现代汉语研究语料库系统的总体设计*

孙宏林

(北京语言学院语言信息处理研究所，100083)

摘要：本文讨论了建立现代汉语计算机语料库的必要性，并从文本的选择、加工和检索等方面具体探讨了建设一个200万词规模的汉语句法语料库的原则和方法，该语料库将经过分词、词性标注和句法结构标注的处理。

Abstract: This paper discusses the necessity for building large computerized corpora of Chinese and presents some considerations on constructing one such large annotated corpus which consists of 2 million words of contemporary Chinese and will be annotated for word-segmentation,part-of-speech and syntactic structure from such aspects as text selection, annotation and data retrieval.

§1 目的与意义

在语言研究工作中，语言学家往往把大量的时间都花费在语言材料的搜集工作上。传统的搜集材料的方法是：语言学家先阅读大量的语言文本，然后把所需要的材料记录在卡片上。这种方式不仅速度极慢，而且效率极低，因为：①这种搜集材料的工作往往是为了一个特定的研究课题或工程而做的，因而一旦一个研究工作结束这些材料就难以再得到利用；②一个研究者或研究集体所做的资料别人往往难以使用；③对一大堆卡片的分类、整理十分困难；④对材料的统计分析十分困难。这种落后的工作方式耗费了语言学家太多的宝贵时间，严重地阻碍了语言科学的发展。

现代电子计算机的出现给语言学家带来了福音。随着计算机技术的飞速发展，计算机的运算速度越来越快，存储量越来越大，价格越来越便宜。今天许多语言学家都有条件使用计算机，有的还拥有自己的个人计算机。在这种情况下，为语言学家提供一个大容量的计算机语料库的工作就成为迫在眉睫的任务了。计算机语料库与人工搜集的语料相比具有以下优点：①语料的规模可以很大，现在国外的语料库许多都在千万字以上；②语料可以共享，一旦一个语料库建成以后，有关的研究者都可以使用；③语料的检索方便快捷；④语料的统计快速准确。

从60年代世界上第一个计算机语料库诞生以来，语料库的建设日益受到各国语言学家的重视。特别是最近一二十年来，围绕着世界几个主要语言（特别是英语）建成了一大批计算机语料库，这些语料库为语言的描写研究、语言教学和自然语言处理都作出了突出的贡献。

现代汉语的研究只有几十年的历史，还处在十分幼稚的阶段。我们对汉语规律（特别是语法）的认识还比较浅，这里的原因是多方面的，但其中一个十分重要的原因就是我们对汉语的语言事实（或真实使用情况）还缺乏全面的深入的调查。

现代汉语研究语料库系统就是在这个背景下提出来的一个课题。它的目的是为现代汉语的研究工作者提供一个研究的平台——一个大容量的、经过各种标注处理的现代汉语语料库和一套功能强大的检索统计软件，使他们抛开做苦力式的搜集材料的工作，而把主要精力投入到对语言材料的分析上，并在此基础上总结汉语的规律，建立汉语的规则体系。

经过标注的语料库可以满足现代汉语研究特别是语法研究对于语料的多层面的需求，主要包括：字频、词频统计，个别词语的用法，某一类词的用法，某一种短语的结构与功能，句型的构造等等。

在大规模语言事实调查基础上建立起来的规则体系除了理论上的价值以外还具有重大的实用价值，这表现在两个方面：

①汉语教学，特别是对以汉语为第二语言的学生的教学。长期以来，教学所依据的语法规则基本上是以西方的语法理论为蓝本在汉语语法学家的语感和小规模零星的语言事实调查的基础上建立起来的。这种语法体系或失之粗疏，或失之允当，难以满足语言教学的需求。基于语料库的研究将会使建立汉语的详解语法成为可能，这种详解语法的特点是：a.全面性。它包括句型、短语结构、词语用法等。b.细密性，它不满足于建立一个语法体系，而是对语法构造的各个层面的规律作出详细的描写，并在此基础上给出合理的解释。c.定性分析与定量分析相结合，每一条语法规则都有相关的统计数据。

②自然语言处理。由于自然语言现象纷繁复杂，传统的基于规则的方法难以满足处理大规模真实文本的要求，特别是庞大的规则体系难以维护，基于语料库统计的方法将为形式规则体系的建立和维护提供方便的条件。

§2 系统的构成与开发流程

2.1 系统的构成

从用户的角度看，语料库系统主要包括两大部分：

①语料库。语料库是存储在介质上的经过选样的一系列现代汉语真实文本的集合。从语料的加工程度的不同可以分为：a、粗语料库：包含原始文本（只改正文字、用法上的明显错误）；b、精语料库（包括经过分词语料库、经过词性标注和经过句法分析的语料库）。

②包括各种检索、统计功能在内的语料库应用软件包。它的主要功能包括：a、任意范围的字频统计；b、任意范围的词频统计；c、关键字的任意长度的上下文检索(KWIC)；d、关键字的例句检索；e、语法标记的上下文检索；f、语法结构的检索与统计。

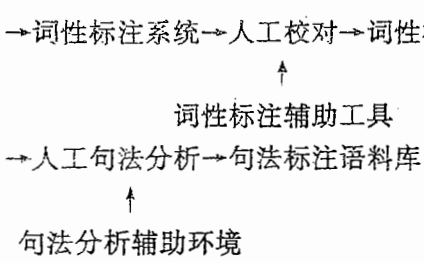
从开发者的角度来说，除了以上两大部分以外，还有一个重要的部分就是语料库开发工具。这些辅助软件工具主要包括：a、汉语自动分词系统；b、汉语自动词性标注系统；c、汉语句法分析辅助环境。

2.2 系统的开发流程

原始语料库→分句→分词系统→人工校对→分词语料库

↑

分词校对辅助工具



§3. 语料的选择与抽样

3.1 语料选取的原则

建造语料库，首先碰倒的就是语料的选择问题。我们正生活在一个"文本爆炸"的时代，现实中的语料浩如烟海，尽管计算机的外存可以很大，但是不可能也没有必要把所有的语料都纳入语料库中，这就要求我们根据一定的原则对语料加以选择。我们在选材时遵循以下几条原则：

①语料内容的广泛性。我们这个语料库的内容是通用的，不涉及专门领域的高深知识，具体来说，大致不超出一个高中毕业生的知识背景。考虑到现代汉语研究的需要，除了收集大量书面语材料之外，还收集一定量的口语材料。在书面语材料中，要考虑各种题材和体裁的内容。在题材上要包括政治、经济、文学、历史、科普、地理、文化、体育、卫生等各个方面；在体裁上除了诗歌以外大致都包括在内，如小说、散文、戏剧、政论、通讯报导等。

②语料的代表性。实际的言语材料在数量上是无限的，要使语料库具有代表性就应该使语料的数量足够地大。计算机大容量的外存为大规模的语料库提供了可能性，另外电子出版的广泛普及和 OCR 的应用也使语料的获得更加方便。但现在语料库的规模受到加工能力的限制。语料库只有经一层层的加工才能得到充分的利用，但现有的语料加工还不能完全实现自动化，需要人的不同程度的介入。考虑到这些因素，我们把本语料库的短期（两年）的目标定为：a、分词语料库：200 万字；b、词性标注语料库：200 万字；c、句法分析语料库：100 万字。

③语料比例的合理性。要使有限的语料反映语言的实际使用情况，就不仅要使语料的覆盖面广，而且要使各种语料的数量之间保持一个适当的比例，以使语料既有广泛性，又有代表性。各种语料之间究竟要保持什么样的比例，严格地说，这不是一个语言学问题，而是一个社会学问题，因为各种语料在普通人日常生活中的比例往往与社会因素有关。比如在许多国家宗教在日常生活中就很重要，但在我国则不是这样，所以汉语的语料库就可以把宗教题材排除在外。当然，我们在考虑语料的比例时，除了要考虑语料的实际使用情况以外，还要考虑到语料采集的难易程度。比如，在我们的日常生活中，口语占有相当重要的地位，它甚至比书面语更重要。但是口语语料的采集相当困难，目前口语的转写还不能实现自动化，所以在目前要得到上百万、千万的录音转写材料是十分困难的。一个弥补的办法是采集一定量的准口语材料，如话剧剧本、电影电视剧脚本、口语教材等。

④语料的规范性。不收方言材料和方言味太浓的文学作品，港台及海外华人的作品也不选用。对于文学作品首先考虑其语言的规范性，不因为作家的名气大或作品的影响大而收入语言不规范的文本。

⑤文本的完整性。不要求样本有统一的长度，不因为追求一致的样本大小而使文本变得不完整，这主要是为了将来篇章研究的方便。

⑥从时间上来说，语料要尽可能地新，确切地说是 1980 年以后出版的。

3.2 抽样的方法

要保证语料的代表性，除了要注意选材的广泛性之外，还必须应用随机抽样的方法。过去国内一些以词频统计为目标的工程也都采用了随机抽样的方法。他们的做法是：先按随机抽样的方法选好文本，然后再把文本录入机器。现在的情况有所不同，因为现在电子出版已十分普及，电子形式的语料很容易获得，所以现在建语料库一般都不采用手工录入的方法，因为这样很不经济。我们的做法是：先收集大量的电子形式的语料，然后根据上面提到的选取原则把不合格的文本剔除出去，再为合格的文本标注文本属性，文本属性主要包括文本的分类、文本的大小、作者、出版时间等，这样就建成了一个文本属性库，最后在这个文本属性库的基础上由机器按抽样的原则由一个程序进行自动的随机抽样。这样既保证了语料抽取的科学性，又大大地提高了效率。

§4 语料的加工

不经过加工的语料库其用处是有限的，大抵只能做字频统计和字符串的检索。要想使语料库发挥更大的作用，就必须对语料进行种种加工。这些加工大致包括分词、分句、词性标注、句法标注、语义标注、语用和篇章标注等。目前后两类标注还比较困难，因为究竟标注哪些内容，以及怎样标注，大家的认识还不太清楚，所以我们暂时把这两项内容列为待研究的内容。下面着重对分词、词性标注和句法分析三个方面加以讨论。

4.1 分词

4.1.1 自动分词　分词是汉语的特殊问题，因为汉语不实行分词连写，汉语的书面形式就是一个个汉字串。要对汉语进行处理，首先要把连续的汉字串离析为汉语的词串。汉语的自动分词技术在 80 年代末趋于成熟，目前国内有好几个实用的分词系统，这些系统的分词准确率一般都在 90%以上，但在分词上还存在着一些问题，其根本的问题是如何从语言学角度准确地定义词。对于词的定义问题语言学界争论了几十年，至今仍没有一个令人满意的答案。我们不能满足于目前语言学界对词的认识，而要深入地调查汉语文本的实际情况，从语言学角度制定一个分词原则，并根据这个原则形成一个汉语的词表，这对于提高分词系统的精度是十分重要的。

4.1.2 计算机辅助校对　再好的分词系统也不可能做到百分之百地正确，要使语料库得到正确的分词结果，就必须有人工校对。人工校对是分词过程有机的组成部分，如何保证人工校对的准确性对于分词质量是至关重要的。如果没有计算机辅助要保证这一点是很困难的，因为人工校对往往需要很多人参加，各人对词的认识不尽一致。即使是同一个人，今天做的跟昨天做的就可能不同。这就需要一个客观的约束机制，以保证分词结果的一致性和准确性。计算机辅助校对主要利用一个词表和一个在线(on-line)的分词规范。在做人工校对时，遇到一个词表中没有的词，系统都会提醒校对者，直到校对者确认，才会把一个生词存入词表，并加上特殊标记。

4.2 词性标注

4.2.1 自动标注　基于概率统计模型的自动词性标注系统在英语语料库的建造中得到应

用，并达到了令人满意的效果。清华大学黄昌宁教授等首次将这种概率模型应用到汉语中，建立了汉语自动词性标注系统。该系统在一定规模内测试对于开放语料标注的正确率达到 96%以上。我们打算将标记集作适当修改之后把这一系统应用到我们的语料库开发中。

汉语的词类问题在语法语法学界分歧也很大。作为一个面向所有语言学家的语料库系统，我们的标注体系应该能够为各家所用，这就要求我们的标记体系要比一般的词类体系细致得多，因为关于词类的分歧一般都反映在大类的区别上，分类的细化能够在各个分类体系中尽可能地找到共性。

4.2.2 计算机辅助校对　　同样，自动词性标注也不可能达到百分之百的准确率，也需要人工校对。基于和以上同样的原因，也需要一个计算机辅助工具，以保证标注结果的一致性和准确性。辅助工具的功能主要包括标记的合法性检查、词典约束和相类词（同义词或反义词）词性的即时查询等，另外也考虑利用一些确定性的规则以纠正一些错误。

4.3 句法分析

句法分析就是对句子作短语结构的分析，标注的内容有两项：短语的内部结构关系和短语的功能类型。短语结构关系的类型有：主谓、述宾、状中、定中、述补、并列、介宾、后附等。短语的功能类型有：名次性短语、动词性短语、形容词性短语、副词性短语、介词短语等。短语类型实际上对应于短语结构语法中的非终结符，为了使规则体系简明，也为了人工分析的方便，短语类型应尽量地简化。

对句法分析，有一点需要说明。在汉语语法的研究和教学中，句型具有重要的地位，但是我们对语料的句法分析并没有包括句型分析，原因有两个：第一，到底什么是句型，语法学家的认识分歧还比较大；第二，现有的句型概念比较容易从短语结构的分析中获得，比如主谓句就是第一层关系是主谓关系的句子，主谓谓语句就是第一层谓语内部是主谓关系的句子，一些带关键字的句型如“把”字句、“被”字句，“是”字句等更容易获得。

对自然语言进行自动的句法分析目前仍是一个世界性的难题（这也恰恰是语料库研究的一个目标）。根据我们的调查，目前对英语语料库进行大规模句法分析的有两家：一个是由 M. Marcus 领导的小组在美国宾州大学做的，他们的做法是自动分析加人工后处理，其自动分析是不完全的，对一些容易出错的结构如介词结构就不作分析，后处理量是比较大的；另一个是由 G. Leech 领导的小组在英国 Lancaster 大学做的，他们经过了多次反复，最终放弃了自动分析而采用手工分析，辅以计算机辅助环境。对世界上描写最充分的英语作自动分析尚且如此困难，对于汉语来说困难就更大了。我们的设想是：吸取 Leech 小组的经验教训，第一步由人工作句法分析，等标注达到一定量以后再考虑自动标注的可能性。

§5　语料的检索与统计

经过不同层次标注的语料库就成为不同层次上的语言知识库。要使这些知识库得到充分利用，就必须为用户提供一个使用方便、功能齐全的语料库应用软件。考虑到用户大多是对计算机不太熟悉的语言研究者，所以在设计该软件时必须充分考虑界面的直观性、友好性，最好有一个示教程序，使用户很快掌握系统的使用。

5.1 语料的检索

该模块的主要功能是提供在任意的文本范围内检索字、词、短语或语法标记的功能。按检索形式的不同可以分为:

① 词汇索引。即给出一个索引表，表中的每一项包括一个词汇形式及其出处，词汇形式是一个词或一个惯用语，出处即词汇形式出现的文本号及行号。②关键词的上下文检索(Key Word in Context)。这是应用最广泛的一种检索形式，它可以输出一个字、词或短语的上下文，这个上下文可以是：a.关键字前后任意长度的字符串。在分过词的语料库中，长度可以以词的数量计算，在没有分过词的语料库中，长度只能以字的数量或字符的数量计算。b.关键字所在的句子。句子是以一些标点符号结束的字符串。

③语法标记的上下文检索。在作过词性标注和句法标注的语料库中，检索出某一词类标记或句法标记的上下文。这个上下文可以是语法标记，也可以是具体的词或者二者兼有，长度也是可以任意指定的。

5.2 语料的统计

该模块的主要功能是提供在任意的文本范围内统计字、词、短语或语法标记出现情况的功能。按统计形式的不同可以分为:

① 字频统计。可以提供字表和按音序或频序排列的频率表。

② 词频统计。可以提供词表和按音序或频序排列的频率表。

③ 有关文本的统计。如文本的词次、词量、平均词长、平均句长等。

④ 语法统计。包括词类标记的共现频率，句法标记的共现频率，短语结构语法规则集及各个规则的频率统计等。

* 本课题的研究得到国家教委人文社会科学基金和北京语言学院的资助。

参考文献

① Leech,G. 1991. The state of the art in corpus linguistics,in English Corpus Lingustics,(ed.) by Karin Aijmer and Bengt Altenberg. London: Longman.

② Leech,G. and Garside,R. 1991. Running a grammar factory: The production of syntactically analysed corpora or "treebanks",in English Computer Corpora,(ed.) by Stig Johansson et al. Berlin:Mouton de Gruyter.

③ Sinclair,J. 1991. Corpus Concordance Collacation, Oxford University Press.

④ 黄昌宁，苑春法(1992)：国外语料库述评，载《机器翻译研究进展》，陈肇雄主编，电子工业出版社。

⑤ 胡明扬(1992)：现代汉语通用语料库的建库原则和设想，载《语言文字应用》1992 年第 3 期。

⑥ 常宝儒(1989)：现代汉语频率词典的研制，载《现代汉语定量分析》，陈原主编，上海教育出版社。