●语言学:语料库与语言专题

语料库关键词与专业话语国外研究述评*

何安平　郭桂杭
(广东外语外贸大学,广州 510420/华南师范大学,广州 510631;广东外语外贸大学,广州 510420)

提　要:本文通过国外文献述评展示用语料库关键词研究专业话语的意义和优势。首先简述关键词的延伸内涵和本质;然后通过专业语料示例介绍近年基于关键词拓展的多种型式;再通过关于专题隐喻、学科认知论和学科语言教学的 3 个案例剖析关键词用于专业话语分析的实施方法及效果启示。

关键词:语料库关键词;内涵与本质;拓展型式;专业话语;案例刨析;方法与启示
中图分类号:H030　　**文献标识码**:A　　**文章编号**:1000 - 0100(2020)01 - 0018 - 6
DOI **编码**:10.16263/j.cnki.23 - 1071/h.2020.01.002

A Survey of Corpus Keyword Approach in Specialized Discourse Research Abroad

He An-ping　Guo Gui-hang
(Guangdong University of Foreign Studies, Guangzhou, 510420/South China Normal University, Guangzhou 510631; Guangdong University of Foreign Studies, Guangzhou, 510420 China)

This paper focuses on value and advantages of corpus keywords approach in specialized discourse research. After introducing a broaden insight and nature of corpus keywords, the paper demonstrates a number of keyword extended patterns developed in studying specialized discourse in recent years. It further elaborates the implementation and implication of keyword approach by three case studies across areas of topic-based metaphor, discipline epistemology and disciplinary language teaching.

Key words: corpus keywords; insight and nature; extended patterns; specialized discourse; case analysis; method and implication

1　引言

《外国语言文学专业本科教学质量国家标准》关于“外语类专业可与其他相关专业结合,形成复合型专业或专业方向,以适应社会发展的需要”的专业定位推动国内对商务英语等学科英语的专业话语研究(教育部 2018:58)。同时也是国外语料库语言学趋向专门化语篇/语料研究的发展趋势之一(Hunston 2017)。其中,语料库视角下的关键词研究是探究专门学科话语的重要抓手。

2　语料库关键词

语料库视角下的关键词内涵近年出现新的延伸。它涵盖所有基于语料频数驱动、计算机自动提取、能凸显语篇关键性(textual keyness)的单个词或者多字词丛(key-words, key-clusters, key-phrases)(Scott 2014:232, Bondi 2010:3)。其中既包括目标语料与参照语料对比后自动产生的频率显著性高(或显著性低)的单词形关键词(key-words,或称主题词),也包括那些无需与参照语料比照而直接从目标语料自动提取的,但超过预设频数的多字词丛(clusters,或称 n-grams, lexical bundles)。这些关键性词语都被称为既有指向性又有可视性的“探针”(pointer),主要用来探测语料库或语篇的主题内容(aboutness)、文体风格

* 本文系教育部人文社科项目“基于语料库的汉英经济隐喻对比研究”(17YJA740014)的阶段性成果。
作者电子邮箱:fld02@scnu.edu.cn(何安平)

(style)和立场态度(stance)(Bondi 2010:7,Stubbs 2010:23, Scott 2010:51,刘辉 2018:69)。关键词的这 3 种功能表明,它已经不仅是一份词汇清单,而是"具有语篇本质属性"(Scott,Tribble 2006:56)。它们与系统功能语言学的三大元语言功能(ideational, textual 和 interpersonal)异曲同工,因为都分别指向语言"表意"(what)"表结构"(how)和"表态"(why)的本质,也因此赋予关键词可探究语言本体的内涵。

由于关键词产生的机理是对比语料里面有显著频数差别的词语,故浏览国内外众多标题带有 corpus,keywords 字样的文献,发现所使用的语料大都指向某个学科、某种专业职场、或某类机构的专业话语(specialized discourse),即"人们在学术、专业、技术和职业等专门领域的典型语境中使用的语言"(Gotti 2008:24)。而专业话语最突出的表征是在词汇层面(同上:33,65),所以识别那些由话题内容和体裁特征带出的显著性高频词语"肯定成为专业话语描述的根本要素"(Bondi 2010:3)。然而,孤立的单词并非描述意义的最好切入点,语料库语言学视角下的意义单位应设为可有多种变体的短语(Sinclair 2004:29 – 30),所以近年关键词法在专业话语研究中不断涌现出基于关键词拓展的各种短语型式。

3　关键词的多种拓展型式

3.1 单个关键词的拓展型式

基于单个关键词的拓展主要体现为关键词与关键词的共选、关键词与周边语境词的共选、以及众多同类语篇的关键词共享。其中一种可称为"关键词搭配词丛"(keyword collocates cluster),是由单个或一批关键词在 cluster 工具界面呈现的 n-字语词丛。词丛中的搭配词不一定是关键词,但能揭示关键词在语料中的典型相貌,同样比单个关键词更能揭示主题内涵。例如提取题为 Money as Debt 的语篇前几个单个关键词(如 gold, money, bank, claim)的 2 – 4 词词丛,可获得 gold and silver(真金白银),demand for real gold(要银行兑现真金),used as money(拿……当钱使),claim check holders(票据持有人),run on the bank(银行挤兑)等一批金融管理类话题的重要术语。

另一种是"关键主题词"(key Keywords,简称 KKW),是语料库内多个相关的独立文本共享的关键词(Scott 2014:231)。KKW 有助于归纳同类主题或同类体裁语篇的核心词群,以此揭示话题的体裁特征和表述某个话题的典型词汇(李文中 2003: 287, Gerbig 2010: 154)。例如,Gerbig (2010)对比 21 世纪和 20 世纪两个旅游话语语料库的 KKW,发现前者核心词群有 guy(s), locals, tourist(s), backpackers, travelers, tour, walk, ride, driver, hike 等名词,凸现旅游者自身及旅行方式的话题;而后者则凸现旅途的地貌风景,其核心词群是 hill, place, spot, stones, sea, mountain, valley, landscape, trees 等及颜色类形容词。

第三种是"关键主题词的关联词"(KKW associates),指的是 KKW 与关键词的重复同现(Scott,Tribble 2006:85),两者的关联构成围绕某一主题表达而触发的复杂词语网络,甚至揭示说话者对话题的心理认知(李文中 2003:288)。例如,以上提及的 21 世纪旅游语料库里有 3 个 KKW(tourist,traveler 和 backpacker),各自的高频关联词中都有 driver,trip,ride 等关键词;但 tourist 另外独有关联词 beach(海滩),似透出这类游客的休闲愿望。此外,tourist 和 traveler 共享关联词 tour,但 backpacker 却未共享;似透出"背包客"不太关注返回原地式的"巡游",而更注重昼夜时间和起居等活动(因为 backpacker 独有的关联词是 night, day, hours, food 等)。可见 KKW 的关联词分析还能折射出作者对话题内容的情感偏好态度(Gerbig 2010:157)。

3.2 多字词丛的拓展型式

基于多字词丛的拓展主要是对词丛内部的形式结构、词性特征和词序连续性等作进一步分类提取。其中一种称为"关键性短语"(key phrase),专指那些至少含有一个名词,且结构多为 N + N 和 Adj + N 的多字词丛(Panunzi et al. 2008:463 – 468),目的是凸现名词性短语对主题内容的有效揭示。例如,源自维基百科关于 1929 年经济大萧条话题语料中最高频的 4 个关键性短语(即 money supply, bank failure, stock market crash, gold standard)显然要比该语料中 4 个最高频的单个关键词(即 economy, bank, government, depression)更能揭示经济萧条话题的核心内容(同上:264 – 266)。

第二种是"短语框架"(phrase frame)。这是纯粹基于频数自动提取的非毗邻式的多字词丛(2 – 8 字),词丛内部除有一字不同,其余的都相同(Fletcher 2012)。短语框架内的空档(即 *)的填充词通常不是语篇关键词,但却能揭示语篇的体裁风格。例如,Grabowski(2015:276)曾对比药物学中"患者用药活页"(PIL)和"药品特点摘要"

(SPC)两种语篇类型的短语框架。发现 PIL 突显的是由 if you 构建的“虚词类框架”,如 if you* any,if you* to, if you* not,if you* a;显示这类话语直面患者的信息组织功能。而 SPC 突显的是由情态动词 should 构建的“动词类框架”,如 should be* by, should be* in, should be* to;而且填充词多为动词被动式(如 reduced, administered, initiated);显示这类语篇不掺杂个人情感、客观性和规约性较强的话语风格。

第三种是“主题性语序”(aboutgram)。它穷尽语料库在 2 至 12 字跨距内所有词之间的搭配频率,自动生成一批可含排序或位置变体的 2 至 5 词连贯或非连贯语序(即 AB,A * B ,B*** A),目的是廓清专题话语中所有词汇的共选相貌,从中识别高频而且有意义的短语型态,以揭示主题内容和体裁风格(Warren 2010:117 – 118)。例如,Warren(同上)发现香港理工大学工程学语料库(HKEC)最高频的实意词 design(133 次)就有 60% 可构成非毗邻的、词位排序不同的主题性语序;包括 design/structural, building/design, analysis/design, design/tall 等 2 词序列(/表示两词的跨距为 2 – 12 个词)。它们显然要比单个词 design 更清晰展示该学科话语的核心内容。

不论是基于单个关键词,还是基于多字词丛拓展的短语型式,其实都在不断对关键词进行频数上、形式上、或语义上的分类与归纳,其深层的理据是语料库语言学的词汇共选理论和多型态短语理念。这些拓展型式为揭示语篇的关键性,即前述的关键词 3 个功能,提供多样化的分析和诠释视角。下文进一步通过刨析 3 个完整案例,评述专业话语研究中关键词法的实施步骤及成果创新。

4　具体案例刨析

4.1 关键词辅助经贸隐喻探究

Philip(2010)在探讨机构体裁专题话语时提出“主题隐喻”(metaphor themes) 和“关键隐喻”(key metaphors) 两个新概念。前者指在专题语料中一组有明显语义关联,但喻词形式不一的语言隐喻,其靶喻却是该专题话语的关键词。后者指某主题隐喻中的源域词在局部语境内以显著方式与关键词同现的隐喻(同上:188,196)。基于 10 万词次的意大利前国际商贸部长在任期间的讲话和新闻发布语料,该研究采用一系列词频信息分类方法:(1)提取该语料的词频表,对其中排行前 500 个词作词簇化处理(lemmatizing),以避免这 500 词之外的词频表里还有同词根词;(2)对词频表中词次为 3 及以下的低频实意词作大致语义归类;(3)提取该语料的关键词表,也对其中的实意词作语义分类,以便归纳主话题及次话题;(4)在步骤(2)已归好类的低频实意词里识别显著不同于步骤(3)所归纳的主/次话题的语义类别词,用索引行工具调查该类的属下词,看其是否与某话题(如 trade)属下的关键词共选。

结果显示,该语料的关键词表内含有“国际贸易”话题(由 Italy, business, commerce, international, country, China 等排行前 10 位的关键词构成);而低频词表内有一批 war 类词(如 battle, fight,loser,aggressive),它们反复与上述话题的关键词共选,构成 international trade is war 这一主题隐喻(内含 Trade is aggressive behavior, Emerging economies are a threat 等次级概念隐喻)。进一步调查该主题隐喻的批量实例,我们发现其中的源域词与靶域词有相对固定的互选倾向(见下例句的斜体字):在谈及国际贸易时,上述的 war 类词往往与东方新兴经济大国有显著关联。例如,The fear that our businesses will end up as the *loser* in the *globalisation challenge*, especially when faced with the *commercial aggressiveness* of the *Far East*. 而谈及贸易扩张时,与 China India, Asian, Far East 共选的是 invade;与 Italy 共选的则是 penetrate. 例如,But *China* is the real future of the textile industry, because it is true that it has *invaded* us since quotas ended... 由此形成一批立场态度鲜明的关键隐喻。

该案例同时从词频表和关键词表切入,并且拓展为关键词的语境词搭配型式。其特点是聚焦那些与话题核心内容有关的隐喻,从而既化解专业话语的抽象概念,又揭示说话者对话题的隐含立场态度。而两者都是学科阅读素养的核心构成,由此启示我们,在探讨专题话语隐喻时,既要关注从关键词表归纳出来的主次话题词;又要关注整体词频表低频部分那些与主题内容显著不同的语义类别;因为“隐喻的源域词通常不会是专题话语的主题词”(Philip 2012:92)。

4.2 关键词辅助学科认知论对比研究

Malavasi 和 Mazzi(2010)的研究旨在廓清不同学科话语的认知论差异。首先假设:学科话语除了其独特的话题和词汇之外还有其独特的认知模式,即不同学科构建、论证、磋商和传播知识的独特范式。具体落实到学科话语对研究主体、研究内容(研究兴趣)和研究方法的词语表达(同上:169,172)。根据认知论的内涵界定,研究者

分别建立市场营销学和历史学两个论文库各 240 余万词次。首先在两个库的关键词表(两库互为参照语料)中各选出语义内涵分别指向研究主体、研究内容和研究方法的 5 个关键词(历史学的是 he, historians, text, science 和 society;营销学的是 we, research, data, results 和 effect)。接着分别对 5 个关键词作局部语境中与"报告类动词"(reporting verbs)的搭配分析;并且将这些动词分为"研究""认知"和"言说"3 类(Thompson, Ye 1991;Thomas, Hawes 1994)。然后归纳出两个学科的关键词与 3 类动词的搭配型式;再诠释这些型式所传递的认知模式信息。

结果发现,两个学科的认知模式很不一样。例如,从研究主体看,历史学的主体类关键词搭配型式是"historians/he + 言说类动词(如 argue/emphasize/say/claim/tell/state/conclude/report/explains/suggest/stress...)",显示该学科研究者为思辨者的身份特征(arguer);而营销学的搭配型式是"We + 研究类动词(如 use/find/examine/test/observe/demonstrate/study)",显示的是行动参与者的研究身份。从研究方式看,历史学的型式是"written /literary/medieval /authentic... + text + reveal/convey/narrate",似侧重文献研读和权威考证;而营销学的是"research /results/data + suggest /support/focus on/confirm...,似侧重基于实证材料和数据结果作结论。

该案例先将关键词作语义分类并选出代表词,然后拓展其在语境中与某类动词搭配的型式。其特点是找到关键词与抽象概念的关联途径。正如英国 South Sussex 大学的语料库 DNA 研究团队在 2018 年题为"Quantifying Concepts in Corpus linguistics"研讨会上指出:概念可内化于任何语言层面,包括语义、语用、语篇、社会文化和语法,等等。所以,对概念的量化分析途径要从概念的操作定义出发,努力达至概念内容的可视化,即呈现表述概念内涵的语言资源在各段语料库的相貌。其中的关键就是找到词汇语义对概念内涵的映射。

4.3 关键词辅助学科语言教学

Cacchiani(2018)从学科认知论视角探讨经济学术话语的词汇语法和语篇结构机制,并将成果应用于学科语言教学。鉴于经济学研究论文的核心是构建知识,其文本必内含"假设、分析、归纳、诠释、预测"等 5 类话语行为(Merlini Barbaresi 1983:3)。研究者首先提取 90 万词次的经济学论文库的关键词表,并专门关注表内那些语义与上述 5 类行为话语相关的语篇结构类和研究方法类的关键词(如 if, estimate, case, assumed, denote, suppose, then),结果发现 if 的关键值(keyness)位居前列。于是又提取该库的 3 – 5 字词丛,将其分为"研究型""语篇型"和"参与型"(转自 Hyland 2008:13 – 19),同样发现在"语篇型"属下的"框架标识类"词丛中含 if 的词丛最为突出。由此推导该学科话语具有"基于实证作假设,基于条件作预测"的知识构建特点(Cacchiani 2018:18)。

进一步拓展这些 if 词丛的语境发现:if 从句及其主句的动词时态基本不吻合传统英语教科书的语法搭配规则。故转向 Declerck 和 Reed (2001)的"Possible World 理论",从"可能的现实""可能的形式"和"可能的诠释"等视角探讨 if 从句的形式与功能。结果发现,经济学论文中用 if 构建知识的复杂性远远高于其形式结构的复杂性,其中包括以下情况。

(1)事实性假设:If, our survey showed, these debtors are unable to pay their own debts, they are insolvent.

(2)理论性假设:If environmental standards are reduced, production in the pollution-intensive sector becomes more efficient.

(3)修辞性假设:Countries that produce a same commodity usually face different values and signs of this correlation coefficient (Table 1) and, therefore, different gains, if any.

基于以上发现,教师改进对该学科研究生的英语教学。其中包括设问,例如:

(1)当你使用 if 从句时,你能在多大程度上判断该假设"肯定能""有可能""差不多能""几乎不能"实现或为真;

(2)在什么情况下你可以用其他词语替代 if(例如用 assuming, given that, in case)。此外,还设计了学科语境填空,例如:

(设语境为:If debtors are unable to pay their debts, debtors are insolvent. 请做以下选择:)We _____ of insolvent if they _____ pay their own debts.

(a) cannot speak ... are able to

(b) could not speak ... were able to

(c) will not speak ... will be able to

(d) would not speak ... would be able to

该案例的方法特点是同时从单个关键词表和多字词丛表切入,而且都仅关注其中的语篇结构类和研究类的词义类别;然后聚焦显著高频的相

关词丛作拓展语境分析。结果不仅坐实该学科构建知识的典型范式;而且深入探寻该范式的语言表达形式在学科话语与普通话语中的使用差异;进而改进教学设计。其启示为,学科的语言教学要涵盖学科认知范式的内容;要实施学科语境化教学,要结合学生已有的学科背景知识设计语言活动。

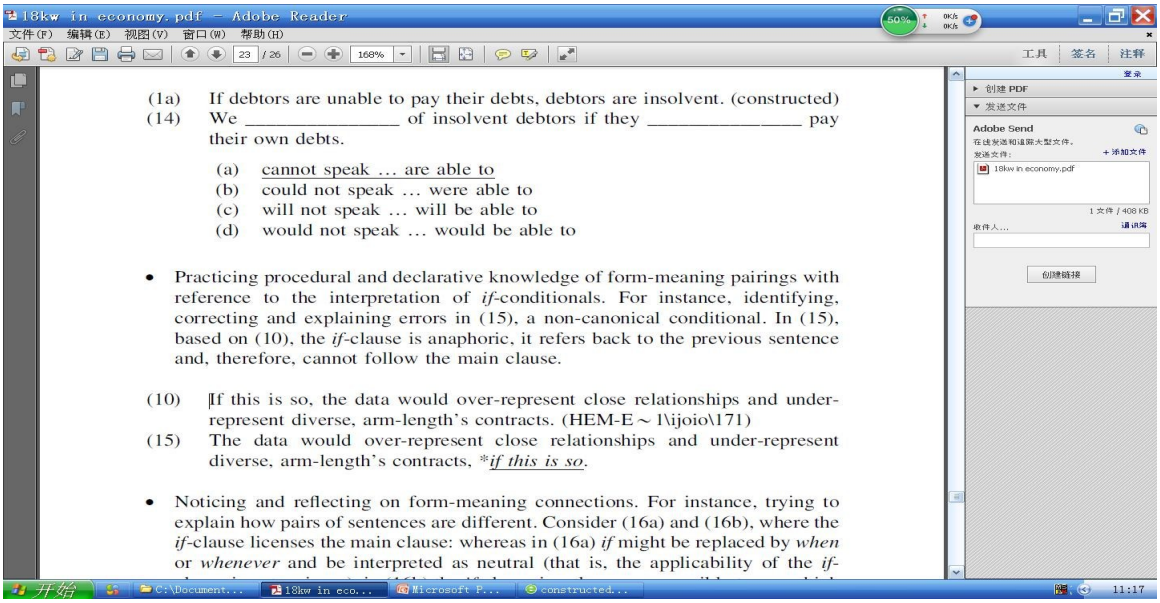

图 1　if 词丛案例图

5　结束语

本文通过阐述语料库关键词的拓展内涵和展示其在专业话语研究与教学中的应用,总结其优势至少有 3 方面。首先,关键词的研究目标直指专业语篇的本质,包括主题内容、研究范式、认知论特征、立场态度以及语篇体裁,等等。由此表明语料库关键词绝不仅仅表明统计学意义上显著性多或少的问题;而是可揭示语篇在“说什么”“怎么说”和“为什么这样说”等本质内涵。第二,关键词的研究路径不同于其它从语篇外部因素入手,或是仅对主观选定的词语作例证式分析方法;而是从语篇最底层的词汇频数入手,自下而上地探索语篇的本质属性;从而使研究结果具有客观性、量化实证性和典型性。第三是应用价值。通过关键词研究获取的专业话语典型特点,包括表述专题核心概念、专业体裁和认知模式等丰富语言资源,可直接应用于学科语言教学。这一点对于我国目前学科英语教学在高校英语教育的比重不断提升,大批通用英语教师正在向学科英语教学转型的现状尤其具有现实意义。

诚然,语料库关键词研究也有自身尚待解决的问题。例如,随着超链接文本的兴起,应如何框定文本的边界,如何对大批量关键词的分类和归纳提供清晰指引,如何解决参照语料库的内容和规模影响关键词提取结果的问题,等等(Scott 2010:52)。可喜的是,近年国外对关键词研究不断有成果创新。其中包括:Rayson(2008)在对目标语料进行自动词性和语义赋码之后,通过提取语料库或语篇的关键词词性类别和关键语义类别,揭示不同社团在同一体裁话语中的核心内容和立场态度差异。Murakami 等(2017)采用“主题建模”(topic modeling)中的 LDA 算法提取专业期刊文章的高概率共选词表以找到各种话题的关键性词群以及关键性语篇、从而揭示话题之间的关联以及话题的历时性变化。Davies(2018)新开发的 140 亿词次 iWeb 语料库(互联网免费检索),既可基于某个单词的在线检索走进所有以该词作为关键词的网页;又可基于若干关键词即时建成专门话题的虚拟网络语料库,等等。这些无不显示出大数据时代语料库关键词研究的广阔前景。

参考文献

教育部. 普通高等学校本科专业类教学质量国家标准[Z]. 北京:高等教育出版社, 2018. ‖ Ministry of Ed-

ucation. *National Standards for the Teaching Quality of Undergraduate Majors in General Colleges and Universities*[Z]. Beijing: Higher Education Press, 2018.

李文中. 基于英语学习者语料库的主题词研究[J]. 现代外语, 2003(2). ‖ Li, W.-Z. A CLEC-based Analysis of Key Words and Associates[J]. *Modern Foreign Languages*, 2003(2).

Bondi, M. Perspectives on Keywords and Keyness: An Introduction[A]. In: Bondi, M., Scott, M. (Eds.), *Keyness in Texts*[C]. Amsterdam: John Benjamins, 2010.

Cacchiani, S. If-Conditionals in Economics Research Articles: From Keywords to Language Teaching/Learning in the L2 Writing-for-Publication Class? [J]. *Corpus Pragmatics*, 2018(2).

Davies, M. The iWeb Corpus — Overview[EB/OL]. https://corpus.byu.edu/iweb, 2018-9-4.

Declerck, R., Reed, S. *Conditionals: A Comprehensive Empirical Analysis*[M]. Berlin: Mouton de Gruyter, 2001.

Fletcher, W. Phrases in English[EB/OL]. http://www.kwicfinder.com, 2019-2-20.

Gerbig, A. Key Words and Key Phrases in a Corpus of Travel Writing[A]. In: Bondi, M., Scott, M. (Eds.), *Keyness in Texts*[C]. Amsterdam: John Benjamins, 2010.

Gotti, M. *Investigating Specialized Discourse*[M]. Bern: Peter Lang, 2008.

Grabowski, L. Phrase Frames in English Pharmaceutical Discourse: A Corpus-driven Study of Intradisciplinary Register Variation[J]. *Research in Language*, 2015(3).

Hunston, S. Corpus Linguistics in 2017: A Personal View[EB/OL]. http://www.birmingham.ac.uk/cl, 2017.

Hyland, K. As Can Be Seen: Lexical Bundles and Disciplinary Variation[J]. *English for Specific Purposes*, 2008(1).

Malavasi, D., Mazzi, D. History v. Marketing: Keywords as a Clue to Disciplinary Epistemology[A]. In: Bondi, M., Scott, M. (Eds.), *Keyness in Texts*[C]. Amsterdam: John Benjamins, 2010.

Merlini Barbaresi, L. Gli atti Del Discorso Economico: La Previsione. Status Illocutorio e Modelli Linguistici Nel Testo Inglese. Parma: Edizioni Zara, 1983.

Murakami, A., Thompson, P., Hunston, S., Vajn, D. 'What Is This Corpus about?': Using Topic Modelling to Explore a Specialised Corpus[J]. *Corpora*, 2017(2).

Panunzi, A., Fabbri, M., Moneglia, M. Multilingual Open Domain Key-word Extractor Proto-type[A]. In: Bernal, E., de Cesaris, J. A. (Eds.), *Proceedings of* 13*th EURALEX international congress*[C]. Barcelona, Institut Universitari de Linguistica Aplicada, 2008.

Philip, G. Metaphorical Keyness in Specialised Corpora[A]. In: Bondi, M., Scott, M. (Eds.), *Keyness in Texts*[C]. Amsterdam: John Benjamins, 2010.

Philip, G. Locating Metaphor Candidates in Specialized Corpora Using Raw Frequency and Keyword Lists[A]. In: MacArthur, F., Oncins-Martinez, J., Sanchez-Garcia, M., Piquer-Piriz, A. (Eds.), *Metaphor in Use: Context, Culture, and Communication*[C]. Amsterdam: John Benjamins, 2012.

Rayson, P. From Key Words to Key Semantic Domains[J]. *International journal of corpus linguistics*, 2008(4).

Scott, M. Problems in Investigating Keyness[A]. In: Bondi, M., Scott, M. (Eds.), *Keyness in Texts*[C]. Amsterdam: John Benjamins, 2010.

Scott, M. Wordsmith Tool Manual (Version 6.0)[M]. Liverpool: Lexical Analysis Software Ltd., 2014.

Scott, M., Tribble, C. *Textual Patterns: Key Words and Corpus Analysis in Language Education*[M]. Philadelphia: John Benjamins Publishing Company, 2006.

Sinclair, J. *Trust the Text*[M]. London: Routledge, 2004.

Stubbs, M. Three Concepts of Keywords[A]. In: Bondi, M., Scott, M. (Eds.), *Keyness in Texts*[C]. Amsterdam: John Benjamins, 2010.

Thomas, S., Hawes, T. P. Reporting Verbs in Medical Journal Articles[J]. *English for Specific Purposes*, 1994(2).

Thompson, G., Ye, Y. Evaluation in the Reporting Verbs Used in Academic Papers[J]. *Applied Linguistics*, 1991(4).

Warren, M. Identifying Aboutgrams in Engineering Texts[A]. In: Bondi, M., Scott, M. (Eds.), *Keyness in Texts*[C]. Amsterdam: John Benjamins, 2010.

定稿日期:2019-12-28　　【责任编辑　谢　群】