黄曾阳先生语料库思想概述*

池毓焕

（中国科学院声学研究所，北京 100080）

E-mail: chiyuhuan@hotmail.com

摘要： 本文是写作中的《HNC 语料库语言学》之第一章略加修改而成，全面介绍了黄曾阳先生关于建设 HNC 语料库的设想，重点阐述了语料库建设必须接受语言学理论指导的思想。

关键词: 黄曾阳; HNC 理论; 语料库

引言

黄曾阳先生有关语料库的论述散见各处。比较集中的论述有：写于 1993 年冬的《HNC 理解处理问答》（以下简称《问答》）、写于 1998 年 5-7 月的《HNC 理解处理的 52 个论题》（以下简称《论题》）和写于 1999 年 9 月的《句类分析的 20 项难点》（以下简称《难点》）。

粗看之下，读者难免产生 HNC 理论与语料库语言学相冲突的印象，即“先生对统计的方法似乎基本持否定态度”（《问答 6》之问，见《HNC 理论》p280）；但具体情况则复杂得多，需要具体分析。

1 《问答》：道不同，但期望仍殷切

《问答 6》明确阐述了 HNC 理论与语料库派的理论出发点之迥异；但出于扶助 HNC 理论及句类分析技术能茁壮成长的良苦用心，仍殷殷期望语料库建设大有作为。

1.1“西语需要通过语料库统计出语词的相关性，而汉语的词库就是这一相关性相当完备的表达。”“至于效果显著的常用搭配知识，本来就存储在你的大脑皮层里，通过词典的诱导，可以比较完整地检索出来，不必借助语料库和统计手段。”

语料库派与描写语言学可谓一脉相承，其思想基础是经验论，关注对语言事实的客观描述；而乔姆斯基的思想基础是唯理论，断定语言能力是人脑所固有的，对语言的研究完全可以借助于内省法。上述引文与乔姆斯基的论断如出一辙。当然，其特定背景是汉语的“字义基元化，词义组合化”现象，因此有汉语的词库蕴含着概念联想脉络一说。

1.2“语料库对于语言个性的揭示，无疑有较大意义。但我们当前工作的重心不应该放在个性方面。”因为 HNC 理论定位于自然语言的理解，而“个性对语言生成或表达至关紧要，所以，有些机器翻译学者

*本项目得到 973 项目（G1998030506）的支持。

不大赞成搞什么语言理论模型，是可以理解的。但个性对于理解和解模糊的影响是比较小的。”“如果你的目的是寻求词与词之间的常用搭配，当然可以利用语料库。然而，常用搭配主要是概念关联性的体现。那么，为什么不从建立概念关联性理论模型这一根本问题着手？”

这段话首先肯定了语料库的作用，但强调HNC理论的出发点是建立处理语言共性的“概念关联性理论模型”。共性处理有以一当十、或“以有限的模型应对无限的语言现实世界”的增效作用，个性问题则可以引入例外处理机制。由此确定：通过语料库统计语言知识之路不是HNC理论的首选。

1.3 “从理解工作的现阶段状况来说，急需一些语法现象的统计知识，可惜得很，词频统计里没有反映。现有的词频是无条件概率知识，而一些条件概率知识更有实用价值。”“汉语的常用字绝大多数有多个义项，但不同义项的运用方式，在现代汉语里却有明显差别。有的义项可以用该字独立表达，有的义项则不能，需要与其他有关的字联合起来，少数甚至不能独立表达。……。因此，在说明字义时，要引入‘独立性’的概念，在各义项里给出相应的标志，这对于新词及单音词的辨认极有参考价值。”“字的总频度与其独立义项的频度没有必然的联系，总频度高的字，其独立义项的频度可能很低。”“理解处理不仅关心字的独立义项的条件频度，也关心字在句首或句尾出现的条件频度。”“以上所说的各项条件频度，当前都没有可靠的资料。我们不得已自己动手做了一点统计。所以，我们是以自己的认识来推进统计工作的。”

如果说前一段是从理论方法论角度区隔HNC理论和语料库学派，这一段则表明了HNC对现有的语言统计工作有一种“恨铁不成钢”的失望和焦虑，以及HNC对语料库的特殊需求。

这还仅仅是在字词层面对语料统计提出了殷切期望。至于句子或篇章层面的HNC语料库建设，当时也有了比较全面的设想。

1.4 “理解系统的完善化，必须经历从简单到复杂，从低级到高级的学习过程。还是让人先来充当妈妈，扶着这个刚学走路的婴儿吧！”“基于扶婴儿走路的观点，我对于语料库有极大的兴趣。需要设计多种多样的语料库，供未来的理解程序使用和学习（检验）。”

也就是说，HNC语料库的首要使命是提供语言仿真环境供理解处理程序使用和学习，一方面通过人工标注的分析结果比较和测试理解程序以图提高，另一方面通过对精选的真实语料进行研究，提取相应的语言知识“供应”给理解程序。正是基于这一点，《论文》系列之十二要专门论述“理解处理的环境仿真”。可惜该文刚开了个头，未写完。这一论文的完成实有赖于“仿真环境”的建成，即HNC语料库及其相应处理和仿真程序的完工。一个具有充分代表性的“仿真环境”实际上构成一个语言研究的公共平台。

1.5 “这种语料库不是语料的堆积，要配备一系列的程序，以配制各种类型的语料。打个比方来说，语料库里不仅应具备山珍海味的好原料，更要配备优秀的厨师，否则做不出美味佳肴来。”“但更重要的是‘厨师’程序。这些程序分为两大类。一类是把原始文本变成各种模糊文本。第二类‘厨师’程序是把原始文本变为各种标准分析文本，基本的分析内容包括：语义块切分、要素判定、语句类型判定、伴随成分判定、语境判定、主题判定等。这些分析工作都基于文字文本来进行，当然这些都是理解本身或其预处理程序，但这些程序最终应纳入语料库的‘厨师’程序库。”

这段引文是对上述“仿真环境”设想的具体化。因为当时的主攻方向是语音文本的理解，连语料的存储都要求由国标码改成音序码；当前主攻文字文本，对第一类厨师程序可暂不考虑。所谓第二类厨师程序，实际上明确了从HNC的需求角度看，熟语料库应标注的内容。除了“伴随成分判定”一说已改称“辅块辨识”外，均应列为HNC语料库的标注项目。后来在《难点》中又提出了分别标注20项难点等新内容，其出发点还在“理解处理的环境仿真”。

从“这些都是理解本身或其预处理程序，但这些程序最终应纳入语料库的‘厨师’程序库”可以看出，当时已经产生了在语料库中融入人工分析和机器分析的成果并让二者交互作用以提高理解程序的能

力的思想。

1.6 “语料库的选材要有全面性和代表性。当然也要注意到重点和雅俗兼有等等。”“从文体来说，四种题材的语料库都要选一些，首先是叙述型语料，纵的方面包括报导、传记、历史、游记等等，横的方面包括叙境、叙事、叙人，其次是论述性语料，纵的方面包括政论、论文、专著，横的方面包括论事、论理、论人。最后是抒情和对话，这主要是从小说取材。”

可以说，语料的代表性是语料库建设的关键问题，直接关系到统计结论的可信度。如果代表性问题解决得好，就能抑制对样本规模的无限膨胀的追求。

从当时的设想看，HNC 语料库还是应该把代表性作为建库目标的，因为代表性是语言研究公共平台的应有之义。

总而言之，《问答》时期对语料库的期望主要是从理论印证和软件测试的角度提出，是一种“扶婴儿走路”的态度，对“大规模真实文本”并没有什么奢望。

2 《论题》：仍关注，重点却是全面提升知性主义

《论题 11-1》写道：“计算语言学必须把自己的立足点转过来，端正主攻方向。在这一转变中，西方语言学的语法传统是一块绊脚石，而所谓的语料库语言学则是一块误导的路标。对语法和语料库的所能和所不能要有一个清醒的认识。”（《HNC 理论》p204）——注：黄先生在《难点》中坦言“这段话有完全否定语料库语言学的语病，容易引起误解”。

《论题 7》则写道：“概念的抽象具体之分及概念节点的系统设计，语义块的主辅之分及语义块基元的系统设计，基本句类的划分及其一级子类的系统设计都存在两可问题，正是处理这一系列两可疑难的艰辛使我对利用机器可读词典寻求语义原语的效果深表怀疑，因为问题的要害不在于寻求最小原语数量，不是一个简单的义项归并问题，而是要寻求自然语言概念体系的理论模式，这里需要深刻的创造性思考，统计方法只能提供工具性帮助，不可能替代思维的创造性作用。”

这两段引文还是表白了对语料库学派的基本态度：对只要归纳不要演绎、只要经验不要理论的倾向深表怀疑。

“在处理上述一系列两可疑难的过程中，曾试图引入疑难度的概念予以定量表述，后来放弃了。因为，这需要统计工具的支持。我虽然不赞成语料库学派的大方向，但我一直关注他们的方法学成果，HNC 理论和技术的发展不久的将来必须在 HNC 知识库和 HNC 语料库的基础上开展这方面的研究。”

在这里，既否定又期待之情跃然纸上，重点表达了尽快建设 HNC 语料库使之成为 HNC 语言学研究的基石的愿望。但《论题 7》最重要之处在于全面阐述了可称之为“知性主义”的 HNC 理论的方法论基础——这个基础不但告诫如何正确对待语料库语言学，还将指导如何从事基于语料库的语言学研究。

知性与理性的区别不是通常的意义，而是康德《纯粹理性批判》中使用的概念，即理性指知其然而不管其所以然，知性指既知其然还要知其所以然。

“为什么要提出语义块的主辅之分？为了建立语句的数学和物理表示式，使语言变成一个 well-defined 的东西。这些表示式就是计算语言学界所孜孜以求的语言模型之一——句子层面的语言模型。”“关于这个模型问题，可以说存在两种态度，一种是得过且过，在短语结构模型的基础上修修补补，不触动它的根本缺陷，希求通过受限的约束避开语言的种种不规范现象，也就是避开对语言本质的探索。另一种是相信乔姆斯基关于自然语言是一个 ill-defined 的东西的说法，脑子里存在大量比喻的和夸张的，乡土的和诗歌的，儿童的和怪诞的例句，并为之困扰而不知自拔，不相信对自然语言的表述可以出现牛顿力学对力学现象或麦克斯韦方程对电磁现象的突破，但是他们不曾想过，如果当年牛顿不是专注于天体的运动，而是专注于羽毛在狂风中的飞舞，麦克斯韦不是专注于电磁场在自由空间中的一般规律，而是专注于方孔的衍射，他们也将一事无成。在建立自然语言模型这一重大探索中，必须紧记在所必为

和有所不为的辩证法，并深思康德的下列两段名言：

理性必须一手拿着原则，拿着那些唯一能使符合一致的现象成为法则的原则，另一手拿着自己按照那些原则设计的实验，走向自然，去向自然请教，但不是以小学生的身份，老师爱讲什么就听什么，而以法官的身份，强迫证人回答他所提出的问题。

自然的最高立法必须是在我们心中，即在我们的知性之中，而且我们必须不是通过经验，在自然里面去寻求自然的普遍法则；而是反过来，根据自然的普遍的合法性，在存在于我们的感性和知性里面的经验可能性条件中去寻求自然。”

从方法论角度看，HNC 语料库要“紧记在所必为和有所不为的辩证法”，不要“脑子里存在大量比喻的和夸张的，乡土的和诗歌的，儿童的和怪诞的例句”，过分追求语料的规模而误以为规模就是代表性，不知一手拿着原则一手拿着实验走向自然。也就是说，要在 HNC 理论的视野中去建语料库、使用语料库并在语料库中得到求证。

3 《难点》：详尽列出 HNC 语料库大纲

《难点》的副标题是《“corc4-3”后记》，也就是黄先生为亲手标注中文语料写的后记，而且专辟第三章写了“HNC 的 CORPUS 观”，可见该著与语料库密切相关。

在第三章中，先对语料库进行了科普式的介绍：“CORPUS 本身没有任何神秘，就是语言或语音资料的电子文本。”

“计算机擅长统计，取得与统计有关的语言知识是 CORPUS 的特长。但是，自然语言处理特别是理解处理所需要的知识，有哪些与统计有关？大脑的语言感知过程利用语言的哪些统计知识呢？甚至可以探问，它利用统计知识么？还应该进一步这样提出问题，语言是随机过程么？如果是，它又是一个什么性质的随机过程呢？”

“统计有各种各样的方式，如何选定统计方式？如何选定条件概率的条件？如何运用采样原理？如何范定大规模真实语料的‘大’？”

“对第一类问题，有两个流行的答案。一个是语词的共现概率，一个是隐马尔科夫过程。对第二类问题，仅对‘大’的规模作过一些探讨。两个流行答案实际上不是直接针对相应科学问题的研究答案，只是一种现成数学工具的运用。从这个意义上说，CORPUS 语言学还没有取得‘学’的资格。”

“CORPUS 的意义不在它本身，而在于如何利用，在于明确：能够从 CORPUS 得到什么知识，不能得到什么知识；能够得到的知识对自然语言理解和生成能起什么作用；什么知识是与 CORPUS 根本无关的，什么知识的获得是可以甚至是必须得到 CORPUS 帮助的。在未明确并基本解决这些理论问题之前，以‘摸着石头过河’的方式采取大规模行动，是无理论指导下的典型盲动。而盲动的研究工作，严格说来，是没有资格称为‘学’的。”

“下面列出与上述问题有关的清单，简称 CORPUS 期望知识清单。”

※ 1 与 CORPUS 无关的概念层面知识

语义网络的宏观构架

基本句类表示式和语义块构成表示式

语句格式知识

基本句类知识的主体

※ 2 与 CORPUS 有关的基础知识

复句与非复句的比例

这一比例与语种和文体的关系（下同，不列）

基本句类、混合句类、复合句类的比率

这一比率与语种和文体的关系（下同，不列）

无分析难点语句与有分析难点语句的比例

20 项分析难点的各自比率

无生成难点语句与有生成难点的比例

6 项生成难点的各自比率

汉语不带上下装的全局特征语义块与带上下装者的比例

英语有名无实的中心动词与形实相符者的比例

※ 3 与 CORPUS 有关的策略研究知识

按难点类型划分，当务之急是分析复杂句蜕块难点

和生成的语义块构成变换难点

※ 4 倚仗 CORPUS 的 HNC 知识库栏目

词语（主要是动词）句类代码

特征语义块构成知识

语义块要素关联性预期知识

体词的语义块构成知识

语言逻辑概念反映射词的语用知识

各类小专家的自给知识库建设

多句类代码动词或词组的语用知识

语义网络概念节点之间的交式及链式关联知识

反映射知识库的建设

※ 5 CORPUS-based 研究平台

基本语境知识框架研究

背景知识框架研究

情景与势态知识框架研究

要点主题分析研究

各种自然语言处理方案的潜力研究

此大纲一出，HNC 语料库语言学的理论问题基本阐述清楚了。实际上，前期有关词语层面如何利用基于语料库的统计知识、为语言研究建立一个统一的研究平台等思想不但包含在这个大纲中了，还具体化了——这既是黄曾阳先生长期深思熟虑的结果，又与 HNC 理论的逐渐完善和成熟分不开。

既然列出这么多“CORPUS 期望知识”，也就是说存在如此之多目前的语料库研究成果尚无能为力的方方面面，从一个侧面说明了黄先生此前对语料库学派的批判的论据。另一方面，更期待在上述理论指导下的 HNC 语料库及其研究尽快开展起来，结出丰硕的果实，让事实说话才更有说服力。

参考文献:

[1] Graeme Kennedy. An Introduction to Corpus Linguistics. Addison Wesley Longman Ltd, 1998.

[2] Douglas Biber, et al. Corpus Linguistics. Cambridge University Press, 1998.

[3] Woods, A., et al. Statistics in Language Studies. Cambridge University Press, 1998.

[4] 黄曾阳. HNC（概念层次网络）理论. 清华大学出版社, 1998.11.

[5] 黄曾阳. 句类分析的 20 项难点. HNC 自然语言理解处理网站:http://www.hncnlp.com/

作者简介: 池毓焕（1967－），男, 福建尤溪人, 博士生, 主要研究领域为语料库, 语序对比研究.

An Introduction to Prof. Huang Zengyang's Thoughts on Corpus Linguistics *

CHI Yuhuan

(Institute of Ascoustics, CAS, Beijing 100080, China)

E-mail: chiyuhuan@163.net

Abstract: It's adapted from the first chapter of a unfinished dissertation titled 'HNC Corpus Linguistics'. There are much misunderstandings of Prof. Huang's thoughts on Corpus Linguistics that are complicated and disputable and a deep analysis is necessary to attain a full-scale comprehension of it. Above all, the article focus on the viewpoint that the construction of any corpora should be guided by one's own theory of linguistics. It's imperative now to build a large-scale HNC corpus.

Key words: Zengyang HUANG; HNC Theory; Corpus

* Supported by the 973 Project under Grant No.G1998030506