基于语料库的英文论文摘要的语言特点研究

陆元雯

（上海交通大学语言文字工程研究所，上海 200240 / 上海交通大学外国语学院，上海 200240）

摘 要：本研究的结果表明，英文研究论文摘要中大量使用复数第一人称 we，较少使用单数第一人称 I，它们在不同学科间使用频率有较大差异。第三人称 it 做主语主要用在被动语态以表示研究结果，不同学科间使用 it 的频率差异不大。被动语态在各学科间使用频率差异不大，主要用于表示研究结果、研究方法和研究内容。高频名词、动词、形容词和副词的用词特点及其出现的典型结构表明，研究结果、研究方法、研究内容、研究目的是英文摘要的核心内容。

关键词：英文论文摘要；语料库；语言特点

中图分类号：H315 **文献标识码**：A **文章编号**：1002-722X（2009）06-0008-06

Linguistic Features of English Research Paper Abstracts: A Corpus-Driven Study

LU Yuan-wen

(Natural Language Processing Institute, Shanghai Jiaotong University, Shanghai, 200240, China / School of Foreign Languages, Shanghai Jiaotong University, Shanghai, 200240, China)

Abstract: Investigating the linguistic features of English research paper abstracts, the present study finds that the first personal pronoun "we" is frequently used, while "I" is used much less in the corpus. The frequency of "we" and "I" differs markedly across different disciplines. "It" as subject is mainly used in the passive voice to report research results, with little variation across disciplines. Frequency of occurrences of the passive voice varies little across disciplines, and it is often used to present research results, methods or contents. The analysis of the ten most frequently occurring nouns, verbs, adjectives and adverbs reveals that research results, methods, contents and objectives constitute the core of the abstracts.

Key words: English research paper abstract; corpus; linguistic feature

0. 引言

20 世纪 80 年代，随着计算机技术的高速发展，语言研究也翻开了崭新的一页。对大规模自然真实的语言进行处理不再是可望不可及的事情。语言研究者采用“数据驱动”的语料库方法探索语言使用的规律，应用检索软件对语料进行统计分析，总结归纳出语言特点。

本文将基于自建语料库，利用 WordSmith 3.0 对研究型论文的英文摘要特点进行分析。首先从高频词入手，发现研究对象，并比较它们在不同学科中的使用特点。然后从这些研究对象及与之共现的高频结构中揭示摘要的核心内容。

1. 英文学术杂志摘要语料库

为本研究专门建立的英文学术杂志摘要语料库收集了文、理、工、农、医 5 大学科英文杂志研究论文的摘要。选用杂志的标准是具有较高的影响因子，如 *Journal of the American Mathematical Society* 和 *Reviews of Modern Physics* 等杂志的影响因子在各自的学科中都是排名第一。每个大学科下包括两个子学科，每个子学科收集的杂志至少有 3 种或 3 种以上，共有语言学、法学、数学、物理、化学工业、一般工业、基础农业、综合农业、基础医学和综合医学等 10 个学科 5 178 篇研究论文英文摘要，库容为 893 969 词。语料的时间跨度为 2001 至 2005 年。详细构成见表 1。

利用 WordSmith 3.0 对该语料库进行统计，所得主要数据见表 2。

收稿日期：2009 - 05 - 04

基金项目：教育部人文社会科学研究规划项目（05JA740020）

作者简介：陆元雯（1968 - ），女，浙江嘉善人，上海交通大学外国语学院副教授，博士，研究方向为语料库语言学。

2. 统计结果与讨论

2.1 第一人称 we 和 I

第一人称复数 we 和单数 I 在语料库中出现的频数分别为 3 446 次和 187 次。We 出现的频数在词频统计中高居第 21 位。以 we 和 I 做主语的句子分别占该语料库句子总数的 11% 和 0.6%。这一结果与一般认为学术论文要少用或尽量不用第一人称来体现客观性的观点显然不一致。它表明，在英文学术杂志摘要中，we 使用得相当多，人们并没有因为要体现文章的客观性而刻意回避它。

但是对各子语料库进一步检索发现，各个学科使用 we 的频率有较大差异。理科中数学杂志使用 we 的频数最高，综合性医学杂志次之，而综合性农业杂志中的频数最低。由于各子语料库的库容不同，将频数转换成频率的结果见表 3。

表 1. 各语料库构成及库容

学科		摘要篇数	库容（词数）
文科	语言学	750	55 627
	法学	273	37 790
理科	数学	628	65 430
	物理	479	54 998
工科	化学工业	697	151 944
	一般工业	431	69 580
农业	基础农业	252	58 478
	综合农业	567	128 637
医学	基础医学	369	90 805
	综合医学	732	180 680
总计		5 178	893 969

表 2. 主要统计数据

形符（token）	893 969
类符（type）	31 796
类符形符比（type/token ratio）	3.56
标准化类符形符比（standardized type/token ratio）	41.41
句子总数（sentence）	30 817

表 3. We 在各子语料库中的频数及频率

学科	频数	各子语料库中的频率（‰）
数学	1 179	18.0
综合医学	959	5.3
物理	390	7.1
基础医学	370	4.1
语言学	174	3.1
化学工业	134	0.9
法学	89	2.4
基础农业	65	1.1
一般工业	55	0.8
综合农业	31	2.4
总计	3 446	

可以看出，We 主要使用在理科和医学杂志的英文摘要中，在文科中使用相对较少，而在农业和工科的杂志中使用得很少。葛冬梅等（2005）曾对电子电气、金融和外科医学 3 个学科中的 150 篇英文摘要中 we 的使用情况做过调查。结果显示，金融学科中 we 的使用频率要远超出另外两个学科。本研究证实了工科杂志的摘要的确较少用到 we，但发现在医学杂志中有不少 we 的出现。不同的结果或许与抽样有关。

另外，本研究显示，文科（如语言学和法学）使用 we 的情况并不是特别多，与预想中因文科主观色彩浓、摘要中会较多使用 we 的想法不符。这种现象还有待进一步研究。

检索发现，与 we 共现词频最高的前 10 个 3 词组合为：

We show that...(193)[①]、we prove that...(88)、we aimed to...(72)、we study the...(57)、we conclude that...(49)、we did a...(47)、we investigated the...(44)、we examined the...(35)、we present a...(33)、we prove the...(33)。

可以看出，we 主要使用在呈现研究内容的句型中。

统计结果显示，I 虽然也出现在英文摘要中，但与 we 相比，使用频率相当低，以 I 做主语的句子仅占语料库句子总数的 0.6%。I 在各子语料库出现的频数及频率见表 4。

表 4. I 在各子语料库中的频数及频率

学科	频数	各子语料库中的频率（‰）
语言学	96	1.7
法学	48	1.3
物理	20	0.4
数学	12	0.2
综合农业	6	0.05
综合医学	3	0.02
基础医学	2	0.02
基础农业	0	0
一般工业	0	0
化学工业	0	0
总计	187	

值得注意的是，I 在语言学和法学英文摘要中使用的频率要远远高于其他学科。这与 we 的使用情况形成鲜明对比。理科中使用 I 的情况同其他学科如医学和农业相比也相对较多。医学和农业杂志中很少使用 I，而在工科的摘要中则没有一例出现。从 I 的使用情况可以推测，不同学科的研究方式存在巨大差异。在文科和理科中，以个人的案头工作进行研究的方式并不少见，而医学或农业科学研究则多数都需要实验室、研究基地、团队合作等才能完成，工科尤其如此。

与 I 共现频率最高的 3 词组合是：

I argue that...(14)、I show that...(7)、In this paper I(will discuss/argue)...(6)、I show how...(4)、In this article I(recommend/argue/explore/focus on)...(4)、I want to(discuss)...(3)、I describe the...(2)，等等。

可以看出，I 同 we 一样也是出现在表示研究内容和研究结果的句型中。在典型的与 I 的 3 词组合中，动词 argue 的频数最高，而与 we 一起使用的典型的 3 词组合中，argue 一词的频数并没有那么高。这在某种程度上说明了文科研究的特点。相对于其他学科，文科研究的特点之一表现在它的结论往往是开放性的。

2.2 第三人称 it 做主语

第三人称单数 it 的出现频率也非常高，它在词频统计中列第 28 位，共出现 1 838 次。其中 it 用做主语的句子出现 1 374 次，占所有含 it 句子总数的 75%。其余 464 次 it 用做宾语，占所有含 it 句子总数的 25%。本节将主要讨论 it 用做主语的用法。

首先，对各子语料库进一步检索发现，it 做主语出现的频率在多个学科中没有特别大的差别。只有在农业和医学杂志的英文摘要中使用较少，以医学中的频率为最低。语言学和法学杂志使用 it 做主语的频率最高。表 5 是 it 在各子语料库中的频数及频率。

表 5. It 在各子语料库中的频数及频率

学科	频数	各子语料库中的频率（‰）
语言学	203	3.6
法学	100	2.6
一般工业	160	2.3
物理	127	2.3
数学	143	2.2
化学工业	297	2.0
综合农业	155	1.2
综合医学	87	0.5
基础农业	78	0.4
基础医学	24	0.3
总计	1 374	

其次，检索发现，与 it 共现的频数最高的 3 词组合有：it is shown...(85)、it was found...(79)、it has been...(53)、it is concluded...(34)、that it is...(34)、it can be...(28)、it is found...(26)、it was concluded...(24)、it is suggested...(23)、it is possible...(22)、it is argued...(21)，等等。

从这些典型句型中的动词可以看出，it 多数用在被动语态中，表达研究结果。

2.3 被动语态

在语料库的词频表（本文略）中，我们发现，使用最多的是系/助动词 BE：is、are、was、were、be 和 been。这些动词除了用在系表结构中，如 found to be the/a...，主要是用来构成被动语态。经过软件检索再加人工识别，在语料库中使用被动语态的句子有 11 863 个，占语料库句子总数的 38.5%，说明摘要中主动语态的使用要多于被动语态。表 6 依次列出了这些动词在语料库中出现频率的高低顺序、构成被动语态的句子数、BE 的各变体用于构成被动语态的百分比以及各变体构成被动语态的句子数量占被动语态句子总数的百分比。

表 6. BE 动词：频数及各变体使用的百分比

BE	词频排序	频数	构成被动态句子数	用于被动态百分比	占被动态句子总数百分比
is	9	7 187	2 287	32%	19%
was	10	7 056	2 940	42%	25%
were	12	6 108	2 851	47%	24%
are	18	3 736	1 419	38%	12%
be	23	3 048	1 350	44%	11%
been	39	1 357	913	67%	8%
being	2 674	264	103	39%	1%
总计			11 863		100%

从表 6 可以看出，超过 2/3 的 been 都用来构成被动语态，was、were 和 be 的使用中有近一半被用到被动语态。此外，由 was 和 were 构成的被动语态在总被动语态中占到近一半的数量（49%），is 虽然在词频排序中最高，但在构成被动语态中只排第 3 位，占 19%。而在 being 的使用过程中，有近四成（39%）被用于被动语态，但由于该词在总语料库中的频数不高，因此构成被动语态的量最少，只占 1%，几乎可以忽略不计。

进一步检索各子语料库发现，各学科对被动语态的使用频率有差异。表 7 列出了各子语料库中被动语态的使用频数及频率。

表 7 显示，在工科的英文摘要中使用的被动语态最多，农业科学使用的被动语态在整个语料库中排第 2 位。联系上面的研究结果，这两个学科 we 和 I 的使用数量都排在整个语料库的后面，可以推测，同其他学科相比，这两个学科的英文摘要倾向于使用被动语态，且以第一人称为主语的主动语态使用较少。另一个现象是，医科中综合医学和基础医学的被动语态在使用数量上有较大差距（12.1∶6.1），其中的原因还有待进一步研究。

下面将具体分析 BE 动词的各个变体构成被动语态的典型结构。

表 7. 被动语态在各子语料库中的频数及频率

学科	频数	各子语料库中的频率（‰）
化学工业	3 097	20.4
一般工业	1 258	18.1
基础农业	961	16.4
物理	692	12.6
综合农业	1 567	12.2
综合医学	2 179	12.1
语言学	667	12.0
法学	346	9.2
数学	548	8.4
基础医学	552	6.1
总计	11 867	

动词 is 用于构成被动语态的典型结构有：it is shown...（90）、is associated with...（82）、is shown that...（80）、is based on...（76）、is used to...（56）、is concluded that...（36），等等。

动词 was 用于构成被动语态的典型结构有：was found to...（122）、was used to...（110）、it was found...（90）、was found that...（89），等等。

动词 were 出现的主要被动语态结构有：were found to...（86）、were used to...（78）、were carried out...（61）、were associated with...（42），等等。

动词 are 用于构成被动语态的典型结构有：are associated with...（35）、are used to...（29）、are compared with...（28）、are based on...（26）、are discussed in...（22），等等。

动词 be 用于构成被动语态的典型结构有：can be used...（104）、be used to...（79）、be used as...（35）、be associated with...（24）、could be used...（24），等等。

动词 been 主要用于构成现在完成时的被动语态，出现的主要结构有：has been studied...（35）、been shown to...（34）、have been used...（31）、been used to...（30）、has been developed...（26），等等。有趣的是，从 been 的检索行发现，过去完成时态出现得很少，而且主要是用在定语从句中。最典型的 3 个结构是：who had been...（15）、that had been...（11）和 which had been...（8）。

从上面所列典型结构中的动词可以看出，被动语态用来表达的内容主要有结果、研究方法和研究内容 3 类。

2.4 用词特点

从语料库的词频表可以看出，最高频率的词多数都是功能词，如前 5 个高频词依次是：the、of、and、in 和 to。由于单个的功能词不能体现其使用意义，本文重点关注名词、动词、形容词和副词的使用特点。表 8 列出了语料库中出现频率最高的前 10 位名词、动词、形容词和副词。

表 8. 前 10 位高频名词、动词、形容词和副词

序号	名词（频数）	动词（频数）	形容词（频数）	副词（频数）
1	patients (2 060)	is (7 187)	high (1 274)	only (846)
2	results (1 791)	was (7 056)	different (1 036)	significantly (725)
3	study (1 771)	were (6 108)	low (888)	respectively (486)
4	temperature (1 060)	are (3 736)	higher (817)	often (200)
5	model (1 055)	be (3 048)	significant (678)	strongly (190)
6	time (1 050)	have (1 652)	new (665)	highly (182)
7	rate (992)	has (1 483)	total (579)	rather (169)
8	data (982)	using (1 479)	potential (548)	mainly (163)
9	effect (935)	been (1 357)	lower (543)	previously (163)
10	activity (924)	used (1 260)	large (512)	especially (161)

1）前十高频名词

仔细观察前十高频名词后会发现一些有趣的现象。由于 patients 和 temperature 很明显是和研究主题有关的词汇，我们这里不去分析它们。位列第 2 位和第 3 位的高频名词分别是 results 和 study。毫无疑问，results 一词主要是用来表达研究结果的。它主要用在以下结构中：results suggest that...（110）、results indicate that...（82）、results showed that...（82）、results show that...（78），等等。这些结构是学术论文摘要中表明研究结果的最典型的句型。

高频名词 study 出现的典型结构有：this study was to...（81）、in the present study...（47）、aim of this study...（32），等等。可以看出，名词 study 主要用于表示研究目的和方法的结构中。

由于本研究所使用的语料库收集的全部是研究论文，model 一词出现频率高也就不足为怪，因为建立和使用模型是做研究不可或缺的一步。从使用 model 的主要结构看也可以证实这一点：of the model...（29）、model for the...（23）、a model of...（18），等等。这些结构主要体现了研究方法。

Time 一词出现的主要结构有：the time of...（47）、at the time...（32）、the same time...（30）、at the same time...（28）、the first time...（25）、and the time...（18），等等。从这些结构看，它主要是指研究过程中某个时刻，或用于描述研究进行过程中某

个时刻事物所处的状态。进一步观察表明，在 time 的 4 词组合中，出现频率最高的 3 个典型结构是 at the time of...(29)、at the same time...(28)和 for the first time...(23)。这说明研究论文不仅强调研究步骤的顺序性和时间性，还强调研究的新颖性。

Rate 一词出现的主要结构有：the rate of...(97)、rate of the...(23)、scalar dissipation rate...(19)、the reaction rate...(17)，等等。可以看出，rate 主要用于专业术语的表述，如：scalar dissipation rate、the reaction rate、heat release rate、the burning rate。

Data 一词出现频率高也很容易理解。在研究论文中，实验数据是最重要的。Data 主要用于：data suggest that...(36)、data from the...(23)、these data suggest...(22)、data were collected...(16)等等，表明 data 主要用于表示数据来源以及数据所显示的研究结果。

为了能更加清晰地看出第 9 位高频词 effect 所表达的内容，我们观察了它最高频的 4 词组合：the effect of the...(42)、had no effect on...(24)、of the effect of...(18)、determine the effect of...(16)、to determine the effect...(16)、investigate the effect of...(15)、to investigate the effect...(13)。从这些结构可以看出，effect 主要用于讨论研究结果或说明研究目的和描述研究内容。

使用第 10 位高频词 activity 主要的 4 词组合有：the catalytic activity of...(18)、antibacterial activity of chitosan...(7)、the antibacterial activity of...(7)，等等。从这些典型结构中可以看出，activity 与名词 rate 一样，主要出现在专业术语中，如：catalytic activity、antibacterial activity，等等。

2）前十高频动词

表 8 的前 10 个高频动词中有 6 个是系动词 BE 的变体，它们是：is、was、were、are、be、been。这些词已在被动语态中讨论过，这里不再赘述。

除了 BE 动词外，另外两个动词 have 和 has 在学术论文英文摘要中主要用来构成完成时态，少量用来表示拥有（如 to have a...）。Have 主要使用的结构有：have been used to...(21)、studies have shown that...(11)、we have investigated the...(10)，等等。从这些结构中可以看出，have 主要用于表示研究方法和结果。

Has 主要使用的结构有：has been shown to...(20)、has the potential to...(12)、there has been a...(10)，等等。可见 has 也是主要用于表示研究结果和方法的句子中。

前十高频动词的最后两个词是 using 和 used。它们分别列在第 8 位和第 10 位。Using 主要使用的结构有：in using hand tools...(7)、was carried out using...(7)、using a combination of...(5)，等等。

Used 主要使用的结构有：can be used to...(49)、have been used to...(21)、be used as a...(20)、can be used as...(20)，等等。

毫无疑问，这两个词主要用于介绍研究方法。它们的高频出现也说明介绍研究方法是英文摘要中的核心部分。

3）前十高频形容词

表 8 所列前 10 个高频形容词 high、different、low、higher、significant、new、total、potential、lower 和 large 都是中心形容词，即它们既能做定语，也能做表语。除 total 一词主要用于表示数量的结构外，如 a total of...(56)、total number of...(11)，其余 9 个按照语义功能可以分成两大类。一类是比较性形容词，如 high/higher、low/lower、different；另一类是评价性形容词，如 significant、new、potential 和 large。

在第一类比较性形容词中，与 high 搭配最多的出现在其后的名词是：temperature(83)、risk(59)、density(29)、grade(28)、pressure(19)，等等。

与 low 搭配频率最高的名词与 high 的基本相同，主要有：temperature(77)、grade(44)、pressure(24)、density(24)、temperatures(19)，等等。有趣的是，risk 没有与 low 搭配，cost 和 energy 没有与 high 搭配。这或许说明研究论文重点关注的有 high risk、low cost 以及 low energy，而非 low risk、high cost 和 high energy。

与 different 搭配频数最高的介词是 from (43) 和 between (11)；副词是 significantly (31)；名词主要有：types (19)、concentrations (12)、methods (11)、levels(9)、conditions(8)，等等。在 8 个频数最高的名词中，表示方法的就有 3 个。

在第二类评价性形容词中，与 significant 搭配频数最高的名词有：differences (63)、difference (41)、effect(27)、increase(26)、effects(18)、reduction (17)、decrease(15)，等等。这些名词多表示变化与不同。

与 new 搭配频数最高的名词有：method(15)、approach(11)、product(8)、applications(8)、results (7)、type(7)、technique(6)，等等。这些名词主要表示方法与结果。

与 potential 搭配频数最高的介词是 for（60）；名词主要有：uses(6)、use(5)、risk(5)、applications(5)，等等。这些名词主要表示方法和结果。

与 large 搭配的名词有：scale（41）、number(19)、class(17)，等等，主要体现量的不同。

从英文论文摘要的前十高频形容词的比较性和评价性的特点，我们可以推测，英文论文大多比较本研究与前人研究，并且对本研究做出评价。

4）前十高频副词

表8所列前10个高频副词 only、significantly、respectively、often、strongly、highly、rather、mainly、previously 和 especially 多数都用来增强语义。下面是这10个高频副词出现的典型结构。

与 only 共现最多的词是 not，即 not only 是 only 出现最多的词组。该词组共出现85次，占 only 使用的10%。此外，only 的典型结构还有：only a small...(13)、can only be...(12)、was the only...(12)、and only if...(10)，等等。

Significantly 主要用在下面这些结构中：was significantly lower...(26)、was significantly higher...(18)、not differ significantly...(15)、not significantly different...(14)，等等。可以看出，与 significantly 搭配的形容词主要是 lower、higher、different 等高频形容词。它的功能是增强这些形容词的语义。

Respectively 一词前面多为数字，以逗号相隔。没有与之共现的其他典型结构。

Often 虽然是高频词，但使用它的典型结构却很少，主要有：often associated with...(4)、are often used...(3)。这说明，often 作为一个简单副词，和它共现的词汇语义范畴很大，使用也相当灵活。

Strongly 主要的结构有：strongly associated with...(14)、strongly related to...(7)、strongly suggest that...(6)、strongly correlated with...(5)，等等。从这些结构中可以看出，strongly 多用于表示联系和关系，也有部分用于表示研究结果。

Highly 出现的典型结构不多，主要有：highly predictive of...(5)、a highly significant...(4)等。Highly 主要用于增强它所修饰的形容词的语义。

Rather 一词主要用在词组 rather than(107)中。该词组占使用 rather 总频率的63%。这也说明 rather 增强被修饰词语义的功能在论文摘要中并不占主导地位。

使用 mainly 的主要结构有：mainly due to...(9)、mainly on the...(6)、mainly by the...(4)、mainly attributed to...(3)、is mainly due...(3)，等等。从这些结构上看，它用于强调原因的居多，如：mainly due to...、mainly attributed to...。

Previously 修饰的主要动词有：reported(22)、published(9)、described(6)，等等，用来描述文献中的研究结果。

使用第10个高频副词 especially 的典型结构不多。被它修饰的词主要是形容词，有：interested(3)、early(2)、low(2)，等等。同副词 often 一样，和 especially 共现的词汇语义范畴很大，使用灵活，没有十分典型的结构出现。

从以上前10位高频名词、动词、形容词和副词的分析中可以看出，研究结果、内容、目的和方法是英文摘要的核心成分。这一点与 Pho（2008）的结果相似。Pho 分析了两种应用语言学杂志和一种教育技术杂志中的30篇摘要。结果表明，这些英文摘要有3部分内容是必须的，即本研究综述、方法描述和结论概述。

3. 结语

本文利用“数据驱动”的语料库研究方法对学术论文英文摘要的语言特点进行了跨学科探讨，以期为国内学者英文摘要的写作和教学提供参考。当然英文摘要的语言特点还有很多，除了本文提到的这些特点，摘要中的时态、类联接、词串的使用特点等等都值得进一步分析和讨论。

注释：

① 括号中的数字代表该词或该词组在英文学术杂志摘要语料库中出现的频数。下同。

参考文献：

[1] 葛冬梅，杨瑞英. 学术论文摘要的体裁分析 [J]. 现代外语，2005，28（2）：138－146.

[2] Pho，P. D. Research article abstracts in applied linguistics and educational technology：A study of linguistic realizations of rhetorical structure and authorial stance [J]. *Discourse Studies*，2008，10（2）：231－250.

（责任编辑　严辰松）