弗里斯对语料库驱动语言学的启示※

周维江,吴建平
(厦门大学 外文学院,福建 厦门 361005)

[摘 要] 美国语言学家查尔斯·弗里斯在语言研究科学性、实证性原则指导下,是第一个尝试语料库驱动研究的人。他的研究启示当代语料库驱动语言学应审慎对待传统词类划分,在对词类体系进行完善与全面否定之间取得平衡;反思新旧语言学理论守成与创新的关系;反思库容与技术的制约效应以及对创新的期盼,在技术可行与文本可用的张力中把握机遇与挑战;反思语料库语言学研究的数据定性归属及哲学立场。

[关键词] 前电子语料库;基于语料库方法;语料库驱动语言学;观察数据
[中图分类号] H0-06 [文献标志码] A [文章编号] 1000-8284(2011)12-0168-04

1. 弗里斯基于语料库的研究

查尔斯·弗里斯(Charles C. Fries,1887-1967)与布龙菲尔德一样拒绝把直觉当作语言研究素材,坚持科学的语言研究必须源于自然、真实的会话,其出发点是反心灵主义的一元论哲学观。他的语言学理论被称作"信号语法",其核心就是把语言看作传递意义的工具,"语言是信号的任意性密码","语言不是意义本身,语言只是用来标志和传递意义的任意性对比模式系统"[1]100。他重要的学术追求之一就是语言描写及语言学理论的实际运用,可以说应用语言学是他着意的追求,而语料库的运用只是达到目的的手段与前提要求,是语言研究范式转变过程中的学术自觉。因此,他与语料库语言学多有不谋而合之处,亦在情理之中。

弗里斯一贯坚持语言研究以实证数据为基础,坚持量化的信息来自精心收集的语料库,这必然导致对具体言语调查与描写充分性、素材筛选标准严密性的重视。

弗里斯重视意义,但是他反对把意义当成句法分析基本工具的通常做法,因为那是不科学的。用意义去分析句子会使我们因懂了词义而戛然而止,无法发现传递意义的精确形式信号。所以,弗里斯句法分析时彻底放弃传统的词类划分依据和标准,一切依据语料(三百个人50 个小时的电话录音、25 万词的词项例证卡片口语库)对英语词类及其形式特征、句子的划分、句子结构模式和结构意义、句子类型、结构层次等进行纯形式的描写,来寻找标示结构意义的形式手段。他对句法范畴重新进行概括,把词类分为 4 个形式类(类似于名词、动词、形容词、副词),另外的 15 类功能词去涵盖其他词类。这里的词类划分与一般实词、虚词划分并不相同,有交叉,因为它基于词语分布、替换能力、功能以及词形因素之上。句子分成单一自由语句、扩展的自由语句和语句序列,还讨论了话轮中的语句功能。尽管在统计手段和索引行显示上存在难以克服的局限,但其自下而上的归纳法、形式标示意义的信号语法理念——(可与 Sinclair[3]7 的主张"There is ultimately no distinction between form and meaning"对比)——显示了其革新的一面,算得上是语料库驱动研究的第一次尝试①。这也符合 Sinclair[3]126 总结的语料库驱动语言学的特征:方法新颖性、语料中心性、理论变革性。

2. 基于语料库方法与语料库驱动方法对比

两种方法的差异是由于语料库驱动方法(corpus-driven approach)的提出而突显的,此前基于语料库的方法(corpus-based approach)是默认的语料库语言学学科方法。Tognini-Bonelli[4]65 把基于语料库的方法定义为:语料库的目的主要用来解释、验证、说明那些并非在大型语料库指导下形成的理论和描写,即语料库的功用在于为先前的想法(pre-corpus beliefs)[2]2 做证明和注脚。其方

※ 本文在写作过程中承蒙厦门大学杨信彰教授的指导和帮助,以及美国中密歇根大学 Peter H. Fries 教授给予资料上的便利和上海市科委"双语词典编纂系统的研发"子课题"英汉双语平行句对语料库建设"项目给予的资助,在此一并致以谢忱。

① Sinclair 的 Collin COBUILD 词典项目被认为是第一场词典学的语料库驱动尝试。

[收稿日期] 2011-10-08

[作者简介] 周维江(1968-),男,山东临清人,副教授,博士研究生,从事词典学、语料库语言学、语义韵研究。

法路径为理论/假设 – 观察 – 验证 – 解释 – 理论。语料库驱动的方法缘起于 Sinclair 为代表的 COBUILD 团队在词典编撰过程中发现大量的证据与传统的语言理论相冲突,因此不得不舍弃传统理论的描写和解释框架,一切从证据开始去发现新型式、新意义方式,建立新的概念和范畴体系的作法[4]86,[5]37。其方法论路径为观察 – 假设 – 概括 – 理论抽象。伴随语料库语言学在研究目标、哲学立场、方法论、理论和描写范畴、知识体积累等方面的发展和进步,语料库驱动语言学的建立被提上日程[3][6][7][8][9][10]。有学者指出两种方法在出发点、研究内容和焦点、研究范式上有本质性差异[10]。语料库驱动方法的核心关注是拒绝旧有脱离语言事实的理论预设,证据源于真实的语言使用,解释基于并限于文本自身,证据优先[4]74。语料库驱动语言学倡导语料库驱动方法。

3. 讨论与启示

在实证主义氛围中,依照描写主义语言学的框架,搜集整理语料、建设语料库为语言研究服务,成了美国结构主义语言学家研究活动的重要部分,其成果在语音、语法、句法方面较为显著。在弗里斯所处的时代,获得语言学学科的独立是语言学家们重要的使命,也因此约束他们并使他们养成注重方法科学性和操作程序规范性的习惯。他们相信客观的材料、科学的方法才能产生科学的结果。当前语料库驱动语言学面临类似的学科发展时期,是机遇也是挑战。弗里斯的语料库研究方法是同时代语言学家中较为突出的,可以对我们协调语言学理论革新与继承的关系提供借鉴。

3.1 对待传统词类划分的态度

尽管传统词类划分有诸多缺陷,如主要词类划分存在从词汇意义、句法功能、形式特征等不同方式加以定义的混乱[11]67,但其合理性也是显而易见的;更重要的是不管定义方式是否一致,他们在人们头脑中根深蒂固,不易更改,而且,词语需要分类的概念同样牢不可破。建立完善的新体系不容易,彻底放弃旧有体系同样不容易。弗里斯低估了旧术语的影响力,他描写、概括的形式类来取代实词类并没有流行起来。

语料库驱动语言学面临相似的选择。尽管旧的词类划分方式及其背后体现的理论有的不合时宜,尽管这些术语不堪意义的重负,同时假使语料库语言学真的发明了一整套合理的词类划分新术语,但要全面替换进来也不是一个简单的谁更具有科学性的问题。更重要的是,现有的词类判定方法整体来看是否有重大失误,还是具体词语的判别失准?Collins COBUILD 词典是语料库词典学的创新,其中有多种方式提供词类信息:语句释义提供判断语境、附加栏明示语法信息。它重要的创新在于以使用频率、以习语原则来修正词类划分的不足,采取词类与功能结合,并非全盘抛弃词类。此外,既然语料库驱动方法对意义研究和分析的核心内容包括词语搭配、类联接、语义倾向和语义韵分析等,语料库驱动研究若放弃词类标注将如何进行类联接研究?毕竟语料库语言学以词汇为切入点,以型式与意义一致关系为中心,词汇共选、共现为体现方式。换言之,处于语料库核心地位的词汇有双重身份认证:词形是形式面孔,用于搭配等研究;词类是身份证,用于类联接研究。频率计算都是以这两者为轴线进行,两者相辅相成,互为补充。如果出于对非基于语言事实旧有理论的反对而把现有词类划分也一并扔掉,那无异于把洗澡水与孩子一起倒掉,过往矫正未必促进学科进步。现在的索引和统计技术还没有达到如人所欲无需标记的程度。

词类划分问题只是如何对待前语料库时代语言理论的一个缩影。语言学学科建立前后的百年,语言理论和实践方法积累了宝贵财富,现代语言描述中的大多数概念范畴经历了时间考验,总体来讲并无大碍。完善、改进、创新是必要,全面否定是不明智的,也不大可能。

Tognini – Bonelli[4]176曾建议语料库驱动研究要摒弃脱离语言事实的理论预设,这是针对以直觉数据为对象的研究,无疑是正确的。但操作中执行标准并不清晰。哪些理论是得自于语言使用的事实、哪些不是?如果不基于语料库就一定得出错误的结论?从何时的理论算起?所有基于语料库得出的结论就都正确?如何判定?我们要放弃哪些理论?所以对待旧理论不宜一刀切。

事实上,理论绝对中立是做不到的。从最初建库的念头、原则、选材、实施、观察、研究、分析至结论归纳的全部过程中,没有理论预设和指导,没有旧有概念框架是不可思议的。"研究对象不是先于视角而给定,恰恰相反,是采纳的视角创造了研究对象"[12]8。况且一元论本身也是理论;没有理论预设,数据说不了话。由此看来,语料库驱动方法需要正确处理与已有理论的关系。

3.2 库容与检索技术问题

《英语结构》项目中,弗里斯依据 25 万词的口语库(用作通用参考库)来研究句法,这在今天看来实在不够大,对语言现象的覆盖难免挂一漏万,尽管弗里斯并未定位在对语法进行全面细致的描写,而只是句法描写的研究。现在认为语法现象研究需要至少一百万词的库容才能有比较全面的描述。由于库容与研究方式对语料库语言学的重大意义,语料库驱动研究第一要求应该归于库容,语料库"小了不美"[7]189。语料库驱动研究只有库容巨大才能弥补对平衡性和代表性的忽略,潜在的规律性才能显现,N 元组才会反复出现。

反过来看,与基于语料库方法相比,语料库驱动方法在库容上并没有多少优势可言,凡是它采用的库在基于语料库的研究那里一样可用,尽管后者并不一味追求大。加上语言使用总体的无限性,驱动研究的语料库具有更多的机遇性(opportunistic)[13]223。大语料库的一个例子就是网络,某种意义上最大的语料库,但目前开发不足,原因是语料驳杂,难于形成"对照模式"。这是大库容与代

表性的对抗。

上述研究中，弗里斯更显捉襟见肘的是采用了语料驱动的研究方式，摒弃旧有框架，一切让语料说话。在手工计数、手工完成词例索引的年代，其难度可想而知。词频的穷尽性、统计的完整性都受到挑战。从技术上而言，自然今非昔比。但是，大库容语料库带来分析难度加大的负面效应：上万行的索引行分析不可能人工完成，若使用 Sampler 又回到代表性本质概念上去。即使有了 Word Sketch 软件可以扫描词频概貌（lexical profile），需要人工概括、解释的问题实质仍在。弗里斯的研究当时因为库容和索引统计技术遭遇巨大障碍，只能部分实现研究目标，留有许多遗憾。今天，在文本获取便利、计算机存储海量、语言自动处理能力增强诸多有利条件，语料库驱动研究前途无量。

3.3 第三人称观察数据

Widdowson[14]6曾把语言数据概括为三种：第三人称观察数据（third person observed data）、第二人称诱导数据（second person data of elicitation）和第一人称内省数据（first person data of introspection），并批评语料库语言学只能研究第三人称数据。对号入座，弗里斯的研究当属第三人称观察数据。

弗里斯的科学语言观强调累积性和可预测性。累积性是指语言分析的方式、过程及假设必须明示出来，供以后的研究者检验和再检验，即科学的语言研究和分析是可重复的。分析的可靠性必然是以收集的自然语流中的语言为基础，即说话者专注于意义的表达，而不是正在使用着的语言本身。Fries 对自然发生语言语料库的坚持即反对以前旧的语言学观点的手段，也是反对后来乔姆斯基生成语法的原因之一。预测性是指科学的基础必须是对孤立现象做出预测。Fries 的“信号语法”强调“语言结构单位的形式描写使人能够有效地预测对照模式在语言社区引起的规律性的识别反应”[15]668，尤其是识别预测性的反应。因此，语料库不仅仅是说话者发出的一套形式信号，也涉及这些形式在互动交际中产生的反应。Stubbs[16]130把实证语言学的原则归结为最大客观性，最少人为干扰，重复事件的显著性。两者观点大体相同。

正是基于客观性考虑，弗里斯一再强调话语语境自然、真实、无干扰，这样的会话语言才是语言研究的对象①。这一点与有的语料库研究学者[3][10][14]不相同，他们强调“从严格的意义上说，所谓的第三人称观察数据的观察者不是真正意义上的第三者。……或说或写就成了所谓的第一人称；或听或读就成了所谓的第二人称；不说不写不听不读根本就不是语言的使用者或参与者；若要观察就不得不听或读。”[10]41如果第三人称数据可以包括或当然成为一、二人称数据，而意义理解和研究是以文本或话语为基础，“关键的手段是释义（paraphrase）”[17]30，研究者或说或写或听或读就成了所谓的第一、二人称，那么，由于数据类别混淆，无法区别什么是自然、真实语料，导致无法判定研究的性质，语料库分析留给人的印象就是“读者反应论”，语料库研究就是观察者自己的主观解读，其实证性和科学性将受到怀疑。

诚然，所有的科学研究都是由人来完成的，难免带有人的痕迹。正因为如此，有意识去控制和避免人为影响才成为区别科学研究客观与否的标准。实证研究数据定性正是以此为前提的。按照结构主义实证研究的思路，语言研究者必须是、最好是旁观的第三者身份，不干扰研究对象的活动以达到所追求的客观。值得注意的是，尽管语言研究者不可避免地参与到文本的解读之中，但研究者不等于交际中语言使用者，因为研究者和原读者分别同原作者构成两个不同的以语言为中介的交际体系之中，交际目的有根本区别，正如武打小说的读者与研究人员的阅读，其关注点或视角大相径庭。证实数据的本质是真人、真语境、真交际目的。同时，原作者或说话人不以研究者为交际对象（口语交际更是如此），研究者也不以自己研究过程中的交际活动（解读、旁听）为研究对象。仿拟 Firth 的话就是：实际的说话人说的是语言，但从来不说语言学（Actual people speak a language，but never speak linguistics）。这是第三人称观察数据与第一、二人称数据的根本区别。研究者的阐释可以成为新的科学话语进入后面语料库，成为一个新交际循环的起点，前提是它必须是基于科学、客观材料之上被接受了的结论。主观的解读是基于语料库方法所识别的客观型式之上。

此外，我们认为语言自治论[8]17，[18]238并没有拒绝原作者，只是不预设她的权威。对其意图的解读是基于文本并限于文本之中，从而揭示语言的社会功能。惟其如此方能体现 Firth[19]19“完整的社会人”（the whole man）和 Sinclair 词汇－语法－语境一体化理念。对第三人称数据的分类定位弗里斯乐于接受。

数据性质之争的背后是立场差异。Stubbs[20]一直说 Widdowson 没有跳出 competence/performance 二分法框框，因而未能理解语料库语言学的研究对象——作为社会符号的语言使用，进而把它与 Chomsky 的 performance 错误地等同起来；是“无的放矢”的“伪命题”[8]39。按照 Firth 的理论，语言使用就是语言学的唯一研究对象。我们认为，从本质上说，它也不等于索绪尔的二分法中的 parole，因为 Firth[19]192不承认其对立面 langue 的存在，因而无从谈起。诚然，不同的哲学立场，对语言研究的范畴界定不会完全一样，但 Widdowson 等人背后的关切不是全无道理的。语料库语言学主张一切基于文本，但文本被语料库语言学赋予了语言所担负的形式与意义、结构与系统、个体与社会、文内与文外（词汇－语法层体现意义

① 现在，语料库语言学已经对口语和书面语区别对待。尽管如此，弗里斯的口语语法研究直到 1999 年出版的 D. Biber 等人基于语料库的研究成果 Longman Grammar of Spoken and Written English（Harlow，Essex：Pearson Education Ltd）才真正被超越。

和语境)的一切责任。问题是文本是语言本身吗?语料库文本的上下文索引行只是语境的一种——语言语境,它能够等同于真实话语语境吗?我们应该认识到文本提供的语境是凝固的、抽象了的语境,话语实施时语言与语境复杂的相互作用并不能得到全面反映[14]7。Stubbs[20]157所说的节点词左右跨距之内的共文本(co-text)与节点词所有索引行上下行之间聚合的互文本(inter-text)构成了节点词的全部共时和历时语境并非Widdowson关注的社会语境,关键在于Widdowson不承认Stubbs等人赋予语言使用的体现形式——"文本"如此全能、无上的社会符号地位,即文本就是社会语境全息、无遗漏的投影。

语料库语言学业已发展起自己特色的视语言为社会行为,形式、意义和功能一体的概念框架,以及调查的、统计的、数据建设与数据处理的特色工作方法。发挥强项并不意味着、也不应该否定和排斥语言的心理现实性及其相关研究[17]26,毕竟,研究鸡下蛋机制的"鸡学"与研究蛋的结构及功能一体化的"蛋学"不大相同。所以,语料库研究应该以开放的胸襟,不必否认自己的弱项,"显微镜"没指望当作"望远镜"用[13]222,发挥自己学科优势,彼此借鉴。

总之,弗里斯以质朴、原始、自觉的实证主义描写语言学的研究模式进行了第一次自下而上语料驱动研究的尝试,留给当代语料库驱动语言学理论上、方法论上的启示与借鉴。

1. 鉴于现有的词类划分体系整体的有效性,鉴于词类划分在类联接、语义倾向、语义韵研究中不可或缺性,鉴于词性划分研究内嵌的目的与手段的一体化融合,鉴于词语分类在人类思维定势和认知方法中的牢固地位,对当代语料库驱动语言学来说,对具体词语判别失准的修正,对整体体系的完善、改进、创新是必要的,但是全面放弃词性划分未必是明智之举,而且其可行性也处于疑问之中。

2. 鉴于理论所具有的抽象性、概括性、预设性、指导性、自成性,鉴于理论命题的逻辑形式与语义内涵的对立与统一,鉴于新旧理论判定标准相对性和局限性,当代语料库驱动语言学应该恰当处理新旧语言学理论继承与扬弃、守成与创新的关系。

3. 鉴于文本材料获取便利性与代表性的矛盾,鉴于包括知识产权在内文本可得与文本可用的矛盾,鉴于大库容语料库具有的揭示潜力与解释难度的矛盾,鉴于索引、词频统计、语义扫描等语言自动处理工具潜力与先天缺陷的矛盾,当代语料库驱动语言学在技术可行与文本可用的张力中面临巨大的机遇与挑战。

4. 鉴于语料库语言学、结构主义语言学与实证主义长久以来的渊源关系,语料库驱动语言学研究的目标之一依然是最大限度的客观性,即,真人、真语境、真交际目的证实数据。数据性质之争折射出各流派哲学立场和学术方法的差异,各语言学分支对语言产品(即文本)的态度是整体语言观的反映:文本应该被赋予怎样的地位、如何阐释。

[参 考 文 献]

[1] Fries, Charles C. Linguistics and Reading [M]. New York: Holt, Rinehart and Winston, 1963.

[2] Sinclair, J. Corpus, Concordance, Collocation [M]. Shanghai: Shanghai Foreign Language Education Press, 1991.

[3] Sinclair, J. Progress and Prospects in Corpus Linguistics [J]. 现代外语, 2004a. (2): 113-128.

[4] Tognini-Bonelli, E. Corpus Linguistics at Work [M]. Amsterdam: J. Benjamins, 2001.

[5] 卫乃兴. 语料库语言学的方法论及相关理念 [J]. 外语研究, 2009. (5): 36-42.

[6] Hunston, S. & G. Francis. Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English [M]. Amsterdam and Philadelphia: John Benjamins, 2000.

[7] Sinclair, J. Trust the Text: Language, Corpus and Discourse [M]. London; New York: Routledge, 2004b.

[8] 卫乃兴. John Sinclair 的语言学遗产——其思想与方法评述 [J]. 外国语(上海外国语大学学报), 2007, (4): 14-19.

[9] 李文中. 语料库语言学的研究视野 [J]. 解放军外国语学院学报, 2010, (2): 37-40.

[10] 濮建忠. 语料库与语言一元化研究 [J]. 《解放军外国语学院学报》, 2010, (2): 41-44.

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

[12] de Saussure, F. Course in General Linguistics [M]. Beijing: Foreign Language Teaching and Research Press, 2001.

[13] Stubbs, M. Words and Phrases: Corpus Studies of Lexical Semantics [M]. Oxford [England]: Blackwell Publishers, 2001a.

[14] Widdowson, H. G. On the Limitations of Linguistics Applied [A]. Applied Linguistics, 2000, (1): 3-25.

[15] Fries, C. C. Structural Linguistics [A]. Encyclopaedia Britannica [Z]. Chicago: Encyclopedia, 1967.

[16] Stubbs, M. On Texts, Corpora and Models of Language [A]. In Hoey, M., M. Mahlberg, M. Stubbs & W. Teubert, Text, Discourse and Corpora [C]. London: Continuum, 2007.

[17] 卫乃兴. 语料库语言学的弗斯学说基础 [J]. 外国语(上海外国语大学学报), 2008, (2): 23-32.

[18] Sinclair, J. New Evidence, New Priorities, New Attitudes, In Sinclair, J., (ed.), How to Use Corpora in Language Teaching. [C]. Amsterdam: John Benjamins, 2004c.

[19] Firth, J. R. Papers in Linguistics. 1934-1951 [C]. London: Oxford University Press. Britannica, Inc, 1957.

[20] Stubbs, M. Texts, Corpora, and Problems of Interpretation: A Response to Widdowson [A]. Applied Linguistics, 2001b. 22(2)

〔责任编辑:曹金钟〕