语料库语言学的理论解析

许家金
(北京外国语大学 语言所 北京 100089)

摘 要:本文试图剖析有关语料库语言学的几个理论问题,以期揭示语料库语言学的本质。第一,语料库语言学是不是独立的新兴学科?第二,语料库提供的数据到底可以给语言学家带来什么?第三,语料库语言学的真正任务是什么,它应该成为怎样一项研究?回答好这几个问题实际上也就回答了语料库语言学的性质和理论地位,以及这项研究向何处去的问题。这将有助于我们更好地从事基于语料库的语言研究和实践。

关键词:语料库语言学;基于语料库的研究方法;理论架构;语料库数据;发展前景
中图分类号:H030 **文献标识码**:A **文章编号**:1000-5544(2003)06-0006-04

Abstract: This paper attempts to answer some theoretical questions of corpus-based language studies. Three theoretical considerations are addressed to capture the nature of corpus linguistics. 1) Is Corpus Linguistics an independent branch under the umbrella of linguistics? 2) What can corpus data offer to linguists? and 3) What kind of research that Corpus Linguistics should be? Answers to the three theoretical questions will conduce to a fuller understanding of the nature of corpus-based language studies and their development.

Key words: Corpus Linguistics; corpus-based approach; theoretical construct; corpus data; future directions

1.0 导言

目前利用语料库从事研究的学者主要有两类。一类是计算语言学家。他们主要从事自然语言处理(NLP)的研究,诸如语音合成、语音识别和机器翻译等等。他们的终极目标并非要揭示语言的本质,而是注重技术层面的研究,比方如何改进算法,如何完善词库的建设以实现技术上的突破。另一类就是语言学家,他们则希望借助这一强大的工具去揭示语言的本质这一学科目标,这正是本文所要关注的。

本文缘起于语言研究者中的一种争论,即"语料库语言学"是否配得上这个名称,它是一种新兴的理论视角还是"暴发户造谱牒"?而本文要探讨的正是有关于语料库语言学研究的理论意义。或者说,语言学家从语料库所提供的信息中到底可以获得什么?这是本文试图解决的核心问题。语料库建设、开发和利用在国内外语言学界日渐兴起,在国内也有相当数量介绍性和综述性文献涌现,为语料库语言学在我国的发展奠定了一定的理论基础。此外,一些语言学者和语言教师(包括外语教师和对外汉语教师)也设计、建立了一些自己的语料库以服务于教学科研。

然而真正对语料库的理论问题进行深入探讨的著述却为数不多。本文将试图剖析语料库语言学的几个理论问题,以期对语料库语言学的优势和不足有进一步的认识。第一,语料库语言学是不是独立的新兴学科?第二,语料库提供的数据到底可以给语言学家带来什么?第三,语料库语言学的真正任务是什么,它应该成为怎样一项研究?回答好这几个问题实际上也就回答了语料库语言学的性质和理论地位,以及这项研究向何处去的问题。这将有助于我们更好地从事基于语料库的语言研究和实践。

2.0 语料库语言学是不是独立的新兴学科?

2.1 语料库语言学是一种理论架构

完全赞成语料库语言学是一种理论架构的几乎没有。只是某些学者比较强调语料库语言学的理论意义。比如,Halliday(1991,1992,1993)指出,语料库语言学作为一种理论架构(theoretical construct),将语

[20] Prince, Ellen F. Toward a taxonomy of given-new information [A]. In (ed.) P. Cole *Radical Pragmatics*[C]. New York: Academic Press, 1981.
[21] Sanford, A. J. & S. C. Garrod. *Understanding Written Language*[M]. Chichester: Wiley, 1981.
[22] Schank, R. C. & R. Abelson. *Scripts, Plans, Goals and Understanding*[M]. Hillsdale, N. J.: Lawrence Erlbaum, 1977.
[23] Shanon, Benny. What is in the frame? — Linguistic indicators [J]. *Journal of Pragmatics*, 1981(5):35-44.
[24] Sidner, Candace L. Focusing and discourse[J]. *Discourse Processes*, 1983(6):107-130.
[25] Sperber, D. & D. Wilson. *Relevance: Communication and Cognition*[M]. Oxford: Basil Blackwell, 1986.
[26] 桂诗春. 新编心理语言学[M]. 上海:上海外语教育出版社, 2000.

作者简介:王军,山东聊城大学外语学院副教授,现在是上海外国语大学英语学院在读语言学及应用语言学博士生,研究方向为英汉语言对比、间接回指等。

收稿日期 2003-03-18
责任编校 薛旭辉

料收集和理论概括统一了起来，从而使我们对语言的理解产生一种质变。这种新的理论架构有助于考察同时作为系统和实例(instance)的语言的本质。因为在 Halliday 的语言学思想当中，实际话语是语言系统的实例再现(instantiation)。而语言系统，或者说是语法体系是一种统计概率(probabilistic)上的自然结果。这一思想与所谓语言学规则是浮现特征(emergent properties)的说法颇为暗合(李平，2002)。也就是说，因为严格设计并创建的语料库所包含的应该是真实文本和真实话语，其中语言实例在出现频率上的优势即是对其背后语法体系的概率体现。另外，我们知道 Halliday 功能主义思想中的一个重要概念就是"意义的选择"(Halliday，1985)。这种意义的选择反映了语言运作的内在机制。语料库辅之以计算工具，便可以将这些机制进行抽象概括从而形成语法。

这里特别值得一提的是，上述思想是与 Chomsky 的心灵主义相对立的。Chomsky 历来认为语言是一种天赋能力，而自然语料都是杂乱无章的。其中包括很多显然不会出现的，或者错误的句子，还有很多诸如迟疑，注意力的不集中和外界的干扰等等。所以他主张我们研究的应该是理想的听话人/说话人的语言能力(Chomsky，1965)。因而 Chomsky 提倡通过内省和诱发的手段来获得语言资料，而反对使用语料库进行语言研究的。

2.2 语料库语言学是一种基于语料库的研究方法

然而，尽管众多语言学家承认语料库对语言研究的巨大贡献，他们并不认为语料库语言学像语言学其他分支一样成为独立的学科领域。

Tognini-Bonelli(2001)对语料库语言学的性质进行了重新思考。她指出语料库语言学并不是一个真正意义上的科学研究领域，只不过是为语言研究提供了一种方法论基础，同时它又给语言学的研究提供了新的哲学思路。所以它是介于理论和方法论之间的一种东西。应该说她的这一观点代表了相当多语言学家的看法。

比如，Leech(1992:105)说过，"……[语料库语言学]倒是更应该被看作是从事语言研究的一种方法论基础。理论上(而且常常在实践当中)语料库语言学与其他语言学分支轻松结合：我们能够借助语料库研究语音学，句法……。"

Leech(1992:105-6)在这里明确指出了语料库语言学的工具性和方法论价值。但同时他又表示：

"语料库语言学不仅界定了一种研究语言的方法论，……而且事实上界定了该项研究课题的一些哲学/理论视角。"

换言之，借助于语料库语言学所提供的方法，语言学家一方面可以验证由内省得到的语言规则，也可以基于语料库提供的数据推演出语法、语用规则。由于不同类型的语料库和丰富的标注手段的出现，语料库语言学使我们的研究视野更加开阔，研究手段也愈加强大了。

综上所述我们认为，基于语料库的研究方法(corpus-based approach)这一提法倒是更能准确地反映语料库语言学的性质和定位。

3.0 语料库提供的数据到底可以给语言学家带来什么？

明晰了语料库语言学的理论定位之后，接下来我们要看一看语料库数据真的能反映语言的本来面貌吗？从现有的语料库的规模、类型来看，语料库所提供的数据有以下特点：

3.1 产品/过程对立问题

从某种意义上讲，语料一旦入库，它所记录的便是语言的产品(product)而不是语言的过程(process)。虽然在定义"语料库"的时候，总少不了提到"自然语言"和"真实文本"这样的概念。但是，实际现场即席话语中的很多鲜活的内容(如：情景语境和文化语境)入库之后即不复存在。毕竟多数语料库所记录的都是文本的或少量的声音信息。由大规模的多媒体数据构成的语料信息还很少见。一则是数据收集的工作量巨大；更重要的是在实际操作中，一旦进行录像，就难免会引起受调查者主观上的注意，从而影响语料的信度和效度。因此在尽可能多地收集多媒体语料的同时，还要认识到收集语料的局限和现有语料的先天不足。面对已有的语料，要想真正研究语言的本质和实际运作，还需借助诸如句法学、话语分析、语用学、社会学、人类学、民俗学等其他理论手段，对语言使用的真实状态进行描写，接近其本质特征。

3.2 取样范畴和代表性问题

毋庸置疑，所有语料库建库人都力图使其创建的语料库足以代表或反映其所要研究的目标语域或整个语团的语言事实。因此在创建初期都会对语料库的设计、取样进行科学的分析(Biber，1993；Biber，1994；Greenbaum，1991；Nelson，1996)。但是有一点我们必须认识到，我们无论如何也无法穷尽"某种语言的全体使用者说出来(或写下来)的和尚未说出来(或写下来的)所有话语"(顾曰国，1999:3)，因为它是一个开放集。我们是无法真正捕获理论上的语言的全貌的。因而为了尽可能地(至少在统计上)反映语言的实际状况，取样的方法在一定程度上可以满足我们研究的需要。与之紧密关联的一个重要概念是"代表性"问题。也就是说所收集的语料是否可以在统计上代表各种类型的真实话语。此时，语料库的大小绝不是一个关键问题。因为在现有技术条件下，一个人可以在数小时之内收集数以亿计(词次)的电子文本。当

然，真实的口语语料的收集则要困难得多。

另外，语料类型的代表性还应与研究需要紧密结合。建库人可以根据需要收集某一语域的口语或书面语的共时语言实例；也可以收集该语域历时的语言实例（比如：赫尔辛基英文文本语料库历时部分）；还可以建用于翻译或对比研究的双语或多语的平行语料库；还有像国际英语语料库（ICE）那样的某一语言的不同变体之间的语料库；还有研究一语、二语或外语学习者（口语、笔语）语料库等等。从理论上来讲，只要我们按照严格的统计取样（辅以前期实验性取样并验证）的办法去收集，就可以获得我们所想要的具有足够代表性的语料。

3.3 有无多重标注和强大的处理工具

语料库标注的好坏，类型的多寡和有无适合专项研究的处理软件很大程度上决定了语料库的有用程度。自从最早的计算机化的 Brown 和 LOB 语料库进行了 POS（Part-of-Speech）标注之后，利用这两个语料进行研究的成果激增。当然我们也注意到迄今为止，只有词汇层次的标注较为成熟，基本上可以实现正确率很高的自动标注。这也就是为何在利用语料库所从事的研究中，词典编纂以及相关的教材开发方面的成果最为显著。

其实语料库标注可以被理解为一种元语言形式。它是对原始语料进行一种初步的静态的注解。词性标注是这样，某些句法层的标注也是这样。曾经还听说有人将所有语料进行主位、述位的标注，以便对其施行系统功能进行分析。因此如果对口语语料再进行音段和超音段的标注，那将会对话语的动态分析带来极大方便（参见：Chafe, 1993; Chafe *et al*, 1991; Du Bois *et al*, 1993）。

说到语料库的处理工具，主要有转写工具（transcriber）、检索工具（concordancer），对齐工具（aligner）和其他一些统计工具等等。这些计算机程序的出现使语料库语言学的定性与定量相结合的方法成为可能。因为一方面它可以根据语言学家的研究需要，生成针对某一种语言形式的数量信息或者分布状况。目前已经有很多比较成熟的语料库工具软件。但是，为了更好地服务于研究的需要，设计符合研究者自身要求的专项软件非常必要。

3.4 异例与不规范语言实例的处理

这是一个极其重要的问题，也是生成语法学派重点攻击的一个方面。Chomsky（1965）指出："语言学理论主要关心的是理想的听话人（说话人），他们处在完全同质的语言社团中，……他不受与语法无关的条件的影响，如记忆力的限制、外界的干扰、注意力和兴趣的转移和口误……"

而语料库所提供的数据因为要体现其真实性（authenticity），所以遵循的原则是尽可能将实际话语原原本本地记录下来。因而说话当中的迟疑、口误都被录入语料库。那么，如此"泥沙俱下"的语料是语言学家所要追求的目标吗？对此，有些学者认为这正是语言的丰富性之所在，反映了语言的本来面目。这一点在理论上似乎无可厚非，然而在实际话语中，支离破碎的句子在出现频率上要远高于语法上规范的句子（Aarts, Jan, 1991: 58），那么这是否意味着我们应该忠实地将这些破碎的甚至是错误的句子写进语法书，或是口语教科书，并告诉孩子们、学生们我们说话时应该如此呢？显然这不是一个理性的处理办法。语法作为一种规范或者一种准则，必须进行一定的抽象和概括。而对于传统语法所忽略的一些话语在动态使用中的一些特征和机制当然也应当补充进去。

综上所述，要想获得理想的语料库数据，我们就应该采取科学的取样方法，尽可能囊括各种类型的口头和书面话语，包括各种语体，语言的各种变体，对语料进行尽可能多的适合不同研究需要的标注。对于口头话语的记录，既要有文字形式的标注，还要有语法、韵律层次等的标注（比如学习者的中介语语料库，就需要错误类型标注）。而且，在均衡取样的情况下，语料库规模应该越大越好。

4.0 语料库语言学的真正任务是什么，它应该成为怎样一项研究？

语料库语言学应该成为怎样一项研究呢？基于前文所讲的语料库的优势和不足，我们认为语料库语言学应该强化其定量定性相结合方法。这就需要语料库语言学家在掌握计算、统计这样的定量方法的同时，注意充分地运用理论语言学方面的最新成果，弥补定量分析的一些弱点。这就对语料库语言学家的素质提出了要求。此外，我们知道基于语料库的语言研究最终是为了解释语言，那么，语料库语言学对语言学会产生什么样的影响呢？

4.1 语料库语言学家的职责

语料库语言学要想真正揭示语言在实际使用中的情况，首先需要有前文所提到的理想的语料库以及相应的工具软件和一整套适合语料库语言学的语言研究理论。这一重任就落到了语料库语言学家们的身上。他们应该是怎样的一种身份或者具备怎样的素质才算比较理想呢？1）语言学家＋计算机科学家；2）具备语言学理论知识的计算机科学家；3）具备计算机知识的语言学家。一般看来，应该是第一种组合，而事实上我们认为第三种才是一种比较切合实际的要求。首先，语料库语言学的研究对象是语言；其次，研究中对计算机知识的要求并不是很高，具备计算机知识的语言学家只需要知道哪些功能在技术上能实现就可以了，编程的工作交由专门的技术人员去完成便可以了。对语言的本质的研究应该是语言学家的终极目

标，语料库语言学家自然也不例外。他们应该在前人近百年的研究的基础上更进一步，完成很多语言学家因为技术手段上的制约想完成却无法完成的任务。

4.2 语料库语言学的语言学地位

自 1961 年最早的 Brown 计算机化语料库的诞生，到了上个世纪八、九十年代语料库语言学掀起的一股热潮传播至今。那么它对语言学的发展到底会产生怎样的影响呢？虽然如前文所述，语料库语言学为语言学的研究提供了新的哲学思路，但我们认为它在语言研究方法论上的意义更加深远。它使我们有办法利用语料库提供的数据将我们对语言规律的朴素的认识(folk beliefs)，上升为语言学理论(linguistic theory)；或者用这些数据来纠正我们常识中对于语言规律理解的种种谬误。这种通过定量定性相结合的方法得出的结果更加有说服力。从目前语料库语言学的发展来看，它还无法获得与其他经典语言学的领域相当的地位。至于将来随着计算机技术和语料库研究方法上的发展，语料库语言学能否为传统语言学带来一些重大变革还将有待时间的检验。

5.0 结语

从前文的分析中我们不难看出，语料库语言学是一门语言研究和相应的计算机技术相伴相生的产物，是语言研究中定量和定性方法相互结合的典范，为揭示语言的本质做出其应有的贡献。虽然说语料库语言学是研究语言的一件强有力的武器，但同时我们必须对它的优势和不足有清醒的认识，它绝不是包医百病的良药。最大限度地开发利用语料库资源，并且充分结合其他相关学科的理论成果，才是对待语言研究的科学的和审慎的态度。

参考文献

[1] Aarts, Jan. Intuition-based and observation-based grammars[A]. In Aijmer, K and B. Altenberg (eds). *English Corpus Linguistics: Studies in Honour of Jan Svartvik* [C]. London and New York: Longman, 1991: 44-62.

[2] Aijmer, K and B. Altenberg (eds). *English Corpus Linguistics: Studies in Honour of Jan Svartvik* [C]. London and New York: Longman, 1991.

[3] Armstrong, Susan (ed). *Using Large Corpora* [C]. Cambridge, Mass: MIT Press, 1994.

[4] Biber, Douglas. Representativeness in Corpus Design[J]. *Literary and Linguistic Computing* ,1993(4): 243-57.

[5] Biber, Douglas. Using register-diversified corpora for general language studies [A]. In Armstrong, Susan (ed). *Using Large Corpora* [C]. Cambridge, Mass: MIT Press, 1994: 179-202.

[6] Chafe, Wallace John W. Du Bois and Sandra A. Thompson. Towards a new corpus of Spoken American English[A]. In Aijmer, K, Altenberg, B. (eds). *English Corpus Linguistics: Studies in Honour of Jan Svartvik* [C]. London and New York: Longman, 1991: 64-82.

[7] Chafe, Wallace. Prosodic and functional units in language[A]. In Edwards, Jane A., Lampert, M. D. (eds). *Talking Data: Transcription and Coding in Discourse Research* [C]. Hillsdale, NJ: Lawrence Erlbaum,1993: 33-43.

[8] Chomsky, Norm. *Aspects of the Theory of Syntax* [M]. Cambridge, Mass: MIT Press, 1965.

[9] Du Bois, John W. Stephan Schuetze-Coborn, Susanna Cumming, and Danae Paolino. Outline of Discourse Transcription [A]. In Edwards, Jane A., Lampert, M. D. (eds). *Talking Data: Transcription and Coding in Discourse Research* [C]. Hillsdale, NJ: Lawrence Erlbaum,1993: 45-89.

[10] Edwards, Jane A., Lampert, M. D. (eds). *Talking Data: Transcription and Coding in Discourse Research* [C]. Hillsdale, NJ: Lawrence Erlbaum, 1993.

[11] Greenbaum, Sidney (ed). *Comparing English Worldwide: The International Corpus of English* [C]. Oxford: Clarendon Press, 1996.

[12] Greenbaum, Sidney. The Development of the International Corpus of English[A]. In Aijmer, K and B. Altenberg (eds). 1991: 83-91.

[13] Halliday, M. A. K. Corpus studies and probabilistic grammar [A]. In Aijmer, K and B. Altenberg (eds). *English Corpus Linguistics: Studies in Honour of Jan Svartvik* [C]. London and New York: Longman, 1991: 30-43.

[14] Halliday, M. A. K. Language as system and language as instance: The Corpus as a theoretical construct [A]. In Svartvik, Jan (ed). *Directions in Corpus Linguistics. Proceedings of Nobel Symposium* 82, *Stockholm*, 4-8, *August* 1991[C]. Berlin and New York: Mouton de Gruyter, 1992:61-77.

[15] Halliday, M. A. K. Quantitative studies and probabilities in grammar[A]. In Michael Hoey (ed). *Data, Description, Discourse: Papers on English Language in Honour of John McH. Sinclair* (on his sixtieth birthday) [C]. London: Harper Collins, 1993: 1-25.

[16] Halliday, M. A. K. *An Introduction to Functional Grammar* [M]. (2nd Edition). London: Edward Arnold, 1994.

[17] Leech, Geoffrey. Corpora and theories of linguistic performance [A]. In Svartvik, Jan (ed). *Directions in Corpus Linguistics. Proceedings of Nobel Symposium* 82, *Stockholm*, 4-8, *August* 1991 [C]. Berlin and New York: Mouton de Gruyter,1992: 105-22.

[18] Nelson, Gerald. The design of the corpus[A]. In Greenbaum, Sidney (ed). *Comparing English Worldwide: The International Corpus of English* [C]. Oxford: Clarendon Press, 1996: 27-35.

[19] Svartvik, Jan (ed). *Directions in Corpus Linguistics. Proceedings of Nobel Symposium* 82, *Stockholm*, 4-8, *August* 1991[C]. Berlin and New York: Mouton de Gruyter, 1992.

[20] Tognini-Bonelli, Elena. *Corpus Linguistics at Work*[M]. Amsterdam and Philadelphia: John Benjamins, 2001.

[21] 顾曰国. 使用者话语的语言学地位综述[J]. 当代语言学，1999 (3).

[22] 李 平. 语言习得的联结主义模式[J]. 当代语言学，2002(3).

作者简介：许家金，北京外国语大学教授，主要研究方向为话语分析、语用学、语料库语言学、计算机辅助语言学习。

收稿日期 2003-01-06
责任编校 王和平