Modern Linguistics 现代语言学, 2022, 10(12), 2937-2943
Published Online December 2022 in Hans. http://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2022.1012397

Hans汉斯

试论语料库辅助语篇分析研究工具及其近十年在国内研究中的应用

王晨缘，李晶洁

东华大学，上海

收稿日期：2022年11月4日；录用日期：2022年12月5日；发布日期：2022年12月13日

摘　要

本文简要介绍了语篇分析及语料库研究的内容，强调两者在研究中的紧密关系。后重点介绍了语料库辅助语篇分析的研究方法中各个重要的分析工具，并展示了其在国内近十年(2013~2022)应用情况。本研究发现虽然总体分析工具的使用有明显提升，但是最新的工具还未被研究者们所注意或充分利用。

关键词

语料库，语篇分析，研究方法，分析工具

Research Tools of Corpus-Based Discourse Analysis and Their Application in Domestic Studies

Chenyuan Wang, Jingjie Li

Donghua University, Shanghai

Received: Nov. 4th, 2022; accepted: Dec. 5th, 2022; published: Dec. 13th, 2022

Abstract

This article briefly introduced the content of discourse analysis and corpus research, emphasizing the close relationship between them. Then, we introduced several important research tools in the research methods of corpus-assisted discourse analysis, the application of which between 2013 and 2022 was also presented in Chinese corpus-based discourse analysis studies. This study found that the frequency of analysis tools use was generally increasing. However, the recently-invented tools have not got enough attention and full use.

文章引用: 王晨缘, 李晶洁. 试论语料库辅助语篇分析研究工具及其近十年在国内研究中的应用[J]. 现代语言学, 2022, 10(12): 2937-2943. DOI: 10.12677/ml.2022.1012397

Keywords

Corpus, Discourse Analysis, Research Methods, Analysis Tools

1. 引言

语篇在英语中可以对应两个词 discourse 和 text，这两个词在文献中经常混用(许家金，2019) [1]，中文翻译时也选择不一。本文的“语篇分析”对应的是 discourse analysis，不限于对语篇衔接等语言结构的表达。语篇分析近几十年得到了较大的发展，吸收了包括语言学、心理学、社会学等学科的研究成果，内涵越来越丰富。对此，语言学等学科的学者展开了很多相关的研究。因为不同领域的不同理论背景，相关研究中采用的方法数量众多(许家金，2014) [2]。本文聚焦语料库辅助下使用的语篇分析方法，主要介绍了三大影响力较大的研究方法：衔接分析方法、体裁分析方法和语域研究分析方法。这三个研究方法同时也是近年研究的热点。为更好体现研究客观性，实现对于大量语料的处理，分析工具就显得必不可少。而多数研究者对于可使用的研究工具，特别是对于新推出的分析工具并不了解。为提升学者对于重要分析工具的了解以及分析工具的使用，本文在衔接分析方法中介绍了 Coh-Metrix 和 TAACO 两种分析工具；在体裁分析方法中介绍了语篇瓦片技术(TextTiling)、TextmithTools、WordSkew 三种分析工具；最后，在语域研究分析方法中介绍了多维分析工具。

2. 理论背景

2.1. 语篇分析理论

语篇分析理论来源众多，其分支和流派也数量众多(许家金，2019：10)。目前来说，语篇分析还没有单一的理论指导，使用的分析方法和步骤也没有得到统一，只能说还属于“尚未定性的学科”(黄国文，2001：01) [3]。现代意义上的语篇分析可以追溯到 Zellig Harris 在 1952 年发表的文章“Discourse Analysis”。[4] Harris 是美国结构主义的代表人物，被认为是美国结构主义的集大成者，所以他对语篇分析的定义也带有明显的结构主义特点。后来 Halliday 在 1985 年出版 An Introduction of Functional Grammar 一书，在系统功能语言学的理论下为语篇分析建立了语法体系(黄国文，2010) [5]，相关研究通常被称为“功能/系统语篇分析”。1994 年 Schiffrin [6]对比了语篇分析中六种主要的研究方法，影响较大。总的来说，语篇分析在 20 世纪 70 年代之后得到了很大的发展。许家金认为语篇分析核心内容包括：话语的结构特征、话语的语义特征、话语的社会属性(2019: 14)。在关于这些核心内容的很多研究中，语料库都作为数据处理的工具得到很多的应用。

2.2. 语料库研究

语料库研究以大规模真实的语料为研究对象，借助计算机技术的发展，揭示语言隐含规律，辅助语言的多个领域的研究。Hunston [7]明确提出“语料库及语料库研究在过去几十年里给语言研究及语言应用研究带来了革命性的变化”(2002: 01)。许家金在其《语料库与话语研究》一书中概括了语料库研究的四个核心特征(2019: 07)：“用”，即尊重语言事实，关注语言用法；“量”，即“量化”手段；“器”，

即概率统计和语料库分析的工具；“聚”，即概率性，体现语言和社会语言学变量间的共选关系。语料库语言学逐渐兴起于20世纪中后期，一方面是因为实证主义和行为主义对收集真实语料的推崇，另一方面计算机技术的发展为大型语料库的建立提供了条件。语料库语言学关注自然语言实例，突出概率数据，能够帮助研究人员提高其研究的客观性。近年来，语料库被用于多个研究领域(雷秀云、杨惠中，2001 [8]；梁茂成，2014 [9]；马晓雷、陈颖芳；2016 [10])，而在这些研究中，语篇分析(话语研究)是“当仁不让的主角”(许家金，2019：05) [1]。但是大规模语料面前，人自身的处理能力就远远不足，必须依靠不同的分析工具。在这种情况下各种分析工具也应运而生，好的工具能够极大减轻研究过程中的工作量，同时帮助研究者深入挖掘语料价值。

3. 研究方法

为揭示本文介绍的语料库辅助语篇分析研究工具近年来在国内的使用情况，本研究采用文献检索方法，以文中重点讨论的Coh-Metrix、TAACO、语篇瓦片技术(TextTiling)、Textmith Tools、WordSkew、多维分析法MDA (Multidimensional Analysis)、MAT多维分析工具为搜索关键词，在中国知网(CNKI)进行搜索，剔除非语料库辅助语篇分析相关论文，共检索得35篇论文。

4. 语料库辅助语篇分析研究工具

首先，我们先来看一下语料库与语篇分析研究的紧密联系。中国著名语言学家桂诗春在2014年的专访中就提到语篇分析和语料库是语言研究的两大支柱[11]，两者的结合一方面能够对文本提出假设，然后用后者来证实，另一方面它们能够处理更多的型式维度(05)。许家金在介绍语料库研究和语篇分析(话语研究)时提到在近年国际范围的语料库相关研究的成果数量和影响而论，其与语篇分析(话语研究)的结合“成绩最为突出”(2019: 01)。语料库研究方法和技术对词汇短语层面的分析是话语研究的切入点，为后者提供了量化的数据(2019: 09)。另一个能体现两者结合研究的流行是相关期刊的创办：2017年斯普林格出版社《语料库语用学》创刊、2018年英国卡迪夫大学出版社创办《语料库与话语研究学刊》、2019年约翰·本杰明出版社出版《语域研究》创刊号。综上可以看出，语料库结合语篇分析是当前的研究热点之一。而衔接分析方法、体裁分析方法和语域研究分析方法是语料库辅助语篇分析研究中的热点。因此本文聚焦于这3种语料库辅助语篇分析研究方法，重点介绍其相关研究工具的具体内容以及应用状况。

4.1. 衔接分析方法

目前专注于衔接连贯的策略和分析的主要有两款工具：Coh-Metrix和TAACO (许家金，2019：89)。以下分别进行介绍。

Coh-Metrix是一个基于网络的文本分析工具，由美国孟菲斯大学的McNamera等人设计。他们认为美国教育部门长期依赖的可读性计算公式(readability formula)使教材中有着大量不连贯的语句，造成学生理解的困难(梁茂成，2006：285) [12]。他们结合计算机语言学和语料库语言学的多种技术，使其工具可有效分析文本的衔接连贯、句法复杂程度、词汇使用特点等等方面(桂林，2010：446) [13]。这个分析工具的研发获得了2002~2005年美国教育研究与改进办公室高达1,425,000美元的资金支持(Graesser *et al.* 2004) [14]。Coh-Metrix有两大优势：一是对现有的文本自动处理工具的充分利用；二是利用LSA技术，对文本的语义进行量化，不再停留在形式层面的计算(杜慧颖、蔡金亭，2013) [15]。虽现在的Coh-Metrix已经可以对106个词汇语法和语义特征进行分析，但是从它一开始的设计初衷可以看出对文本的衔接分析是它的重点。江进林在介绍Coh-Metrix在文本衔接研究中的使用时，用表1清晰展现了与衔接性直接

相关的 Coh-Metrix 维度及特征(2016: 60) [16]：

Table 1. Cohesion-related dimension and characteristics of Coh-Metrix
表 1. 与衔接性直接相关的 Coh-Metrix 维度及特征(江进林，2016：60)

维度	包含的变量	变量数量
指称衔接	相邻句子名词重叠的平均数 相邻句子论元重叠的平均数 相邻句子词干重叠的平均数 所有句子名词重叠的平均数 所有句子论元重叠的平均数 所有句子词干重叠的平均数 相邻句子实词重叠的平均数与标准差 所有句子实词重叠的平均数与标准差	10
连词	所有连词的比例 因果连词的比例 逻辑连词的比例 转折连词的比例 时序连词的比例 扩展时序连词的比例 肯定意义连词(如 also，moreover)的比例 否定意义连词(如 but，however)的比例	9
潜语义分析	相邻句子语义相似度的平均数和标准差 段落内所有句子语义相似度的平均数和标准差 相邻段落语义相似度的平均数和标准差 所有句子语义相似度的平均数和标准差	8

因其功能的强大，操作的相对简便，近十年来 Coh-Metrix 在国内国外都有应用。

虽然总体来看，Coh-Metrix 分析工具得到了较多的积极评价，但是也存在一些问题。比较突出的是其处理文本的数量有限，在线使用时不稳定。这严重限制了其研究语料库的规模(许家金，2019：91)。也是在这样的背景下，单机版软件 TAACO (Tool for the Automatic Analysis of Cohesion)问世。

TAACO 是一款很新的分析工具，Crossley，Kristopher Kyle，和 McNamara 在 2016 年发表文章详细介绍了它的特点、用法等[17]。2018 年 Crossley，Kristopher Kyle，和 Dascalu 又介绍了 2.0 版本[18]。TAACO 是一款免费提供的文本分析工具，易于使用，可在大多数操作系统上运行，与 Coh-Metrix 不同，他可以安装到用户的硬盘驱动器上，批量处理文本文件。不过 TAACO 的安装需要配置 Python 运行环境，但并不需要什么专业的编程知识。Crossley 等人在文章中详细介绍了工具的使用方法：通过双击 TAACO 图标进行启动，TAACO 界面是易于使用且直观的图形用户界面(GUI)，用户需要选择包括感兴趣文件(.txt 格式)的输入文件夹(2016: 1230)。相比于其他的分析工具，TAACO 有很多优势，例如它能够报告数量更多，

包括局部，整体和全文的内聚标记，还能允许用户独立安全地处理敏感数据。第二版本在原来基础上，功能更加完善。

4.2. 体裁分析方法

体裁分析主要指的是“Swales 等人倡导的学术英语话语研究方法”(许家金，2019：92) [1]，有助于对语篇组织模式的解析，对特定语篇具有的宏观认知结构的挖掘(秦秀白，1997：11) [19]。目前，体裁分析多基于 Swales 和 Bhatia 的理论框架，以学术语篇等为研究对象，采用“语步分析”方法(宋仁福，2020：19) [20]。“语步分析”主要是对文章中的引言、方法、结论等话语单位进行分析，相关的计算机语言学和语料库分析工具比较多，例如语篇瓦片技术(TextTiling)、TextmithTools、WordSkew 等等。以下简介提到这三种分析工具。

语篇瓦片技术及相关程序是在 20 世纪 90 年代早期，由美国加州大学伯克利分校信息学院的 Marti Hearst 教授开发。1993 年，Hearst 在 TextTiling:A Quantitative Approach to Discourse Segmentation 一文中对语篇瓦片技术作了介绍[21]。语篇瓦片技术被定义为一种将全长文本文档划分为连贯的多段单元的方法。相比于“语步分析”的流行，TextTiling 的使用很少(宋仁福，2016：86) [22]。但是宋仁福在 2016 年的文章中强调了语篇瓦片技术的优势：利用自然语言处理技术，在自动切分文本子话题、统计等方面，“标准客观、前后一致”(94)。更重要的是，传统的语步分析在处理大量文本时费时费力，而语篇瓦片技术作为计算机自动分析处理程序，省时省力，优势显著。

梁茂成和刘霞在 2014 年介绍了一种新的语料库分析工具 TextSmith Tools [23]。他们认为作为语料库对比标准方法的主题词分析法，在用于分析词汇或多词单位在语料库文本的不同部分的分布差异时显得不足，所以为了分析词汇短语特有的语篇特征和组织功能，他们设计并开发了这一新的工具(梁茂成、刘霞，2014) [4]。TextSmith Tools 主要功能包括文本切分、主题词分析和个案分析，能够对大规模的文本进行语篇结构的分析，且用户界面设计友好，是对话语语步分析的一个推进(许家金，2019：93) [1]。

2016 年 Michael Barlow 发文介绍了 WordSkew 文本分析工具[24]。该程序当前可用于 Windows。根据 Barlow 的描述，WordSkew 显著特征在于可以指定话语的单位和分割话语的方式，针对每个切分的部分给出搜索的结果(105)。许家金指出 WordSkew 仍然关注的是索引分析，只是用研究者设定的百分比去呈现检索得到的频数信息(2019: 93)。

4.3. 语域研究方法

在语篇分析中，语域、语体、文体等是一些近似的概念，在各种文献中经常出现混用的情况。2009 年 Biber 和 Conrad 专门对语域、语类、文体三个术语进行区分，这三个术语也代表着三种分析语篇的视角(03) [25]。其中，Halliday 等人倡导的语域分析受到了广泛的关注[26]。根据他们的观点：“语言因使用的情境而变化”，语域指的是“这种因使用而导致的各类语言变体”(胡春雨、谭金琳，2020：67) [27]。语域变异的研究中，使用的主要方法是多维分析法 MDA (Multidimensional Analysis). 多维分析是由 Douglas Biber 在其 1984 年的博士论文中首先提出。后来 Biber 在 1988 年出版了《口语及书面语间的语域变异》一书，对该方法进行了改编扩充。多维分析基于大量的语言特征的统计，运用因子分析方法，对语言的潜在维度进行考察(胡春雨、谭金琳，2020：67) [27]。Biber 曾使用过 41 和 120 多个词汇语法特征，但是最经典的还是 67 种语法特征(许家金，2019：96) [1]。但是这一方法有个很大的不足：使用的相关软件并不对外开放。2013 年 Andrea Nini 开发了 MAT 多维分析工具，打破了这一壁垒。新的多维分析内嵌 Stanford POS Tagger 词性标注工具，语言特征的提取和统计都能后台自行批次处理，操作流程大大简化。因多维分析为语域研究开辟的新的思路，所以在很多的领域都有广泛的应用。胡春雨和谭金琳在 2020 年的文章中就总结了国内外的应用状况。

5. 研究工具在国内研究中的应用

根据以上论述，在三个主要语料库辅助语篇分析方法中都有高效的分析工具。这些工具不仅使得对于大型语料的语篇研究成为可能，也为研究提供新的思路和方法。图 1 展示了这些分析工具近 10 年在国内相关研究中的应用情况：

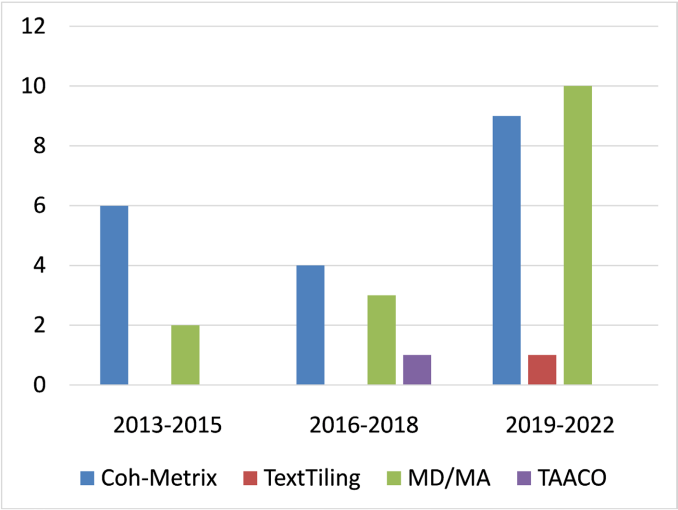

Figure 1. Application of analysis tools between 2013~2022
图 1. 2013~2022 年各分析工具运用情况

总体来看，由于 TAACO，Textmith Tools 和 WordSkew 推出的时间都不久，所以国内结合其使用开展的语篇研究极少，如刘芳芳结合 Coh-Metrix、TAACO，分析中国英语专业学术议论文中衔接手段的使用(2018) [28]。而 TextTiling 虽然推出的时间已经过去很久，但是并没有在国内引起关注，只有宋仁福在 2016 年发表的论文中详细进行了介绍，并在 2020 年针对英文语料库研究论著前言的多维分析中，使用了这一技术。与之相比，Coh-Metrix 和多维分析工具受到研究者更多青睐，并且使用频率近三年有明显增长。这一方面体现分析工具在语料库辅助语篇分析中的作用越来越受到研究者的肯定，另一方面也体现了更多新的研究工具还未被研究者熟知和运用。

6. 结论

语篇分析和语料库都是目前研究的重点方向，随着计算机技术的不断发展，语料库分析方法中可使用的分析工具的不断开发升级，两者结合方面的研究将能产出更多的学术成果，而在这些研究中，方法和分析工具的重要性显得格外突出。本文首先介绍了语篇分析和语料库的背景知识，两者的特点使其结合成为必然。后面介绍了 3 种语料库辅助语篇分析研究方法(衔接分析方法、体裁分析方法和语域研究分析方法)中的重要分析工具。多数研究方法的分析工具都由国外提出，然后传入中国，结合中国学者的智慧与中国的研究特点发挥其作用。但是近年也有中国学者提出新的分析工具，这个现象是非常可喜的。总体来说，分析工具的使用越来越受到研究者的重视，但是对于新工具的运用却远远不够，并没有充分发挥分析工具的功能。通过本文的介绍，更多学者能够了解分析工具的功能，提高分析工具的利用效率。

参考文献

[1] 许家金. 语料库与话语研究[M]. 北京: 外语教学与研究出版社, 2019.
[2] 许家金. 许家金谈语料库语言学的本体与方法[J]. 语料库语言学, 2014(2): 35-44, 113.

[3] 黄国文. 功能语篇分析纵横谈[J]. 外语与外语教学, 2001(12): 1-4, 19.

[4] Harris, Z.S. (1952) Discourse Analysis. *Language*, **28**, 1-30. https://doi.org/10.2307/409987

[5] 黄国文. 语篇分析与系统功能语言学理论的建构[J]. 外语与外语教学, 2010(5): 1-4.

[6] Schiffrin, D. (1994) Approaches to Discourse. Basil Black-Well, Oxford.

[7] Hunston, S. (2012) Corpora in Applied Linguistics. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781139524773

[8] 雷秀云, 杨惠中. 基于语料库的研究方法及 MD/MF 模型与学术英语语体研究[J]. 当代语言学, 2001(2): 143-151, 158.

[9] 梁茂成. 语料库、平义原则和美国法律中的诉讼证据[J]. 语料库语言学, 2014(1): 25-33, 110-111.

[10] 马晓雷, 陈颖芳. 基于共词分析的语料库语言学研究现状分析(1971-2015) [J]. 语料库语言学, 2016(1): 41-54, 116.

[11] 桂诗春. 语料库语言学答客问[J]. 语料库语言学, 2014, 1(1): 1-15, 110.

[12] 梁茂成. 学习者书面语语篇连贯性的研究[J]. 现代外语, 2006, 29(3): 284-292.

[13] 桂林. 基于计算机评估的 L1 和 L2 作文词汇水平对比研究[J]. 外语教学与研究, 2010, 42(6): 445-450.

[14] Graesser, A.C., McNamara, D.S., Louwerse, M.M. and Cai, Z. (2004) Coh-Metrix: Analysis of Text on Cohesion and Language. *Behavioral Research Methods, Instruments, and Computers*, **36**, 193-202. https://doi.org/10.3758/BF03195564

[15] 杜慧颖, 蔡金亭. 基于 Coh-Metrix 的中国英语学习者议论文写作质量预测模型研究[J]. 现代外语, 2013(3): 293-300.

[16] 江进林. Coh-Metrix 工具在外语教学与研究中的应用[J]. 中国外语, 2016(5): 58-65.

[17] Crossley, S.A., Kyle, K. and McNamara, D.S. (2016) The Tool for the Automatic Analysis of Text Cohesion (TAACO): Automatic Assessment of Local, Global, and Text Cohesion. *Behavior Research Methods*, **40**, 1227-1237. https://doi.org/10.3758/s13428-015-0651-7

[18] Crossley, S.A., Kyle, K. and Dascalu, M. (2019) The Tool for the Automatic Analysis of Cohesion 2.0: Integrating Semantic Similarity and Text Overlap. *Behavior Research Methods*, **51**, 14-27. https://doi.org/10.3758/s13428-018-1142-4

[19] 秦秀白. “体裁分析”概说[J]. 外国语, 1997(6): 9-16.

[20] 宋仁福. 英文语料库研究论著前言的多维体裁分析[J]. 语料库语言学, 2020(1): 18-31, 113.

[21] Hearst, M.A. and Plaunt, C. (1993) Subtopic Structuring for Full-Length Document Access. In: Korfhage, R., Rasmussen, E. and Willett, P., Eds., *Proceedings of the* 16*th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Association for Computing Machinery, New York, 59-68. https://doi.org/10.1145/160688.160695

[22] 宋仁福. “语篇瓦片叠压”(TextTiling)技术解析[J]. 语料库语言学, 2016(2): 86-95, 116.

[23] 梁茂成, 刘霞. 语篇内部的短语学特征分布模式探索——以学术论文为例[J]. 解放军外国语学院学报, 2014, 37(4): 1-11, 22.

[24] Barlow, M. (2016) WordSkew: Linking Corpus Data and Discourse Structure. *International Journal of Corpus Linguistics*, **21**, 105-115. https://doi.org/10.1075/ijcl.21.1.05bar

[25] Biber, D. and Conrad, S. (2012) Register, Genre and Style. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511814358

[26] Halliday, M.A.K. (1994) An Introduction to Functional Grammar. Edward Arnold, London.

[27] 胡春雨, 谭金琳. 中美企业致股东信语域特征的多维分析[J]. 外语与外语教学, 2020(6): 64-75.

[28] 刘芳芳. 基于语料库的中国英语专业学生议论文写作中衔接手段的使用与发展及其与作文质量关系的研究[D]: [硕士学位论文]. 大连: 大连外国语大学, 2018.