从词语到话语：通过语料库开展话语研究*

许家金
（北京外国语大学 中国外语教育研究中心，北京 100089，北京）

提　要：本文首先简述了话语的几个核心特征：话语连贯、语境特征和互动性，兼及这些特征的跨语体变异和话语的主观情态与社会性。接着本文着重介绍了如何利用语料库方法围绕前述五个方面开展话语研究。文章结合实例详细说明了语料库方法应用于话语研究的优势和适用范围。

关键词：话语特征、话语研究、语料库方法

中图分类号：H030　　文献标识码：A

1. 引子

目前语料库技术主要以词汇索引、词表、搭配、主题词等为主要分析手段，相关研究也主要停留于词汇和浅层句法。然而，若要对语言充分描写，需要超越词汇、短语、句子，对话语层面的语言表现有所作为。过去 5 年多，国内外涌现了一批基于语料库的话语研究成果。这一点可从最近几次语料库语言学大会的主题及专题讨论得到印证。比如 2004 年美国应用语料库语言学研究会（The American Association for Applied Corpus Linguistics，AAACL）主办的第五届北美语料库语言学研讨会（The Fifth North American Symposium on Corpus Linguistics）会后出版的论文集名称为“Corpus Linguistics beyond the Word: Corpus Research from Phrase to Discourse”（2007），旗帜鲜明地指出了语料库研究需要拓展至话语层面。2005 年 26 届 ICAME（International Computer Archive of Modern and Medieval English）和第六届 AAACL 联席会议，2009 年 American Association for Corpus Linguistics （AACL）等会议发言中都有大量基于语料库的话语语用研究。2008 年 ICAME 29 更是将大会主题明确定为 Corpora: Pragmatics and Discourse。同一时期，国际间出版的基于语料库的期刊论文、专著、论文集，着眼于话语研究的也大幅增加。可见，经过数十年发展，学者们已不满足于词汇和短语层面的语料库研究，希望基于此前词汇研究的成功经验，开展更多话语语用方面的研究，使语料库方法在语言研究中发挥更为重要的作用。以下将综合文献和我们自己开展的研究，简要介绍基于语料库的话语研究的主要选题、研究设计和方法，并结合案例加以说明。

2. 话语与话语特征

2.1 话语及话语研究

介绍基于语料库的话语研究之前，先简述一下我们对话语及话语特征的认识。一方面，这有助于后续内容的展开，更重要的是，话语研究具有浓厚的跨学科性质。其理论源流涉及诸多学科，因此话语分析流派众多，各家对话语研究的认识不尽相同。这里我们采用 Schiffrin, Tannen & Hamilton（2001：1）《话语分析手册》导言中对以往话语研究的分类：1）有关超句单位的研究；2）有关语言使用的研究；以及 3）有关与语言和交际相关联的社会实践行为及意识形态方面的考察。三类研究中，第一类研究的关注点是话语的结构特征。简单说就话语所关涉的长度和跨度，即话语不限于一词一句，更强调词句间的连缀协同，从而完成意义的连贯表达。第二类强调话语研究考察的是自然发生的、真实交际中的口语和书面语。在有关语言使用这一点上，话语研究与语用研究有共同的研究焦点，它们都关注言语行为、指称、情态表达等，有时两者不易截然分开。第三类研究取向是话语研究相对较新的发展。它

强调话语应在更广的社会文化情境中解读才够充分完整。另一方面，一些学者也主张通过话语来探究社会文化的现状及其变化发展，甚至寄希望于改变不合理的社会现状。后一种即所谓的“批评话语分析”。以下将在这几类研究基础上，归纳话语所具备的一些核心特征。

2.2 话语特征

话语研究中通常比较注重考察话语的衔接与连贯、语境特征和互动性等核心特征。

衔接主要指话语中有助于意义传达的词汇语法手段。英语中的衔接手段较有影响的是 Halliday & Hasan（1976）归纳的 5 种常见类型，即指称形式、替代、省略、连接和词汇衔接。限于篇幅，这里不一一解说。对于 Halliday & Hasan（1976）有关衔接手段的概括和分类，后来学者做了厘清和扩展。但应该说，英语衔接手段在词汇语法层面基本上还是脱不了 Halliday & Hasan（1976）五分法的框架。各种衔接手段的运用使得话语能够意义连贯。连贯是话语的核心概念之一。任何话语都应具备这一特性。当然，连贯的实现需要借助各种语言、非语言手段，包括语境特征和互动性特征等的共同作用。

语境特征也是话语研究中不可忽视的基本特征。Malinowski（1923：310-311）将社会语境分为情境语境（situational context）和文化语境（cultural context）。而在语言分析中，这一语境二分法还要包括语言单位间的上下文语境，更大的语言单位（段落与段落、前后篇章、不同文本等）之间的互文性（intertextuality），以及文本之外的社会文化语境等。

另外，互动性也是话语的一个重要特征。互动性在口头会话中体现得较明显，因为会话需要双方（或多方）的参与、合作才能顺利推进，完成交际意图。书面话语也存在互动性。作者总因特定读者而写，或描述，或倾诉，或论证，总希望将信息或想法传递给读者。在实际话语中，互动性也必然通过一定的语言形式得以体现。

话语特征众多，其中衔接连贯、语境特征和互动性最为不可或缺。其他话语特征多是从这几个特征分化衍生而来。下节将通过实例阐释如何借助语料库方法对上述话语特征加以研究。

3. 语料库在话语研究中的应用

语料库同话语研究可以兼容有其内容和方法上的合理性。它们都关注使用中的真实语言，语料库研究和话语研究都兼涉语言内容（linguistic data）和分析方法（analytical tool）两层含义。文献中常用 synergy（联合）这个词来形容这两个兼容领域的联姻。

3.1 基于语料库的话语研究方法概要

从前文对话语及其特征的梳理来看，话语连贯、语境特征和互动性都会通过一定词汇语法特征体现出来，因而可以通过语料库技术提取和分析相关的词汇语法特征开展话语研究。

在介绍基于语料库的话语研究方法之前，先简单分析一下基于语料库的话语研究同传统话语分析的主要区别。

传统话语分析方法，多基于单篇或少量文本，针对某个或某些主题做细致解读和逐个分析。这种分析往往细致、深入、透彻，但无法得出适用性更广的结论。比如，Halliday（1981）有关尼安德特人话语的及物性分析，论证十分有力，但所得结论只适用于 Golding 的小说，或者说只能说明 Inheritors 那一篇小说。其他话语分析论著所附的示例（如 Halliday & Hasan 1976：340-355；Hoey 1991：246-257；Gee 1999：119-148 等）也都只是针对片段语例的分析。

而基于语料库的话语研究有两方面优势：一是语料库作为数据库，容纳了大量的（往往具代表性的）某类话语的集合。从中发现的语言事实、话语特征相对于少量文本，更能推而广之；其次，语料库分析方法中有其相对成熟的语言特征提取和统计方法，可以帮助发现和回答话语现象之间的关联、差异等问题。而按 Baker（2006：10-14）的归纳，基于语料库的量化分析为主的方法可减少研究者的主观偏见和发现大量话语中呈现出的累积效应

（incremental effect of discourse，即通过大量文本发现共性话语现象）。另外，Baker（2006：14-15）还提到，根据所掌握的语料不同，基于语料库的话语研究除了发现大量语料中稳定的共性特征外，会发现一些异例或反例（resistant）的存在，以及随着时代发展体现出的话语变化；同时，语料库方法还可作为传统话语研究方法的补充，语料库方法（特别是语料库驱动方法）还十分有助于发现新的话语现象和研究课题。

3.2 基于语料库的话语研究及常见选题

一般认为基于语料库的话语研究属定量研究范畴。事实上，完整的、合理的基于语料库的话语研究都要包含研究者定性的、阐释性的概括及分析。可以说基于语料库的话语研究是定量定性相结合的研究方法。

Biber, Connor & Upton（2007：12-14）将基于语料库的话语研究分为自上而下型和自下而上型两种。前者从一定的理论前提出发，然后借助语料库方法，从大量语言事实中获取语言实例，经分类、概括等，通过量化数据对语言进行充分描写并回答研究问题；后者没有理论预设，完全从文本数据中自动挖掘语言事实，词表、主题词表、词语搭配等属于这一范畴。而从利用语料和对量化数据的依赖程度来看，基于语料库的话语分析可分为基于语料库的（corpus-based）和语料库驱动的（corpus-driven）研究取向两种。

语料的使用上也可分为两大类。一种是从特定选题出发，针对全球化、全球变暖、金融危机、法律诉讼话语、学术口语、学术论文、英语教材等等，收集大量的相关语料，这些称为专门用途语料库（specialized corpora）。第二种是利用现成的通用型语料库，比如反映英国英语一般特点的“英国国家语料库”（British National Corpus，BNC）和反映中国学生英语状况的一些学习者语料库。语料的选择由研究目的决定。有的研究需要对语言的一般状况加以描写，则需要从整个语料库中提取数据；为回答特定领域语言现象，则可以利用专门用途语料，或从通用语料库中抽取部分语料构成子库。

表 1 话语的典型特征及语料库研究思路

话语特征	分析思路	语料库技术	研究选题举例
衔接连贯	关注词汇语法衔接手段，某个（些）词项在多个文本中的分布	词项的单独、批量检索	所有名词；所有地点名词；情态动词；篇章连接词与作文成绩相关研究
语境特征	点——线——面方式扩展语境	词项检索、主题词、搭配、框合结构等，结合语体信息等元信息	按性别、国别等开展的各种词汇、短语、语法结构分析
互 动 性	关注互动词汇、短语	互动性词汇、短语批量检索、批量词块提取再筛选、主题词分析后分类再筛选等	停顿、修补、反馈语、书面语作者/读者能见度、口语化、互动词块

除了话语的三个核心特征，语料库方法还特别适合分析这些特征在不同语言变体间的异同，以及不同时间维度上的变化。这一方面，比较有代表性的有语域变异（Biber 1988，2006；Biber *et al*. 1999）、跨学科学术话语对比（Hyland 2004，2006）、语料库文体学（Semino & Short 2004）和中介语对比分析（Granger 1998；Granger *et al*. 2002）等。

另外，如 2.1 节所述，实际语言交际中，说话人通过衔接连贯、语境特征和互动方式，旨在传达一定的主观认识（subjectivity），或展现某种社会意识形态。因此，话语特征中蕴

含的说话人的主观性（如有关评价、情态、立场的研究）和社会性（身份认同、性别差异、意识形态）也是话语的重要方面。语域变异和主观性研究以往更多见于社会语言学文献。近一二十年，话语研究与社会语言学界限已不那么清晰。

以上为了论述方便，将基于语料库领域的相关课题做了上述分类。实际中各层面往往交织在一起。

3.3 研究案例

衔接连贯

以英语单词 something 为例，在 http://corpus.byu.edu/bnc/中检索 something [aj*]，可以得到表 2 中的结果，限于篇幅，这里列出排前 5 的形容词。我们发现 something 这个极普通的英文单词，后接的形容词常常是表示令人关注的事。要么是“出了什么问题（wrong）”，要么是“有新发现（new）”，要么是“有一些特异的事发生（different、special）”等。可见，对一个中性的平常的词，通过语料库也可能发现一些隐藏其背后的主观性。Hunston（2002: 62）将这种通过一些固定常见词语获得语言背后的话语信息的做法称为“探针法”（using probes）。比如，她用 something ADJ about him/her 去语料库中检索，获得了人们在谈及男性和女性时一般用什么样的形容词。

表 2 something 后接的形容词

something	wrong	501
something	new	383
something	different	333
something	special	253
something	similar	194

在表 2 基础上，还可以通过 http://corpus.byu.edu/bnc/的 chart 功能，获得 something [aj*]的语体分布（如图 1），something ADJ 的表达多见于口语体和与口语体接近的小说语体。

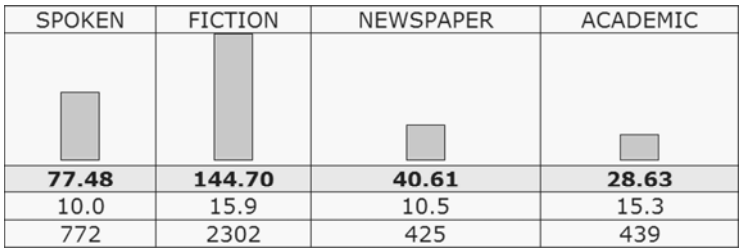

图 1 something ADJ 的语体分布

更有趣的是，进一步分析得出，学术（ACADEMIC）语体中，something ADJ 出现最多的是 something new，计 41 次。这正符合学术研究的宗旨——探索新知。可见，即便通过单个词语的跨文本分析，也可以得出其话语含义。这里突出体现了学术文本的语体特征。某种意义上说，特定语体的存在是众多相关词语和结构共同作用的结果，something ADJ 恰好是其中之一。那么，在连续话语内部，众多词项是如何共同作用从而保证意义的连贯表达的呢？这是话语研究中的一个常见的基本问题。下面我们以研究较多的衔接手段为例，简述如何通过语料库方法观察话语连贯。

这里介绍两种方法，一种是类似何安平、徐曼菲（2003）有关口语小品词的研究，该研究以 19 个词项（just，like，okay，oh，right，well，I know，I mean，sort of，kind of 等）

作为检索项。这些词语是口语中经常用作起承转合的衔接手段。通过批量检索（file-based concordancing）的方式，得到 19 个词项在语料库中的分布和频数。在此基础上，按出现位置（话轮的前、中、后）和话语功能（接续、转移话题、修补等）分别讨论。作为学习者语料库研究，还可以对比中国学生和英美学生，以及不同母语背景英语学习者在上述结构和功能方面的异同。

第二种方法是考察衔接词项同外部特征的相关性。比如，可以对表达话语连贯的关连词（如 but，therefore，in fact，on the other hand 等）与学生英语作文的成绩做相关分析。我们曾用 82 个类似的关连词对 120 篇中国学生作文进行批量检索，并与这 120 篇作文的成绩做相关分析，得到相关系数为 0.21，即弱的正相关。如果这个结论在更大范围的语料中得到证实的话，将是一个有意思的发现。0.21 的弱相关说明，英语老师们经常鼓励学生们使用的关连词，未必对整篇文章的连贯有显著作用。换言之，单有形式上的连贯是不够的，文章内容上的连贯也是十分重要的，甚至是核心的。

语境特征

上节谈到如何从单个或多个词项出发，分析其频数、分布及话语功能。这类研究的观察点可以是 something，fact，happy 等抽象词语，也可以是相对确定的词，如 China，Jewish，知识分子等，还可以是关连词、（元）话语标记、情态表达等某一类词，甚至可以是某一个语法构式，如“名词 and 名词”及被动语态等。然而只关注这些词汇语法单位的频率和分布是不够的。按照话语研究的思路，我们需要结合语境来综合认识相关话语的特征。2.2 节介绍过，语境可以是局部上下文，语言单位（段落与段落、前后篇章、不同文本等）间的互文性，以及文本之外的社会文化语境。

结合语料库语言学的方法，这里提出“点-线-面语境扩展分析法”。分析始于词项的检索，检索项可以是单个或多个单词、短语、句法结构，这是所谓的“点”；进而，可以在索引行里分析上下文特点，主要做法是观察搭配、类联接、语义倾向和语义韵，这是所谓的“线”；最后，利用语料库提供的元信息和其他文内标记，可以对检索项在文本中出现的相对位置（如可利用索引词图，concordance plot 功能），分析检索项的来源或原始说话人的身份和态度等主观性特征和其他社会语言学特性（如性别、社会阶层、说话场景等）。点、线加上外部语境，共同构成话语分析的面。

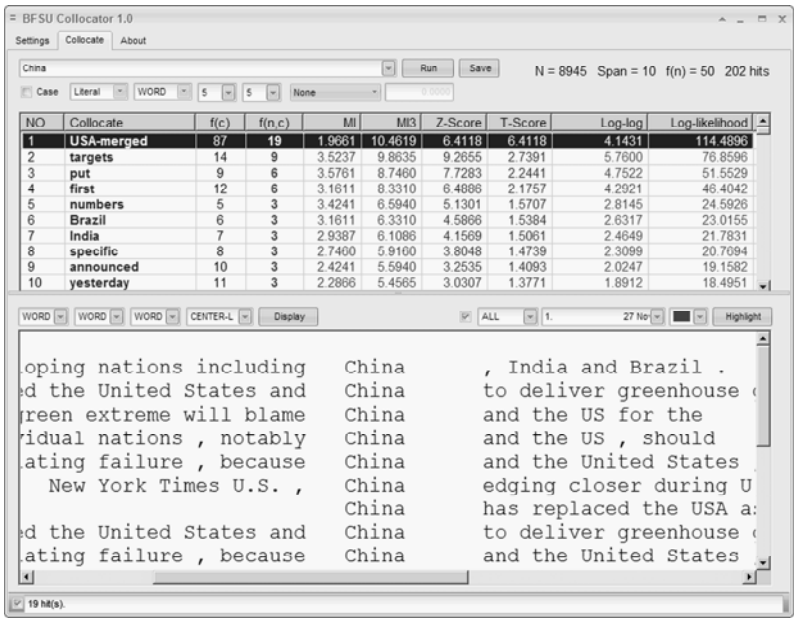

图 2 China 在哥本哈根气候大会语料中的搭配词分析

图 2 中的例子为有关中国在哥本哈根气候大会上的角色分析，可以算作媒体话语中的中国形象研究的一个简单案例。所用语料为 25 篇哥本哈根气候大会的相关新闻报导。这里以单词 China 为检索词，借助 BFSU Collocator 软件，计算出 China 左右跨距各 5 个词范围内的强搭配词。按对数似然比（log likelihood）的搭配力指标排序，得到 USA（这里将 US，U.S.，USA，U.S.A.，States 等合并为一个词项，因此命名为 USA-merged；同时在操作时标点和介词等虚词被删除），targets，put，first，numbers，Brazil，India，specific，announced，yesterday 等核心语境共现词语。通过这些与 China 紧密共现的词语，我们可以推知，中国和美国在大会期间常被相提并论，因为两国是最大的碳排放国。这次会议上，媒体所关注的与中国最相关的议题是中国作为与美国以及印度（India）和巴西（Brazil）等发展中国家和新兴经济体对减排所需承担的责任。而会议期间，最热门的议题之一便是前一日（yesterday）中国第一次（first）宣布（announced，put）的具体（specific）减排指标（targets，numbers）。排除背景知识不论，上述结论的得出，主要依据的是 China 这个词的语境共现词。为获得对上下文的准确认识，双击第二列 Collocate 的某个搭配词，在下方窗口即可得到索引行信息。进而，在每一个索引行的右侧会显示该行所在的新闻文本。在对 China 的词语搭配的同时，BFSU Collocator 还支持对检索词的类联接的考察。在对词语搭配的深入分析时，可以对搭配词语按语义归类，分析其语义倾向和语义韵。

如需要，我们可以按国别对新闻机构加以分类，从而获得诸如西方媒体，中国媒体，或发展中国家媒体对中国气候大会报导中对中国形象的建构。比如我们可以发现中国的英文媒体会强调中国是一个负责任的大国，会按自己的评估制定务实的减排目标；而西方媒体会指责、质疑甚至妖魔化中国在减排中所扮演的角色。

同理，上述分析方法还可用于分析奥运、中国经济、金融危机、四川及海地地震、农民工、春运等话题的相关研究中（可参看许家金、赖辉 2009；许家金 2009a，2009b 有关知识分子、小沈阳和施惟可（SWECCL）的话语身份建构研究）。

互动性

互动性之所以重要，是因为它指向了一个基本事实，即话语都是有听众和读者的。因而，在进行话语分析，无论是独白式的，还是对话式的，都不能将话语互动的另一方忽略。话语双方的互动总会通过特定的词汇语法形式体现。因此，我们便可以通过抓取互动语言形式从而探究话语的互动性。

从语料库的视角，我们可以通过检索相关的互动话语形式，进而从语言特点和话语功能角度分析其特点。前面提到过的何安平、徐曼菲（2003）的研究其实也属于这一类。Petch-Tyson（1998）和文秋芳等（2003）、文秋芳（2009）等则是对学生的书面语中的互动性做了语料库考察，得到了有价值的发现，比如学习者书面语有口语化倾向，作者/读者能见度高等。

以下介绍一下许家金、许宗瑞（2007）基于学习者语料库所做的互动话语词块的研究。

这一研究利用 COLSEC 语料库中的学生口语语料，自动提取其中 2-6 词的复现词块，然后人工筛选出其中具有典型话语互动作用的词块，如 I think，I don't know，you can see that，first of all，by the way，more or less，or something like that，if you like 等。筛选出的互动词块按功能表现被划分为认知传递、内容指向、语气调节、认知制约四类，并从形式和功能方面与英语母语者的口语语料 ICE-GB 口语部分提取出的相应互动词块，做了对比分析。从而得出结论：形式上，中国学生用于话语互动的英语短语形式单一，往往是汉语的简单对译，还特别表现得以自我中心。最后一点在基于中国学生的类联接研究中也得到印证（许家金、熊文新 2009：20-21）；话语功能上，中国学生在英语口语中表现得过于直率生硬和缺乏技巧，往往"慷慨陈词"、"直抒胸臆"，缺少缓和语气的表达形式。

从上述对衔接连贯、语境特征和话语互动的分析例证中可以看出，话语的几个核心特征

不容易完全分离开。在分析词语的衔接时，很自然会考虑到其话语连贯与语境的同构作用；同时，不论是文本还是口头话语，我们还必须认识到，真实话语也都是说话双方同构的结果。因此，说话双方的互动也是经常需要考虑的。

在描写话语特征时，可以专注于单个语言变体的话语特征的描写和分析；同时，为揭示某些话语特征的特点，我们需要针对不同语料加以比较。比如，中西方媒体对制裁伊朗的报导的比较，中国英语学习者的英语作文与英美大学生的作文的比较，口语和书面语的比较，英汉语话语特征的比较，法律文本与通用文本的比较，演讲比赛中的定题演讲与即兴演讲的比较，甚至学术语体内部不同话步（move）之间的对比等等。

不管是对单个语体的话语特征的分析，还是跨语体、跨语域、跨语言的话语特征比较，话语研究都试图寻求词汇语法形式及其语用功能，语言形式与主观情态和社会文化内涵之间的关联。这其中，语料库的主要作用是借助计算机手段发现和提取相应的话语特征，并提供频率数据和话语特征的分布情况。

3.4 话语分析选题的选择与发现

研究课题的选定，可以源自日常观察或偶然发现，更多的是通过文献阅读获得。然而，因为有语料库这样的基于概率的量化分析方法，我们也可以借助其帮助发现和挖掘隐藏在文本之下的话语现象。这里特别要介绍的语料库驱动的选题发现方法。

就目前的语料库技术而言，语料库驱动的方法主要包括词表、主题词、词块、主题词块、框合结构等。这些方法的特点是不预设任何语法项目和理论框架。简言之，分析开始前，我们并没有确定的检索项或分析对象，往往是穷举式地列出所有词项，包括单词、短语或非连续的词汇序列。

以主题词和主题词块为例，我们通过比较 BNC sampler 中的口语和书面语子库，得到了排位靠前的主题词是 you，er，I，yeah，erm，it’s，oh，got，it，know 等，排位靠前的主题词块为 you know，I mean，I think，I don’t，don’t know 等。这些自动生成的词汇和短语可以作为我们进一步分析典型的英语口语词汇和短语。这种方法得到的词表和短语列表具有穷尽性和标准一致性。而传统上选择口语性词项的做法主要是基于以往文献和直觉，难免挂一漏万。

4. 小结

通过概念梳理和案例分析，本文简述了如何借助语料库以及相关分析手段进行话语研究。我们将诸多话语特征加以提炼，概括为衔接连贯、语境特征和话语互动三个核心特征，同时这些特征还可以进行不同语体、语域等的对比分析。另外，我们还特别强调，话语研究的一个基本任务就是在语言形式和功能之间建立联系。语料库方法在其中的作用就是对语言形式的量化描写，并辅助完成话语形式和功能之间关联的建立。而研究问题的最终的回答，还是在于研究者对量化数据的解读。

基于语料库的话语研究，从操作角度看，可以很简单，比如从一两个单词入手，如 man 和 woman（Pearce 2008）；也可能需要设计很复杂的检索和计算，如 Biber（1988）、Biber *et al*. 2007）的多特征/多维度（MF/MD）语体变异研究。可以是主要限于语言层面的，比如个别词汇衔接手段的语料库考察（Flowerdew & Mahlberg 2009），也可以是社会层面的批评话语分析和话语身份认同研究。

因为基于语料库的话语研究可以在不同层面开展，因此要求研究者具有相关领域的知识，例如体裁分析、批评话语分析、社会语言学等方面的方法和认识。需要补充一点，一些学者开始利用包含图片或音视频的语料开展多模态话语的研究，但这种研究在分析方法与以文本为主的话语研究差异较大。因此不是本文的讨论重点。

最后，我们希望提出基于语料库的话语研究的一些局限，语料库方法对语言的描写主要还是立足于词项（Mahlberg 2007：193）。对话语现象的分析解读主要还是在于研究者。语料库在其中的核心作用是获取和发现包含研究现象的语言事实。其次，语料库方法更偏重于形式特征的抓取，只是选取了计算机容易提取的特征进行。研究不应该因计算机易于提取作为其出发点。最后，我们必须认识到，语料库方法主要来说是一个"望远镜"（Partington *et al*. 2004：144；Stubbs 1999，引自 Hunston 2002：20），在需要对文本进行细读时，研究者依然需要回到单个文本，去解读字里行间的隐含意思。

对于基于语料库的话语研究，我们的态度是，应尽可能发挥计算机给我们带来的便利，将话语研究从少量文本的细读延伸至话语特征的全局性宏观分析。

最后，回到本文的标题"从词语到话语"，它有两层含义：一是，语言描写应扩展其视野，不应局限于词汇、短语和语句，也应着眼于更长的语言单位；其二，在语料库技术的辅助之下，可以从词语提供的信息推知话语特征。

附注

* 本文撰写得到教育部人文社会科学研究项目"基于语料库的中国大学生英语口语话语特征研究"（项目号：08JC740002）资助。

** 刘霞、熊焱帮忙校对了文章初稿，在此表示感谢。

参考文献

[1] Baker, P. *Using Corpora in Discourse Analysis* [M]. London: Continuum, 2006.

[2] Biber, D. *Variation across Speech and Writing* [M]. Cambridge: Cambridge University Press, 1988.

[3] Biber, D. *University Language: A Corpus-based Study of Spoken and Written Registers* [M]. Amsterdam: John Benjamins, 2006.

[4] Biber, D., S. Johansson, G. Leech, S. Conrad, & E. Finegan. *The Longman Grammar of Spoken and Written English* [M]. London: Longman, 1999.

[5] Biber, D., U. Connor & T. Upton. *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure* [M]. Amsterdam: John Benjamins, 2007.

[6] Fitzpatrick, E. (ed.). *Corpus Linguistics beyond the Word: Corpus Research from Phrase to Discourse* [C]. Amsterdam: Rodopi, 2007.

[7] Flowerdew, J. & M. Mahlberg (eds.). *Lexical Cohesion and Corpus Linguistics* [C]. Amsterdam: John Benjamins, 2009.

[8] Gee, J. *An Introduction to Discourse Analysis: Theory and Method* [M]. London: Routledge, 1999.

[9] Granger, S. (ed.). *Learner English on Computer* [C]. London: Longman, 1998.

[10] Granger, S., J. Hung & S. Petch-Tyson (eds.). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* [C]. Amsterdam: John Benjamins, 2002.

[11] Halliday, M. & R. Hasan. *Cohesion in English* [M]. London: Longman, 1976.

[12] Halliday, M. Linguistic function and literary style: An inquiry into the language of William Golding's *The Inheritors* [A]. In D. Freeman (ed.). *Essays in Modern Stylistics* [C]. London: Methuen, 1981: 325-360.

[13] Hoey, M. *Patterns of Lexis in Text* [M]. Oxford: Oxford University Press, 1991.

[14] Hunston, S. *Corpora in Applied Linguistics* [M]. Cambridge: Cambridge University Press, 2002.

[15] Hyland, K. *Disciplinary Discourses: Social Interactions in Academic Writing* [M]. London: Longman/Ann Arbor: University of Michigan Press, 2004

[16] Hyland, K. Disciplinary differences: Language variation in academic discourses [A]. In K. Hyland & M. Bondi (eds.). *Academic Discourse across Disciplines* [C]. Bern: Peter Lang, 2006: 17-48.

[17] Mahlberg, M. Lexical items in discourse: Identifying local textual functions of *sustainable development* [A].

In M. Hoey, M. Mahlberg, M. Stubbs, W. Teubert (eds.). *Text, Discourse and Corpora: Theory and Analysis* [C]. London: Continuum, 2007.

[18] Malinowski, B. The problem of meaning in primitive languages [A]. Supplement I in C. Ogden & I. Richards. *The Meaning of Meaning* [M]. (of the 10th ed. (1972)), 1923: 296-336.

[19] Partington, A., J. Morley & L. Haarman (eds.). *Corpora and Discourse* [C]. Bern: Peter Lang, 2004.

[20] Pearce, M. Investigating the collocational behaviour of MAN and WOMAN in the BNC using Sketch Engine [J]. *Corpora*, 2008, 3(1): 1-29.

[21] Petch-Tyson, S. Writer/reader visibility in EFL written discourse [A]. In S. Granger (ed.). *Learner English on Computer* [C], 1998: 107-118.

[22] Schiffrin, D., D. Tannen & H. Hamilton (eds.). *The Handbook of Discourse Analysis* [C]. Oxford: Blackwell, 2001.

[23] Semino, E. & M. Short. *Corpus Stylistics: Speech, Writing and Thought Presentation in a Corpus of English Writing* [M]. London: Routledge, 2004.

[24] 何安平，徐曼菲. 中国大学生英语口语 Small Words 的研究 [J]. 外语教学与研究，2003,（6）：446-452.

[25] 文秋芳. 学习者英语语体特征变化的研究 [J]. 外国语, 2009,（4）：2-10.

[26] 文秋芳，丁言仁，王文宇. 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，2003,（4）：268-274.

[27] 许家金. 粉丝眼中的小沈阳：基于词语搭配的认同话语建构研究 [R]. 11 月 12 日北京大学外国语学院语言学沙龙报告，2009a.

[28] 许家金. 中国英语专业大学生英文口头叙事中的自我形象：词语搭配研究视角 [R]. 12 月 26 日–27 日“首届全国学习者语料库专题研讨会”主旨发言，2009b.

[29] 许家金，赖辉. 基于语料库的知识分子形象的话语建构 [R]，第三届“当代中国新话语”国际研讨会宣读论文，5 月 15 日–16 日，2009.

[30] 许家金，熊文新. 基于学习者英语语料的类联接研究：概念、方法及例析 [J]. 外语电化教学，2009 ,(3): 18-23.

[31] 许家金，许宗瑞. 中国大学生英语口语中的互动话语词块研究 [J]. 外语教学与研究，2007, (6): 437-443.

From Lexis to Discourse: Using Corpora for Discourse Studies

Xu Jiajin

(National Research Centre for Foreign Language Education, Beijing Foreign Studies University, Beijing 100089, China)

Abstract: This paper reviews the major aspects of discourse and categorizes them into three strands of research efforts, namely, cohesion and coherence, contextual features, and interactivity of discourse. Moreover, inter-textuality/register-variability and subjectivity and socialization of discourse are also discussed. The main part of the paper illustrates with sample analysis the corpus methods applied to the above-mentioned discourse features. The paper ends with the strengths and weaknesses of using corpora for discourse studies.

Keywords: discourse features, discourse studies, corpus method

收稿日期：2009-12-03; 修订稿：2010-02-26

作者简介:

许 家 金：北京外国语大学中国外语教育研究中心专职研究员，副教授，博士。主要研究方向为：话语分析、语料库语言学。