语料库研究学术源流考*

北京外国语大学　许家金

提要：本文通过一手文献，特别是新发现文献，围绕“用”“量”“器”“聚”四项设计特征，考证语料库研究的核心概念和理论源流。语料库研究哲学背景深厚，语言学发展脉络清晰，将其全然视作研究方法有悖史实。语料库研究的演进与语言学发展史高度同步。20 世纪上半叶可视作语料库研究 1.0 时代，也可称为前电子语料库时代，它与（美国）结构主义语言学几乎可以等同视之。20 世纪五六十年代至今的语料库研究 2.0 时代委身于形形色色的功能语言学研究之中。在大数据技术推动下，21 世纪的语料库研究 3.0 时代，借助集成丰富语境信息的海量数据，定能完成语料库研究“用”“量”“器”“聚”的全面提升，从而将语言学塑造成领先学科，于语言描写和阐释方面皆有大成。

关键词：语料库研究核心概念、哲学渊源、语言学渊源、语料库语言学、学术史

[中图分类号] H0-06　[文献标识码] A　[文章编号] 1000-0429(2017)01-0051-13

1. 语料库研究的设计特征

当今语言研究中，语料库相关的学问，即便不是显学，也称得上热门无疑。对此，学界有激赏，也有质疑。考订语料库语言学发展源流，适逢其时。这有助于学界客观认识语料库研究现状，及其在整个语言学发展中的历史定位。

中国的语料库研究，同诸多学科领域一样，是舶来学术，其发展与西方语料库研究传统有着不可割裂的渊源。因此，本文将重点对西方语料库研究的发展加以梳理，从而知源明流。厘定学术谱系，撰写学术史，是严谨而审慎的工作。因此，有必要于开篇就本文谈及的内容、范围及写作原则加以界定和说明。首先是对“语料库”一词的理解。目前的语言学文献一般认为，大规模电子文本数据库为语料库。然而，19 世纪中叶至 20 世纪初，便有学者（如 Pitman 1843[1]；

* 本文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）的资助，同时承蒙卫乃兴教授、李文中教授、李民博士提供细致中肯的反馈，在此一并致谢。

Kaeding 1897/1898；Thorndike 1921；陈鹤琴 1922 等）按一定取样方案收集大规模文本，以数理统计方法，获得词频、词序、文本分布信息。这些工作，即使以今天的眼光看，虽全为手工计数，但其发现丝毫不逊于计算机时代的语料库成果。因此，本文将计算机普及之前，依一定取样原则采集、记录在纸面或磁带上，并有明确（应用）语言学目的的成规模语言素材汇集，也视作语料库。此类语料库，Francis（1992）称之为“前计算机时代语料库”（language corpora B.C.[2]）。对这些“古董级”语料库的认定，于探讨学术史意义重大。只有这样，才可以贯通历史，将先贤与今人的学术关联起来。

不将“电子化”作为判定语料库的必要条件，有助于我们更深入地探讨语料库研究的本质。我们将语料库研究的设计特征概括为：用、量、器、聚（详见许家金 2014：35-36）。“用”即尊重语言事实，关注用法。正是在这个意义上，Teubert（2007：57）把语料库研究表述为“言语语言学”（parole linguistics）。因此不难理解，语料库研究以对语用的充分描写为根本。但若仅从研究对象（即外在语言或言语）来界定，语料库研究尚无法同社会语言学、话语分析、语用学等领域区分开。语料库研究在“量”上的要求，特别是通过“量化”手段研究语言，是其更为关键的设计特征。这里的“量”，是指语料库研究更关心大规模语言实例中呈现出的使用趋势。语料库研究主张语言是概率性现象（Halliday 1991）。Halliday（1992）指出概率性既是对语言实例的描摹，也是一种理论构念。在语料库研究中，这种概率性需要通过数理统计及计算机分析来实现，即“器”。在理论层面，语料库研究很大程度上体现为一种聚集或共现关系，可概括为“聚”。与此相关的语料库研究，主要受到伦敦学派“语境论”及社会语言学等理论的影响。其中围绕搭配或词语共现开展的短语学研究，有着清晰的弗斯语言学背景；另外，如火如荼的语体变异、各种基于语料库的话语研究，或多或少受到 Labov 为代表的变异社会语言学的影响。前者更多关注语言成分之间的关联，其理论贡献在于语言意义；后者关注语言成分与社会语言学变量之间的共变，其理论聚焦点在于社会意义。这四项设计特征使得语料库研究自成一体，也揭示出它与其他领域的学科交叉特点。

除了“用”“量”“器”“聚”几项设计特征外，在判别和界定语料库研究时，还会遇到我们称之为“大写的语料库语言学”（big C Corpus Linguistics，即狭义的语料库研究）和“小写的语料库语言学”（little c corpus linguistics，即广义的语料库研究）的现象。大写语料库语言学通常都要对语料库进行穷尽式分析，而不像后者仅从语料库中引述符合个人研究所需的语言实例。这即是 Quirk（1960）所谓的“穷尽阐释”（total accountability）原则。这一点也是上述“量”的原则题中应有之义。

综上，“用”是基础，“量”是关键，“器”是途径，“聚”是语言学理论归宿。

本文基于第一手文献，特别是一些此前未被提及过的文献，梳理语料库研究中核心概念和理论视角的发展源流。限于篇幅本文仅讨论欧美所开展的英语语料库研究。有关汉语语料库的成就可参阅 Xu（2015）。

2. 语料库研究核心概念考

2.1　Corpus

英文 corpus 这一概念的发展历程大致经历了如下几个阶段：

第 1 阶段：表示文本汇集。约 400 多年前，corpus 指文学作品集萃、宗教经典全集、文献汇编[3]。

第 2 阶段：表示作为研究资料的文本汇集。19 世纪中后期到 20 世纪二三十年代，主要出现在（对比）语文学、历史学、神学、法学、人类学研究文献中，表示作为研究资料的文本汇集，多指古代典籍文献。其中以人类学家 Malinowski（1922）谈的 *corpus inscriptionum Kiriwiniensium*（基里维纳人语言素材）与今天的语料库概念最为接近。

第 3 阶段：表示作为语言学研究资料的文本汇集。20 世纪四五十年代在（美国结构主义）语言学文献中，开始出现 corpus 单独使用[4]表示语言研究素材库的含义。从目前查到的文献看，以结构主义语言学家 Harris（1947）为早。其使用语境如下：

> When such comparisons are carried out for a large *corpus*, we obtain morphemic segments which are repeated in various environments throughout the *corpus*.（同上：175，斜体系本文作者添加，下同）

第 4 阶段：表示作为语言学研究资料的电子文本汇集。20 世纪 60 年代以后，以布朗语料库（1967）为代表，逐渐确立了按一定取样方案采集、服务于语言研究目的的电子文本库这一概念。

汉语文献中“语料库”这一中文译法，大致是对照计算机术语“数据库”推衍而出[5]（杨惠中 1981：30；杨惠中、黄人杰 1982：60）。

2.2　Corpus linguistics

Corpus linguistics 的提法，进入英文学术话语也有确切记载。有文献表明，corpus linguistics 这一概念的出现不晚于 1959 年。其使用语境如下：

> This certainly is the assumption behind linguistic statements derived wholly from texts. By definition, this means *closed corpus linguistics*, formerly known as philology. But anthropological linguistics operates with an *open corpus* ....（Voegelin 1959：216）

1959 年的这一用法很有意味，它将从对比语文学到美国结构主义这前后两个世纪的语言学研究都视作语料库研究，其共性是研究结论应全然基于文本。对比语文学所考察的梵语、拉丁语、德语材料，很多不再使用，只存乎文献，因此是封闭型的；对真实发生的鲜活语言使用则是开放式语料。后者是如今语料库概念的非标记项，因此，open corpus linguistics 中的 open 自然也就略去。

2.3 Concordance

Concordance 一词的历史较为久远。各种文献都指出，大约在 1230 年（又 1262 年），枢机主教 Hugh of Saint-Cher 编制的《拉丁文圣经通检》（*Concordantiae Sacrorum Bibliorum*，又 Vulgate，另有学者认为所编为希伯来文圣经索引），即是给出圣经中所有词汇用例和出处的一种工具书，称为“词汇索引”。这项工作是词汇索引的滥觞。这种工具书，又称“逐字索引”（刘殿爵等 2000）。由“逐字索引”字面可见，这是集词表和索引为一体的信息呈现方式。除了不是由计算机生成外，它同语料库研究中的词汇索引并无二致。这种工具书通常是为研读文学和宗教经典而编。我国学术界上世纪二三十年代兴起一股“索引运动”热潮。其中以蔡廷干（1922/2014）所编《老解老》和洪业 1930 年前后开始编纂的“哈佛燕京引得丛书”为代表（洪业 1932），它们开创了我国词汇索引（曾译“堪靠灯”，又称“串珠”“通检”）编制的先河。Ohlman 和 Luhn 在 1960 年前后几乎同时分别研制出计算机化“带语境的关键词”（Keyword in Context，简称 KWIC）的索引技术（Luhn 1960），后为语料库索引工具（concordancer）汲取，成为语料库工具的主要功能之一。计算机索引工具的出现，将以往学者们皓首穷经方可完成的工作，变得分秒钟内即可实现，使得索引成为深入研究文本的起点。

2.4 Collocation

英文 collocation 这一概念的发展可以分为三个阶段。

第一阶段：400 多年前，从 collocation 表示“事物并置”这一含义开始，便有词语并置搭配的用法（参见 Simpson & Weiner 1989：487），它是对一种普通语言现象的描述。

第二阶段：1933 年前后，collocation 成为具有重要教学价值的应用语言学术语。1933 年英国学者 Harold Palmer 在日本出版的《有关英语搭配的第二阶段中期报告》（*Second Interim Report on English Collocations*）是一部较早系统描写英语核心词语搭配行为的学术文献。据 Cowie（1998a）考证，这份报告的实际编写者是词典学家 A. S. Hornby。该报告封面印有：

A collocation is a succession of two or more words that must be learnt as an integral whole and not pieced together from its component parts.

Cowie（同上：13）曾指出 Palmer（1933）是一本“被严重忽略了的经典之作”。另据 Cowie（1998b）考证，20 世纪 40 年代俄罗斯的短语及搭配研究兴起，对英语搭配研究也产生过积极影响。

第三阶段：上个世纪五六十年代，搭配发展成为具有理论语言学地位的专业术语，它强调的是语言成分之间的结伴、相互期待和相互预见关系。

这一阶段搭配概念的发展，当归功于 J. R. Firth，他认为“搭配”是“意义的多维方式”（Firth 1951/1957：194）中的一个维度。他还在另一作品《1930-55 年语言学理论要览》（A synopsis of linguistic theory，1930-55）中提出了“识词于其所友！”（You shall know a word by the company it keeps!）（Firth 1957/1968a：179）的说法。当然，“识词于其所友”的理念也见于其他早期文献，详见梁茂成（2014：26-27）的相关考证。

另外值得一提的是 Firth（1951/1957：194）在文中所谈的是“‘搭配’生义”（meaning by “collocation”），而非“搭配”这一术语本身。这一点很关键，这里所谈的“搭配生义”正是 Firth 语境意义观在词汇层面的实现机制——词语意义与它前后结伴的语境词语互相依存。在差不多相同的时期，美国结构主义语言学家 Harris（1954）提出的“分布假说”（distributional hypothesis）也主张语言成分的意义由其所处的语境决定，两者异曲同工。

最后要特别指出 Halliday 对词语搭配研究的贡献。Halliday（1961：276）在 Firth 提出搭配生义之后，较早提出了搭配的概率观（probabilistic collocation），并使用了 node（节点词）、collocate（搭配词）、span（跨距）这些概念，学界沿用至今。Halliday（1966：158）和 Sinclair（1966：415）还分别用实例说明了搭配词的统计算法。Sinclair *et al.*（1970/2004）尝试了卡方检验、叶茨校正的卡方检验、费舍尔精确检验和泊松分布 4 种搭配强度的算法。此后，语料库研究对于搭配的统计计算不断衍生，并发展成为独具解释力的短语学理论。

以上探讨的“语料库”“词汇索引”“搭配”等概念，正与语料库研究“用”“量”“器”“聚”这些本体特征相对应。厘清这些概念的发展脉络，有利于贯彻和运用语料库理念，开展实证研究。

3. 语料库研究的哲学和语言学源流

3.1　*语料库研究的哲学源流*

语料库研究的哲学基础是“经验论”，又称“经验主义”，是西方认识论的两大

主流之一。经验论源自英国，始于培根，并由洛克、贝克莱、休谟继承发扬。培根倡导由行为或外在现象归纳获得有关人类和世界的知识。而洛克提出的“白板说”更直观地说明了人类后天知识获得的机制（Russell 1946）。相比世界其他地区，语料库研究的发展在英国更为如火如荼，应当说这与英国的经验论哲学土壤不无关系。欧洲大陆的语料库研究起步也相当早，语料库学者分布地域范围也相当广。总体来说，欧洲大陆在语言学的认识论方面，经验论和唯理论兼而有之。

而美国的情况似乎有所不同，20 世纪中叶至今，美国语言学以生成语法为主导，其核心认识论是笛卡尔、莱布尼茨等人倡导的“唯理论”。美国语料库研究不乏亮色，但却未能走向语言学的中心舞台。这与生成语法领军人物 Chomsky（1957：14-17，97，1965：3-4）断然否认言语使用和语言产品的语言学价值有直接关系。生成语法认为语言是人本官能，社会属性不是语言研究的首要关切，因为日常语言充斥着语误和不规范用法，不宜作为语言研究对象。

经验论同唯理论的针锋相对，即“习性”与“本性”之争。在对语言本质的认识上，前者视语言为“事实”，后者视之为“心智”。在研究实践中，前者采取自下而上的归纳分析法，后者采取自上而下的演绎分析法。某种程度上，在相同研究视域内两者矛盾很难调和。

在探讨语料库研究哲学渊源时，有些学者（如 Louw 2011；何安平 2013）会溯及弗雷格、罗素、维特根斯坦、卡尔纳普等分析哲学家。这其中，主要是后期维特根斯坦思想——“意义即使用”与语料库研究的“用法观”较为契合（Firth 1957/1968b：138-139），其他几位属于逻辑实证学者，他们并不太关注日常语言。

若将经验论和唯理论看作宏观哲学原理的话，与语料库研究有关的还有一些中观、微观哲学原则。比如，一元论、格式塔观念和阐释学。Firth，Halliday，Sinclair，Stubbs，Teubert 是一元论的坚定支持者。这一派学者不认同有关“能力”（competence）与“使用”（performance）的区分。Halliday（1991：34）有关“气候”和“天气”的比喻最为直观形象。他指出人们每天体验或观察到的天气变化，累积起来，以更大的时间尺度看待，就是气候。某个地区的气候具有一般性或稳定性，是一种概率性特征，经年累月的微小天气变化是气候的基础。天气变化也会调节和影响气候。两者是同一个现象的两种观察角度。Saussure 所区分的“语言”与“言语”，Chomsky 所区分的“能力”与“使用”，Halliday 区分的“系统”（system）与“实例”（instance），不过是一币两面。有关语料库研究与阐释学的关联，Teubert（2010：199）谈得最多。这一观点循着语义研究的主线展开讨论。Teubert 指出“话语之外无意义”，所谓的事实也都是话语建构的事实。在这一点上，正合 Sinclair（2004）“信任文本”（trust the text）的观点。这里谈的一元论、格式塔和阐释学，很大程度上，还是一元论的问题，即语言之运用

与语言之本体，实不可分。研究“用”等同于研究“体”。

3.2　语料库研究的语言学源流

3.2.1　结构主义语言学与语料库研究 1.0 时代

如前所述，语料库研究在当今美国语言学界不是主流。然而在 20 世纪初至五六十年代美国结构主义一统天下，其核心任务就是基于真实语料的语言描写，特别是对印第安语言的描写。因此，美国结构主义又被称为“描写主义”（descriptivism）。其代表人物是 Boas，Sapir 和 Bloomfield。

真正对现代语料库研究产生直接影响的是所谓后布龙菲尔德学者（post-Bloomfieldians），代表人物有 Harris，Hockett，Pike，Twaddle。美国布朗大学 Francis 和 Kučera（Twaddle 在上世纪五六十年代正是布朗大学语言学系负责人），以及密歇根大学 Fries 等人是结构主义思想影响下从事语料库研究的早期重要代表。Francis 和 Kučera 专注于书面语语料的收集和描写，创建了最早的电子化英语语料库——布朗语料库（Francis & Kučera 1967）。很长时间里，布朗语料库俨然是平衡书面语语料库的事实标准，乃至 LOB 语料库和 BNC 语料库的建设都直接受其影响。Fries 更多关注口语语料的收集与描写（参见 Fries 1952）。学界对 Fries 在口语语料库研究方面的贡献有所忽视（Fries & Fries 1985）。他基于真实语料库的研究实践早于布朗语料库十多年，而 Fries 基于英语口语语料创编英语语法的这种做法，不但领先美国，甚至不输英国的“英语用法调查”。

在英国，伦敦大学学院（UCL）的 Quirk，Greenbaum，以及后来赴兰卡斯特大学任教的 Leech 都秉承了结构主义语言描写的传统。其中 Quirk 和 Leech 曾在美国接受语言学训练，Greenbaum 也在美国任教多年。与这几位学者共事并持相似语言观的还有一些北欧学者，如 Svartvik 和 Johansson 等。由这些学者主持创建的语料库包括“英语用法调查”、LOB 语料库、ICE 国际英语语料库家族、BNC 语料库等。他们对于英语的描写，口语和书面语并重，某种程度上更注重口语描写。他们立足共时语言研究，也十分重视历时语言考察。

上述努力的首要成绩是系列重要语料库的创建。在此基础上，一批基于语言描写的语法书应运而生。比如 Quirk *et al.*（1985）主编的《英语语法大全》（*The Comprehensive Grammar of the English Language*）[6]和 Biber *et al.*（1999）主编的《朗文口笔语语法》（*Longman Grammar of Spoken and Written English*），以及与之相配套的教学语法。

另外，如今各大主流语料库期刊、大量语料库研究专著、重要语料库国际会议，有关某个或某类语言特征的描写，仍然是热门选题。相关的语言特征描写

方法也不断改进。可见，结构主义对于语料库研究影响至今，然而在美国因为 Chomsky 理论的横空出世，结构主义理论路径的语料库研究却在英国和欧洲大陆得以发扬。这实在是因为生成语法理论与语料库研究针锋相对。Chomsky（1957：14-17，1965）明确主张：1）句子是否合乎语法不能同田野调查得来的语料划等号；2）句子合乎语法不必跟它是否语义通畅划等号；3）不能将英语句子合乎语法与否等同于统计出来的语言规律。Chomsky 只认可语句合乎语法的可能性（possibility），不认可语言使用的概率性（probability）。这些是对“用”“量”“器”“聚”等语料库研究设计特征的根本否定，可谓水火不容。然而，当争论双方各持不同的哲学观、认识论时，一方对另一方的否定，并不能真正驳倒对方，不过是道不同而已。

生成语法是作为结构主义反对者角色出现的。然而，结构主义语言学的辉煌时代已成历史。继结构主义而起的是功能语言学。功能语言学是当今语料库研究真正的学科背景。

3.2.2　功能语言学与语料库研究 2.0 时代

本文所谈功能语言学是广义的（参见 Siewierska 2011），包括伦敦学派的弗斯语言学及其衍生流派系统功能语言学、话语分析、社会语言学、语用学、认知语言学、类型学等。功能语言学者主张语言的普遍规律应从人的认知能力、语言的交际功能或历史演变等方面去寻求解释（陈平 1987：7）。这就不难理解，为什么功能语言学者十分关注语境、关注语言变异，因为他们不作非黑即白的断然（possible）结论，而是基于对语言事实的挖掘，归纳得出或然（probable）结论。

功能语言学者认为应当在语境中考察语言，关注语言使用和话语交际功能，才能综合了解语言的运作机制。在英国，Firth 所创的伦敦学派语言学影响至今。以 Sinclair 为代表，主要在伯明翰大学从事（过）研究的一批学者，如 Hunston，Teubert，Louw 等人，在 Malinowski 和 Firth 的“语境语义观”（contextual theory of meaning）指导下，提出语言描写和解释的“共选”（co-selection）理论。这一理论主张基于语料库，通过词语搭配、词类联接、语义倾向和评价特征，寻求对语义生成机制的解释，即“扩展意义单位”模型。这一原创性分析框架，倡导摒弃既有语言学理论，倚赖文本内部共文（co-text）和语境信息，对局部短语结构乃至语篇意义的建构作出解释。Sinclair（2004：164）称这种词汇主义语言观为“词汇语法 $_{\text{Sinclair}}$”（lexical grammar），多少有些类似徐通锵（1994）所提出的汉语研究“字本位”观点。准确地说，Sinclair 的“扩展意义单位”语料库语言学观是一种“短语本位”的全新语言观。

Halliday 开创的系统功能语言学，在初始阶段就与语料库研究有着不解之

缘。如本文第 1 节所述，他注重概率语言观，重视词汇语法的语言学价值，提出“词汇语法 Halliday”（lexicogrammar）的概念。如前所述，有关搭配研究的一些核心概念，如节点词、跨距及搭配的统计算法都由 Halliday 首创。早在 Sinclair（1991）提出“习语原则”（idiom principle）和“开放选择原则”（open choice principle）之前，Halliday（1966：152-153）就指出词语在横组合关系上的制约构成搭配关系，在纵聚合层面构成开放式的集合关系（set）。两者共同界定语言结构和系统。Halliday 主张的全局制约条件（global constraints）和局部制约条件（local constraints）正是他提出的系统与实例的互补性。系统功能语法因注重自身理论体系建设，并未就语料库相关理念作更多拓展。然而，系统功能语言学创立之初就与语料库研究方法高度兼容。

如今国际语料库研究领域，学者们用力最勤的是话语分析和社会语言学研究。这其中仍然可见语境论的影响。上世纪七八十年代话语分析研究蓄势成长，其主要策源地在英国和欧洲大陆。起初，学者们受篇章语言学影响，较多关注话语内部的谋篇布局，立足于文本语境。近一二十年话语分析的主体已转向研究话语背后的社会语境——话语权力或意识形态，即所谓的批判话语分析。基于语料库的批判话语研究是过去一段时期内语料库研究的主流。英国兰卡斯特大学的 McEnery，Baker 等学者是这一研究潮流的代表学者。而 Labov 上世纪六七十年代开创的变异社会语言学，近一二十年在语料库方法的强力支撑下，生发出强大的学术活力。美国语料库研究的领军人物 Biber（1988）基于多特征多维度（MF/MD）方法所考察的语域变异（register variation）是变异社会语言学的一个经典语料库研究应用。《朗文口笔语语法》是变异社会语言学框架下的代表作品，其本质是在结构主义语言描写体系之上增加了语域变异维度，从而在结构之外，增加了社会功能维度。

语用学研究目前也普遍提倡采用真实语料和量化实证方法。语用学的研究领域，特别是欧陆学派，与话语分析和社会语言学，交集甚广。《语用学学刊》（*Journal of Pragmatics*）是国际语言学期刊中采用语料库方法最多的期刊之一。

当今，认知语言学与语料库研究的结合愈发紧密。《认知语言学》（*Cognitive Linguistics*）杂志副主编 Janda（2013）统计发现，该刊物从 2008 年开始，出现明显的量化和实证转向。她注意到 2008 年以后认知语言学期刊论文半数以上运用语料库或其他实证手段。涉及的研究领域有隐喻、转喻、构式语法、象似性等。之所以出现这样的变化，跟认知语言学阵营内部很多学者认同语言认知的体验基础有直接关联。他们主张基于用法的语言学观（usage-based approach），认为语言是一种动态浮现（emergent）现象，是众多使用者频繁运用语言后固化（entrenchment）的结果。

语言类型学作为热门语言学领域，关注跨语言变异和共性，致力于对语言结构和功能规律的描写和概括。当前类型学研究还主要基于前人对各语言的语法描写。类型学研究往往要涉及几十种以上的语言，很可能是极罕见的语种，相关的大规模语料库还不具备。随着语言信息化和计算机技术的发展，语料库及技术在类型学研究中必有用武之地（参见 Liu 2010）。

语料库与词典编纂、计算语言学（机器翻译）、语言习得、翻译研究等应用语言学领域的结合，也有相当长的历史，且产出了大量重要成果。限于篇幅，此处不再赘述。

4. 结语

目前的语料库研究是在数字化时代对（美国）结构主义语言学的续写，以及对功能语言学的发扬。它有自身清晰的哲学观、语言观和坚实的学科发展历史。将语料库研究全然视作研究方法（McEnery & Hardie 2012：1），有悖史实。将它仅仅看作计算机技术，更是严重低估了其语言学地位。借用计算机领域的命名方式，我们可将 20 世纪上半叶称为语料库研究 1.0 时代，这一时期的语料库研究与（美国）结构主义语言学几乎可以划等号（另见梁茂成 2015：15）。20 世纪五六十年代至今是语料库研究 2.0 时代，它委身于形形色色的功能语言学研究之中。在大数据背景下，语料库研究 3.0 时代已现端倪，它将会全面升级过去一个世纪以来语言描写和研究的范式。在更大规模电子语料这一基本前提下，充分记录语言结构和功能特征，基于“关联数据”（linked data）模型，整合结构、语义、语境变量、语言类型学属性等信息，辅以智能查询和分析模块，从而完成语料库研究“用”“量”“器”“聚”的强化和升级。时下兴起的语言研究多因素分析法也是迈向语料库研究 3.0 时代，在“器”方面的技术革新。相信不断完善的关联数据模型必将开启语料库研究和语言学的未来。

当前语料库研究中还存在一些值得注意的现象。比如：1）语料库研究实践中存在重词汇短语、轻句法语义的情况，这种局面随着技术革新，应会得到改观。2）相关研究论文的标题中出现“语料库”字样的情况有所减少，语料库理念和方法由标题而转入正文，不再作为一种时髦的学术标签，这可以视作是学科成熟的标志。3）研究实践已经指明，有关语料库研究的本体和方法的争议已渐有共识，语料库研究者既要成为数据采集者，又要争做理论建构者。语料库研究很大程度上是“具有很强方法论导向的语言学分支”（Leech 2011：158）。

中国语料库研究者，应切实立足国情，坚持开展汉语语料库研究、汉语中介语语料库研究、中国英语学习者语料库研究和双语对比与翻译研究。语料库研

究相对没有晦涩复杂的理论，中国学者可以较快上手，甚至后来居上，为世界语料库研究的发展做出贡献。

注　释

1. Issac Pitman 爵士于 1843 年自行出版的 *The Phonotypic Journal* 上刊载了一份英语词频表。其中表 1 依字母排序，表 2 依词频降序排列。该词表基于 20 本书，每书取 500 词，共得 1 万词。以此为语料统计得到词表大致如下：the（675 次）、and（413 次）、of（396 次）……。该表完成于 1838 年，于 1843 年首次公开发表。该词表刊出后，有读者来信寄来另一项早于 Pitman 词表 20 年的词频表。可见此类工作早已有之。
2. 这里的 B.C. 是 Francis 的戏仿。B.C. 在英语中本是“公元前”（before Christ）的缩略，Francis 此处指 before the use of computers（Francis 1992：17）。
3. 据 Radding & Ciaralli（2007：35）考证，公元 528 年拜占廷帝国国王查士丁尼一世（Justinian I）诏令编修多部民法法规。1583 年，法国法学家 Dionysii Gothofredi 集印而成《民法汇编》，拉丁原文 *Corpus Iuris Civilis*，又作 *Corpus Juris Civilis*（Gothofredi 1583/1828）。另外，本节只追溯到 corpus 与语料库相关的“文本”义，不探讨 corpus 本初的“身体、尸体”义。
4. 20 世纪四五十年代，在语言学文献中已时常能见到 a corpus of material, a corpus of data 的用法。但 corpus 单用作语言研究材料汇集这一含义，大约仍以 Harris（1947）为早。
5. 目前所能查到最早使用“语料库”这一中文表述的文献为杨惠中（1981：30）。汉语中“语料”一词的使用要更早一些，至少不晚于 1951 年，参见赵元任（1951：25）。
6. 严格来讲，因为并未全面参引真实语例，并未运用统计数据，《英语语法大全》并非完全意义上的基于语料库的语法。

参考文献

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

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

Chomsky, N. 1957. *Syntactic Structures* [M]. The Hague: Mouton.

Chomsky, N. 1965. *Aspects of the Theory of Syntax* [M]. Cambridge, MA.: The MIT Press.

Cowie, A. 1998a. A. S. Hornby: A centenary tribute [A]. In T. Fontenelle, P. Heligsmann, A. Michiels, A. Moulin & S. Theissen (eds.). *Euralex '98 Proceedings* [C]. Liège: Université de Liège. 3-16.

Cowie, A. 1998b. *Phraseology: Theory, Analysis, and Applications* [M]. Oxford: Clarendon Press.

Firth, J. 1951/1957. Modes of meaning [A]. In J. Firth (ed.). *Papers in Linguistics: 1934-1951* [C]. Oxford: OUP. 190-215.

Firth, J. 1957/1968a. A synopsis of linguistic theory, 1930-55 [A]. In F. Palmer (ed.). *Selected Papers of J. R. Firth 1952-59* [C]. Bloomington: Indiana University Press. 168-205.

Firth, J. 1957/1968b. Ethnographic analysis and language with reference to Malinowski's views [A]. In F. Palmer (ed.). *Selected Papers of J. R. Firth 1952-59* [C]. Bloomington: Indiana University Press. 137-167.

Francis, N. 1992. Language corpora B.C. [A]. In J. Svartvik (ed.). *Directions in Corpus Linguistics* [C]. Berlin: Mouton de Gruyter. 17-32.

Francis, N. & H. Kučera. 1967. *Computational Analysis of Present-day American English* [M]. Providence: Brown University Press.

Fries, C. 1952. *The Structure of English* [M]. New York: Harcourt, Brace & World.

Fries, P. & N. Fries (eds.). 1985. *Toward an Understanding of Language* [C]. Amsterdam: John Benjamins.

Gothofredi, D. 1583/1828. *Corpus Juris Civilis Romani (Tomus Primus)* [M]. Neapoli: Apud Januarium Mirelli Bibliopolam.

Halliday, M. 1961. Categories of the theory of grammar [J]. *Word* 17: 241-292.

Halliday, M. 1966. Lexis as a linguistic level [A]. In C. Bazell, J. Catford, M. Halliday & R. Robins (eds.). *In Memory of J. R. Firth* [C]. London: Longmans. 148-162.

Halliday, M. 1991. Corpus studies and probabilistic grammar [A]. In K. Aijmer & B. Altenberg (eds.). *English Corpus Linguistics* [C]. London: Longman. 30-43.

Halliday, M. 1992. Language as system and language as instance: The corpus as a theoretical construct [A]. In J. Svartvik (ed.). *Directions in Corpus Linguistics* [C]. Berlin: Mouton de Gruyter. 61-77.

Harris, Z. 1947. Structural restatements: II [J]. *International Journal of American Linguistics* 13: 175-186.

Harris, Z. 1954. Distributional structure [J]. *Word* 10: 146-162.

Janda, L. (ed.). 2013. *Cognitive Linguistics: The Quantitative Turn* [C]. Berlin: Walter de Gruyter.

Kaeding, F. 1897/1898. *Häufigkeitswörterbuch der deutschen Sprache* [M]. Berlin: Self-published.

Leech, G. 2011. Principles and applications of Corpus Linguistics [A]. In V. Viana, S. Zyngier & G. Barnbrook (eds.). *Perspectives on Corpus Linguistics* [C]. Amsterdam: John Benjamins. 155-170.

Liu, H. 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks [J]. *Lingua* 120: 1567-1578.

Louw, B. 2011. Philosophical and literary concerns in Corpus Linguistics [A]. In V. Viana, S. Zyngier & G. Barnbrook (eds.). *Perspectives on Corpus Linguistics* [C]. Amsterdam: John Benjamins. 171-196.

Luhn, H. 1960. Keyword-in-context index for technical literature [J]. *American Documentation* 11: 288-295.

Malinowski, B. 1922. *Argonauts of the Western Pacific* [M]. London: Routledge & Kegan Paul.

McEnery, T. & A. Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice* [M]. Cambridge: CUP.

Palmer, H. 1933. *Second Interim Report on English Collocations* [M]. Tokyo: The Institute for Research in English Teaching, Department of Education, Japan.

Pitman, I. 1843. List of words from which grammalogues may be selected [J]. *The Phonotypic Journal* 2: 161-163.

Quirk, R. 1960. Towards a description of English usage [J]. *Transactions of the Philological Society* 59: 40-61.

Quirk, R., S. Greenbaum, G. Leech & J. Svartvik. 1985. *The Comprehensive Grammar of the English Language* [M]. London: Longman.

Radding, C. & A. Ciaralli. 2007. *The* Corpus Iuris Civilis *in the Middle Ages: Manuscripts and Transmission from the Sixth Century to the Juristic Revival* [M]. Leiden: Brill.

Russell, B. 1946. *History of Western Philosophy* [M]. London: George Allen and Unwin.

Siewierska, A. 2011. Functional and cognitive grammars [J]. *Foreign Language Teaching and Research* 43: 643-664.

Simpson, J. & E. Weiner. 1989. *The Oxford English Dictionary (Second Edition) Vol. III* [Z]. Oxford: Clarendon Press.

Sinclair, J. 1966. Beginning the study of lexis [A]. In C. Bazell, J. Catford, M. Halliday & R. Robins (eds.). *In Memory of J. R. Firth* [C]. London: Longmans. 410-430.

Sinclair, J. 1991. *Corpus, Concordance, Collocation* [M]. Oxford: OUP.

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

Sinclair, J., S. Jones & R. Daley. 1970/2004. English lexical studies: Report to OSTI on Project C/LP/08 [A]. In R. Krishnamurthy (ed.). *English Collocation Studies: The OSTI Report* [C]. London: Continuum. 2-204.

Teubert, W. 2007. Parole-linguistics and the diachronic dimension of the discourse [A]. In M. Hoey, M. Mahlberg, M. Stubbs & W. Teubert (eds.). *Text, Discourse and Corpora* [C]. London: Continuum. 57-87.

Teubert, W. 2010. *Meaning, Discourse and Society* [M]. Cambridge: CUP.

Thorndike, E. 1921. *The Teacher's Word Book* [M]. New York: Teachers College, Columbia University.

Voegelin, C. 1959. The notion of arbitrariness in structural statement and restatement I: Eliciting [J]. *International Journal of American Linguistics* 25: 207-220.

Xu, J. J. 2015. Corpus-based Chinese studies: A historical review from the 1920s to the present [J]. *Chinese Language and Discourse* 6: 218-244.

蔡廷干，1922/2014，《老解老》摘登 [J]，《语料库语言学》(2)：81-90。

陈鹤琴，1922，语体文应用字汇 [J]，《新教育》(5)：987-995。

陈　平，1987，描写与解释：论西方现代语言学研究的目的与方法 [J]，《外语教学与研究》(1)：1-15。

何安平，2013，从“意义即使用”哲学观到语料库的“意义单位”探究 [J]，《外语与外语教学》(3)：44-48。

洪　业，1932，《引得说》[M]。北平：燕京大学图书馆引得编纂处。

梁茂成，2014，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》(1)：25-33。

梁茂成，2015，梁茂成谈语料库语言学与计算机技术 [J]，《语料库语言学》(2)：15-25。

刘殿爵、陈方正、何志华(编)，2000，《颜氏家训逐字索引》[Z]。香港：香港中文大学出版社。

徐通锵，1994，“字”和汉语的句法结构 [J]，《世界汉语教学》(2)：1-9。

许家金，2014，许家金谈语料库语言学的本体与方法 [J]，《语料库语言学》(2)：35-44。

杨惠中，1981，计算机辅助词典编纂 [J]，《外国语》(1)：29-32。

杨惠中、黄人杰，1982，JDEST 科技英语计算机语料库 [J]，《外语教学与研究》(4)：60-62。

赵元任，1951，台山语料 [J]，《历史语言研究所集刊》23：25-76。

收稿日期：2016-03-11；修改稿，2016-12-03

通讯地址：100089 北京市 北京外国语大学中国外语与教育研究中心