

语料库语言学

CORPUS LINGUISTICS

1

Vol. 1 No. 1
第1卷 第1期

2014

北京外国语大学中国外语教育研究中心

Annotation
AntConc
BNC
Chunk
CLEC
Cluster

Collocation

Congram

Concordance

Context
co-occurrence
Corpora

Corpus
corpus-based
corpus-driven

Frequency

lexical item
KWIC
Wordlist
Keywords
Chunk
Node
Text
Lemma
POS
Tagging
Lexis
N-gram
Annotation
Phraseology
Idiom principle
Open Choice principle
Pattern

Units of meaning

Semantic preference
Semantic prosody
Sinclair
Data-driven learning
Meta-data
Regex
WordSmith
CLEC
SWEECH

外语教学与研究出版社

FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS



《语料库语言学》

2014年 第1卷 第1期

目 录

发 刊 词

创刊寄语

学者聚焦

语料库语言学答客问..... 桂诗春 (1)

研究论文

英语搭配框架的意义单位再探..... 何安平 (16)

语料库、平义原则和美国法律中的诉讼证据..... 梁茂成 (25)

四大名著汉英平行语料库的宏观语言特征研究..... 刘泽权、刘鼎甲 (34)

中国学者应用语言学英语论文中的词块研究..... 郑红红 (47)

研究综述

语义韵研究的理论、方法与应用..... 陆 军 (58)

基于CiteSpace的国内语料库语言学

研究概述 (1998-2013)..... 刘 霞、许家金、刘 磊 (69)

研制开发

人文社科学术文本俄汉平行语料库的创建与研究..... 陶 源 (78)

基于CiteSpace的国内语料库 语言学研究概述（1998-2013）*

北京外国语大学 刘 霞 许家金 北京外国语大学/燕山大学 刘 磊

提要：本文基于文献计量学工具CiteSpace对1998年至2013年发表于CSSCI刊物的语料库相关文献进行量化分析和可视化呈现。1998-2013年间，我国外语界语料库语言学的发展沿着两条主线展开，一是基于学习者语料库的中介语研究，二是基于双语语料库的翻译及对比研究。从时间上看，起初是以学习者语料库研究为主导，之后双语语料库研究更为受人关注，最近三年，国内语料库语言学研究出现了许多新的研究点，出现了跨学科、跨领域的特性。

关键词：语料库语言学、CiteSpace、中介语研究、翻译及对比研究

1. 引言

本文所关注的是国内较重要的语言学刊物登载语料库研究论文的情况。其数据来源于南京大学中国社会科学研究评价中心开发的《中文社会科学引文索引（CSSCI）数据库》。我们以“语料库”为关键词检索得到708条相关文献，截止2014年1月7日数据采集当日，708条文献的发表时间覆盖1998年至2013年初。显然，14余年，与语料库有关的语言学文献远不止于此。本文立足CSSCI数据库，集中考察高影响因子文献，应能一定程度上反映国内语料库语言学研究的大致面貌。诚然，CSSCI刊物之外不乏好文，或许今后可以基于中国知网（CNKI）的期刊数据再作进一步分析。为践行“量化、客观”原则，我们借助文献计量学分析工具CiteSpace（Chen 2006, 2012；冯佳等 2014），对CSSCI中检索得到的语料库语言学文献数据做了可视化的呈现。

2. 国内语料库语言学研究文献的数据分析

2.1 1998-2013年国内语料库语言学研究总体趋势

图1所示为1998年至2013年CSSCI来源期刊登载语料库相关论文数量的分布走势。其逐年递增的总体趋势，一目了然。在数据所覆盖的14余年中，自2003年开始，似乎出现了明显的增长势头，到2010年到达顶峰，截止数据采集当日，该数据库仅收集了2013年

* 本研究得到国家社科基金项目“基于双语语料库的汉语复杂动词结构英译研究”（12CYY060）和教育部“新世纪优秀人才支持计划”（NCET-12-0790）的资助。梁茂成教授、李文中教授、熊文新副教授对本文初稿提出了细致中肯的意见和建议，特致谢忱。

第1、2期的文献。由于文献数量不全导致了2013年数量的减少。

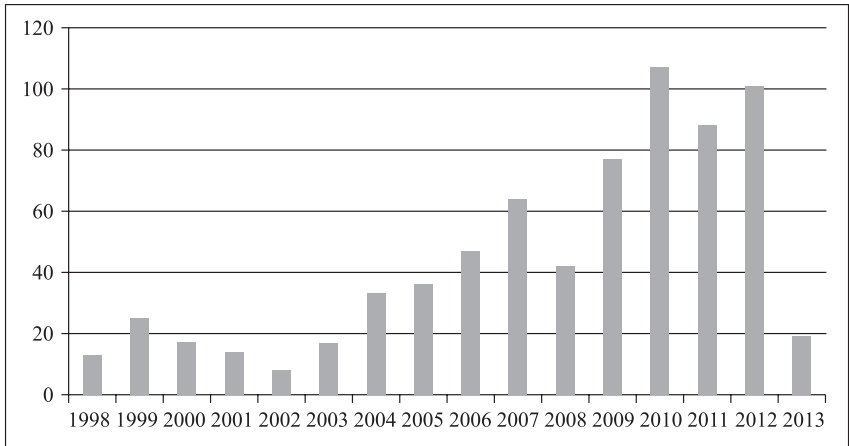


图1. 国内语料库语言学研究文献的时间分布

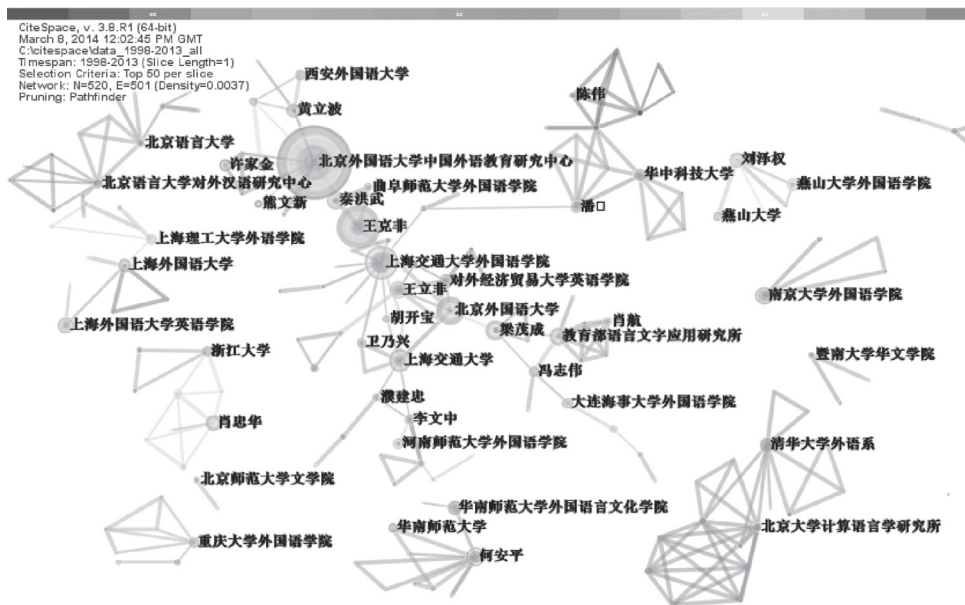
708篇语料库相关研究分布于不同的学科。其中最多的是语言学（586篇），其次是教育学（50篇）、图书情报学（29篇）以及新闻学与传播学（10篇）。语料库在语言学研究中的应用从1998年至今日益增多，但在其他领域的应用在近三四年间才得到推广。本研究的数据显示，教育学和图书情报学领域的第一篇语料库相关研究发表于2002年，新闻与传播学领域更晚，直到2007年才出现第一篇。但从2010年到2013年初，这三个领域应用语料库的研究几乎占据了14余年的一半，如教育学（24/50），图书情报学（13/29），新闻学与传播学（6/10）。由此可见，语料库已经越来越多地被其他领域的研究者掌握并应用到各自的研究中。

国内CSSCI期刊中，登载语料库语言学文章最多的前10个刊物及其刊文数量为：《外语电化教学》（双月）93篇，《外语教学与研究》（双月）57篇，《现代外语》（季刊）39篇，《语言文字应用》（季刊）38篇，《当代语言学》（季刊）32篇，《外语界》（双月）32篇，《外语学刊》27篇，《外语与外语教学》（双月）27篇，《外国语》（双月刊）25篇，《中国外语》（双月）23篇。语言学尤其是外国语言学类CSSCI刊物似乎都愿意接受语料库语言学研究稿件，这一定程度上也可看出各家刊物对语料库语言学这种新式研究方法的认可度。

对国内语料库语言学研究的总体趋势有了一个大致了解之后，我们将借助CiteSpace这一科学计量学的方法，进一步考察语料库语言学研究的基本情况，试图探测其发展趋势或动向，并以可视化的方式加以呈现。根据其节点类型，CiteSpace可以呈现四类可视化图谱，第一类是作者、研究机构、国别；第二类是参引文献（cited reference）之间以及被引作者（cited author）之间的被引关系；第三类是关键词和名词性术语；第四类是研究基金。本文将基于1998到2013年初的数据，分别呈现并讨论前三类可视化图谱。

2.2 国内语料库语言学研究的主要研究单位及学者

图2中的圆圈(即节点)代表研究单位及学者,其大小代表该机构或学者发表的文章数量,数量越多,节点越大。节点由不同颜色的年轮构成,每一年轮对应文章的出版时间,由内到外,年轮对应的时间由远及近。节点间的连线代表作者与作者以及作者与机构之间的联系。为了让图谱中的文字更加清晰,笔者设定阈值为4,即图2中有文字标注的节点,均是发表文章大于等于4的学者和机构。

图2. 国内语料库语言学的主要研究单位及学者¹

1998—2013年间,国内有一些产出语料库语言学研究成果较多的单位或机构,比如,北京外国语大学(及北京外国语大学中国外语教育研究中心)、上海交通大学(及上海交通大学外国语学院)、华南师范大学外国语言文化学院、上海外国语大学、南京大学外国语学院、对外经济贸易大学英语学院、清华大学外语系、华中科技大学、河南师范大学外国语学院、燕山大学外国语学院等。以上基本是外语院、校,它们从事的主要是基于语料库的英语研究。在开展语料库研究的单位中,也有一些以中文为主要研究对象的机构,如教育部语言文字应用研究所、北京语言大学对外汉语研究中心、华中师范大学语言与语言教育研究中心等。有关汉语学界基于语料库开展的研究²可参看Feng(2006)。

显然,图2提供的信息勾勒出了国内从事语料库语言学研究的机构。这些机构往往由一些核心成员形成研究团队,从而能持续产出有影响的研究成果。语料库语言学研究的本质离不开团队协作。团队的存在是语料库建设与研究成果产生的重要平台。国际范围内,兰卡斯特大学、伯明翰大学、伦敦大学学院、比利时鲁汶天主教大学、北亚利桑那大学等无不是学者聚集的团队。只有形成团队,构建学术交流机制,才能不断产生更新更好

的语料库产品和学术成果。

2.3 语料库语言学的知识结构

“一篇文献的被引频次可以在一定程度上反映该文献的影响度”（刘则渊等 2008: 143）。理清国内语料库语言学具有高被引频次的文献，以及他们的被引用情况，能够帮助我们廓清该领域的知识结构，并通过找到文献的被引激增以及转折点，发现该领域的研究动向。图3呈现了国内语料库语言学研究的共被引图谱，并标注出了国内学者引用最多的前11条文献³。根据节点大小，我们不难看出，被引频次从高到低依次为：Sinclair（1991）、杨惠中（2002）、Hunston（2002）、桂诗春、杨惠中（2003）、王克非等（2004）、Baker（1993）、文秋芳等（2003）、卫乃兴（2005）、黄昌宁（2002）、李文中、濮建忠（2001）、何安平（2004）。这些高频次被引文献可分为两大类：一类是关于语料库的基础知识，如Sinclair（1991）、杨惠中（2002）、黄昌宁（2002）；另一类是关于语料库的应用，在这一类中又可分为两类，一是语料库如何应用到教学和学习中，如Hunston（2002）、桂诗春、杨惠中（2003）、文秋芳等（2003）、卫乃兴（2005）、李文中、濮建忠（2001）、何安平（2004），另一类是语料库如何应用到翻译研究中，如王克非等（2004）、Baker（1993）。

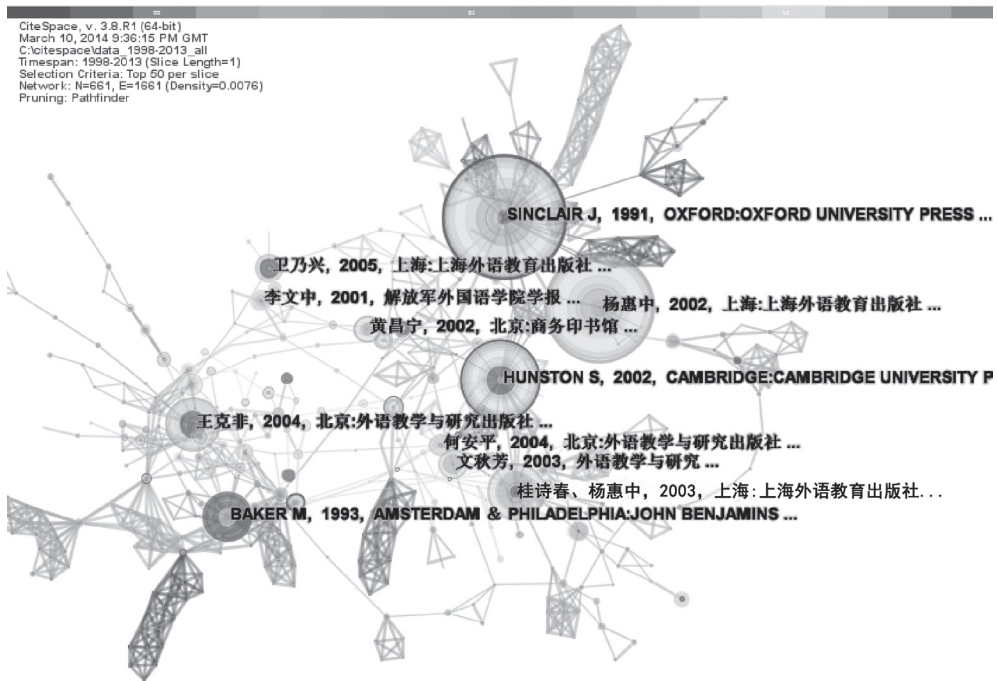


图3. 国内语料库语言学研究的共被引文献⁴

任何学科发展过程中都会经历一些具有重要意义的转折点，语料库语言学也不例外。CiteSpace通过计算每一个节点的中介中心度（betweenness centrality），继而标注出整个网

络图谱中用以连接两个子网络的节点，即转折点。表1列举了国内语料库语言学的十个转折点，并按照其中介中心度由高到低排列。

表1. 国内语料库语言学的转折点文献

	文献名称	作者	出版年代	中介中心度	总被引频次
1	Corpus, Concordance, Collocation	Sinclair	1991	0.41	67
2	Corpora in Applied Linguistics	Hunston	2002	0.33	43
3	Corpus-based Translation Studies: The Challenges That Lie Ahead?	Baker	1996	0.23	10
4	Corpus and text: Basic principles	Sinclair	2005	0.21	2
5	Corpus linguistics and translation studies: Implications and applications	Baker	1993	0.17	5
6	Corpus Linguistics	McEnery & Wilson	2001	0.16	12
7	If you look at...Lexical bundles in university teaching and textbooks	Biber <i>et al.</i>	2004	0.15	4
8	Prolegomenon to a theory of translation	Frawley	1984	0.15	2
9	The Role of corpora in investigating the linguistic behaviour of professional translators	Baker	1999	0.11	6
10	Dimensions of Register Variation: A Cross-linguistic Comparison	Biber	1995	0.11	2

此外，图3的知识图谱中还有一些在1998–2013年之间被引激增的文献。“通过对被引激增文献的考察，我们可以追踪某一学科和研究领域的热点及其历时演变”（Chen 2012: 597）。在本研究的数据中，我们探测到12个被引激增的文献，如图4所示，他们均是在过去13余年间语料库语言学研究的热点。最近三年出现引用激增的文献有Baker（1993）、Baker（2000）以及Tognini-Bonelli（2001）。前两条文献足以说明将语料库应用到翻译研究是当下的研究热点，后者则体现了近期语料库语言学领域中关于两种研究范式的争论，正是Tognini-Bonelli（2001）首次对基于语料库和语料库驱动的研究范式进行了区分。通过进一步观察这三条文献的施引文献（citing reference），甚至进一步锁定引文内容，便可证实我们的猜测。但如果想要更加直观地了解语料库语言学的研究热点，还需要通过节点类型为关键词（keyword）的知识图谱来呈现。

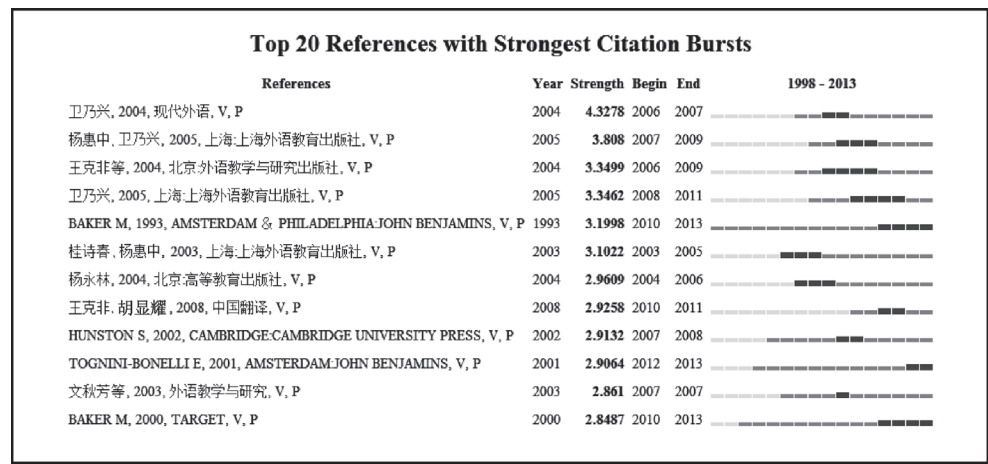


图4. 国内语料库语言学研究被引激增的文献

2.4 国内语料库语言学研究热点

关键词是每一篇文章核心内容的浓缩，如果某一关键词在该领域的文献中重复出现，就可以被视为研究热点，再将它们与所出版的时间相联系，就能发现某个领域在特定时期的研究热点。

图5直观地列出了国内语料库语言学研究的部分关键词，如果将出现频率最高的前50个关键词进行合并和分类，我们可以大致将国内语料库语言学的研究热点分为三个研究取向：中介语研究（如“外语教学”、“学习者语料库”等）、翻译研究（如“平行语料库”、“翻译共性”等）和语料库语言学的研究议题（如“语义韵”、“搭配”等），前两者的出现频率远远高于后者，而且在相当长时期内，中介语研究和翻译研究都会是我国外语界语料库语言学研究两个主要方面。

图5中的信息还可以按年份来观察。图中左上方以翻译为主的节点（“平行语料库”、“语料库翻译学”等），揭示出近期国内语料库研究的主要关切点，而右下部分的关键词（“英语教学”、“学习者语料库”等）表明这些研究点势头趋缓。研究热点的转变反映了我国外语界语料库语言研究的两大方面，即学习者语料库研究和双语语料库研究。学习者语料库研究的巅峰时期在过去的5-8年以前，最近五年左右，双语语料库成为国内语料库语言学研究的一个新的热点。另外，语料库语言学在中国发展至今，有关词块、搭配和意义单位的探讨在1998年至2011年期间一直为大家所关注，这与国际上语料库语言学研究趋势基本吻合，即有关短语学（phraseology）的探究始终是热门话题。其中有关意义单位的讨论出现更晚一些，到最近几年才陆续产出了一批成果。

图5. 国内语料库语言学研究的关键词⁵

还有一些最近三年新出现的关键词，由于出现频率少，节点不易显现，如“口译语料库”、“汉语中介语语料库”、“web语料库”、“社会语言学”、“中国立法语言语料库”、“中文文本情感分析”、“汉语方言”、“手语语料库”、“文献计量”等，但这些关键词更能体现目前语料库语言学的研究趋势。我们不难看出，语料库逐渐成为了各研究领域的工具，越来越多的学者带着各自的理论视角投入基于语料库的研究当中，因此语料库语言学领域产生了许多新的研究点。

3. 结语

本文通过分析基于CiteSpace获得的语料库语言学研究机构和学者图谱、共被引图谱以及关键词图谱，发现1998-2013年间，我国外语界语料库语言学的发展沿着两条主线展开，一是基于学习者语料库的中介语研究，二是基于双语语料库的翻译及对比研究。从时间上看，起初是以学习者语料库研究为主导，然后双语语料库研究逐渐受到关注，最近三年，国内语料库语言学研究出现了许多新的研究点，出现了跨学科、跨领域的特性，如与文献计量学、社会语言学、汉语语言学的结合。当前国际语料库语言学研究也呈现出一些新的动向，如语料库分析与话语分析、语用分析、功能语言学研究，以及认知语言学研究结合起来，使语料库语言学的边界得到了较大拓展（参看McEnery & Hardie 2012）。我国语料库语言学研究者也应跳出学科的藩篱，重视从领域之外汲取营养，才能发掘出更多有价值的研究选题，作出更深入的研究。

本文对1998–2013年间国内语料库语言研究所作的综述，主要基于量化数据和计算机自动分析，虽然有其自身的优势，但同时也有不足之处。一方面，CSSCI数据库本身存在一定的缺陷，所以基于该数据库的分析需要注意以下几点：第一，该数据库并未收录所有的好文，不同时间的检索结果并不一致；第二，某些文献的发表年代有误；第三，同一文献的引用格式不同（如作者的姓名和出版机构的拼写格式），这有可能影响网络节点之间的连线。对以上问题，我们都进行了详细的手工排查和批量处理，但仍不能保证没有错误。另一方面，我们需要清醒地认识到，学术评价不是单凭文献引用率可以最终裁判的。文献的高频引用更多反映的是学术关注度，而不是学术品质本身。譬如说，我们不能简单地认为，从文献引用率看，学习者语料库热度已退，因此我们就放弃了对学习者中介语的探究。相反，我们更应认真思考，如何突破中介语对比分析法的局限，探索出新的研究增长点。事实上，中介语研究仍然大有可为。同理，我们不能简单地认为某位学者被引用率高，其研究的水准一定高。文献引用有其特点，比方说，某领域的综述类文章远比某专题的实证研究更易被引用，出版早的文献较新文献引用率会偏高。

概而言之，基于文献引用的综述让我们可以更全面地了解“普罗大众”的学术关注，使我们在开展研究时，不至于闭目塞听、闭门造车。对于学术观点和研究价值的判定，新的研究选题的发掘，则在于多读、多问、多思、多行。

注释

1. 图2以一年为一个时间分区（time slice），每一时间分区内提取被引频次最高的前50篇文献的共被引数据，采用寻径算法（pathfinder），对单个时间切片和复合时间切片进行剪裁后生成知识图谱。

2. 本文基于CSSCI期刊的数据分析，未对汉语语言学和外语语言学做学科区分。从发文数量看，近10年左右国内外语界明显多于汉语界。然而，汉语界有一些重要团队、重要学者做出了许多出色的语料库研究成果。譬如，国家语言资源监测与研究中心的一些分中心：北京语言大学纸媒语料库杨尔弘教授团队、中国传媒大学有声媒体语料库侯敏教授团队、暨南大学海外华语语料库郭熙教授团队、厦门大学教材语料苏新春教授团队、华中师范大学网络媒体何婷婷教授团队、中央民族大学的民族语言赵小兵教授团队，以及南京师范大学董志翹教授、陈小荷教授的国家社科重大招标项目古汉语语料库等。更早的还包括陈小荷教授20世纪90年代初在北京语言大学开展的“八五”教委规划项目用于留学生汉语教学的“汉语中介语语料库系统”。中文方面理论研究包括张普教授的动态流通语料库理论以及在此基础上的语言资源监测，此外还有服务对外汉语教学的北京大学袁毓林教授团队的国家社科重大项目等。

3. 阈值选为14，即被引频次大于等于14的文献节点才出现标注文字。

4. 图3以一年为一个时间分区，每一时间分区内提取被引频次最高的前50篇文献的共被引数据，采用寻径算法，对复合时间切片进行剪裁后生成知识图谱。

5. 图5以一年为一个时间分区,每一时间分区内提取被引频次最高的前50篇文献共被引数据,采用最小生成树(minimum spanning tree),对复合时间切片进行剪裁后生成知识图谱。另外,为保证其他节点更加清晰,我们隐去了图谱中“语料库”这个最大节点。

参考文献

- Baker, M. 1993. Corpus linguistics and translation studies: Implications and applications [A]. In M. Baker, G. Francis & E. Tognini-Bonelli (eds.). *Text and Technology: In Honour of John Sinclair* [C]. Philadelphia: John Benjamins. 233-250.
- Chen, C. 2006. CiteSpace: Detecting and visualizing emerging trends and transient patterns in scientific literature [J]. *Journal of the American Society for Information Science and Technology* 57(3): 359-377.
- Chen, C. 2012. Predictive effects of structural variation on citation counts [J]. *Journal of the American Society for Information Science and Technology* 63(3): 431-49.
- Feng, Z. 2006. Evolution and present situation of corpus research in China [J]. *International Journal of Corpus Linguistics* 11(2): 173-207.
- Hunston, S. 2002. *Corpora in Applied Linguistics* [M]. Cambridge: CUP.
- McEnery, T. & A. Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice* [M]. Cambridge: CUP.
- Sinclair J. 1991. *Corpus, Concordance, Collocation* [M]. Oxford: OUP.
- Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work* [M]. Amsterdam: John Benjamins.
- 冯佳、王克非、刘霞, 2014, 近二十年国际翻译学研究动态的科学知识图谱分析 [J], 《外语电化教学》(1): 11-20。
- 桂诗春、杨惠中, 2003, 《中国学习者英语语料库》[M]。上海: 上海外语教育出版社。
- 何安平, 2004, 《语料库语言学与英语教学》[M]。北京: 外语教学与研究出版社。
- 黄昌宁, 2002, 《语料库语言学》[M]。北京: 商务印书馆。
- 李文中、濮建忠, 2001, 语料库索引在外语教学中的应用 [J], 《解放军外国语学院学报》(2): 20-25。
- 刘则渊、陈悦、侯海燕等, 2008, 《科学知识图谱方法与应用》[M]。北京: 人民出版社。
- 王克非等, 2004, 《双语对应语料库: 研制与应用》[M]。北京: 外语教学与研究出版社。
- 卫乃兴, 2005, 《语料库应用研究》[M]。上海: 上海外语教育出版社。
- 文秋芳、丁言仁、王文字, 2003, 中国大学生英语书面语中的口语化倾向——高水平英语学习者语料对比分析 [J], 《外语教学与研究》(4): 268-274。
- 杨惠中, 2002, 《语料库语言学导论》[M]。上海: 上海外语教育出版社。

通信地址: 100089 北京市北京外国语大学中国外语教育研究中心(刘霞、许家金)
100089 北京市北京外国语大学中国外语教育研究中心/066004 河北省秦皇岛市燕山大学外国语学院(刘磊)

Research on semantic prosody: Theory, methodology and application

..... LU Jun (58)

This paper reviews the development of the theoretical framework, research methodology and application areas of semantic prosody. It begins with a brief description of multiple schools of thoughts on defining semantic prosody, followed by an analysis of their contribution to linguistic studies and the relationship among them. Then, research methods are reviewed from the perspectives of their theoretical basis, potential improvement and future development. The third part deals with the application areas of semantic prosody in terms of their principle, theoretical contribution and future research direction. Finally, it is concluded that a basic theoretical framework and research methodology for semantic prosody research have been established, which has a strong demand on further extended research.

An overview of corpus linguistics research in China (1998–2013): A CiteSpace based analysis

..... LIU Xia, XU Jiajin & LIU Lei (69)

This article reviews corpus based language studies drawing on the CSSCI indexed journal articles between 1998 and 2013. The bibliographic data have been quantified and visualized with the bibliometric tool CiteSpace. The results show that corpus based language studies have been carried out along major lines of research: corpus based interlanguage analysis and bilingual corpus based translation studies and contrastive linguistics over the 16 years span. During the earlier years in the time span of the review, focus of attention was devoted to learner corpus research, while bilingual corpus based studies have become more popular in recent years. The cross-fertilization of different disciplines and research fields highlights the corpus research in last three years.

Construction of and research on parallel Russian-Chinese corpora: With special reference to academic texts of social sciences and humanities

..... TAO Yuan (78)

Parallel corpora construction is a prerequisite for descriptive translation studies. Although there are a variety of such corpora, bilingual or multilingual, both at home and abroad, there are few