retrieve the most frequent 4-word clusters in a self-built corpus of academic papers from SCI written by Chinese scholars, i. e. SCI-CH Corpus and to analyze the structural characteristics of these clusters from the perspective of both types and tokens. Besides, this paper makes contrast with native scholars the use of the frequent 4-word clusters in the four sub-corpora of the SCI-CH Corpus. It is found that the most frequently used clusters on the whole are those that incorporate prepositional and verb phrase fragments; the most frequently used clusters in each sub-corpus by Chinese scholars differ significantly in tokens from those used by native scholars, especially clusters from noun phrase with of-phrase fragment, prepositional phrase with of-phrase fragment, the active and passive structures and adverbial clause fragment. Therefore, this study can provide some guidance for Chinese scholars to write academic papers at higher levels.

Keywords: Chinese scholars; SCI academic paper; 4-word clusters; structural characteristics