DOI :10.16499/j.cnki.1003-5397.1996.03.001
语言文字应用 1996 年第 3 期(总第 19 期)

现代汉语语料库研制

刘连元

提要 现代汉语语料库是通用型语料库,采用系统选材方式,共选取 1919 年至今的现代汉语语料 7000 万字。本文着重介绍了该库的总体设计原则、选材原则以及汉语语料库的有关规范问题。

国外语料库开发通常包括五个阶段:规划(planning) 、设计(design) 、选材(selection) 、建库(creation) 和标注(annotation) 。现代汉语语料库参照国外语料库开发规范,结合汉语特点进行研制。

一 规 划

现代汉语语料库是由国家主管部门主持建立的,面向全社会的应用需求。根据在全国十几个科研和教学单位的调查,现代汉语语料库将主要服务于以下五个方面的实际需要。

1 .语言文字信息处理。近年来,语料库研制在各国计算机语言文字信息处理中,日益受到重视。正如英国语料库语言学家 John Sinclair 1991 年所指出的,"现用语言的计算机语料库为信息科学各个分支的工作展现了新的前景"[①]。国内信息界专家普遍认为,应建立一个大型的现代汉语语料库。基于大规模真实语料,一方面可以对汉语进行定性与定量相结合的分析研究,为计算机汉语处理提供系统的数据和规则,另一方面又可直接支持基于语料库的各类计算机应用系统的开发。

2 .语言文字规范和标准的制定。面向信息处理和语言文字社会应用的规范和标准,需要经过足够数量的语言文字材料的分析和检验,否则其科学性难于保证。

3 .语言文字的学术研究。没有大型语料库的现代化手段,涉及大量语言文字材料的宏观课题难于开展。国外从 60 年代以后,各主要国家分别建立了不同语种的语料库,并且兴起了称之为语料库语言学(Corpus Linguistics) 的分支学科。"语料库语言学可以描述为基于文本语料的语言研究"[②]。这种研究基于大规模真实语料,又借助于计算机处理手段,其应用价值日益受到重视。

4 .语文教育。现代汉语语料库对现代汉语具有很强的综合分析能力,可以为各类语文教材的选材和语文教学大纲的制订提供依据,从而提高语文教育的科学性、系统性和现代化水平。

5 .语言文字的社会应用。大型通用语料库社会应用范围很广泛,例如,供国内外各界检索汉语这个大语种的各种信息和数据;供辞书编纂部门查询新词新语、词汇使用的语境和例句,等等。

语料库的用途是确定语料库类型的主要依据。关于语料库类型的划分，目前国际上仍无固定的说法。但是，由于用途不同，出现了类型不一的多种语料库，则是一个客观事实。Donald E. Walker 在《语言生态学》一文中[③]，将语料库划分成四种类型：1.异质型(Heterogeneous)。这种类型的语料库广泛收集和存储各种语料，语料选取并不要求依据某个事先确定的选材原则，所选语料一般只按其原貌存储。如美国计算语言学会倡议建立的 ACL/DCI 大型语料库，英国牛津大学计算中心建立的 OTA 文本档案库，都是异质的语料库；2.同质型(homogeneous)。与异质语料库相反，这类语料库所收语料必须具有同一属性。如美国 TIPSTER 语料库只存储与军事有关的文本，德国波恩大学建立的 Kant 语料库只收集作家 Kant 的著作；3.系统型(systematic)。依据事先确定的选材原则和比例选取语料，这类语料库强调语料选取的系统性、均匀性和合理性，力求具有广泛的代表性，以真实反映一个特定语种或特定范围的语言事实全貌。如英国的 BNC 语料库，美国的 Brown 语料库，以及由英国 Lancaster 大学倡议、由挪威 Oslo 大学完成的 LOB 语料库，都属于这一类型；4.专用型(specialized)。专门服务于某个特定用途的语料库，如美国为研究儿童心理语言学而建立的 CHILDES 语料库，为珍藏人文科学重要著作和资料而建立的北美人文科学语料库等，都属于专用型语料库。

现代汉语语料库要满足信息技术、标准研制、学术研究、语文教学和社会应用等多方面的需求。这些方面的应用都涉及现代汉语的语言事实全貌。因此，现代汉语语料库应是一个规模较大的通用语料库，应能真实地反映现代汉语在文字、词汇、语法、语义等方面的全貌。也就是说，现代汉语语料库是一个系统型的语料库。

根据该库的类型和用途，确定其库容量为 7000 万汉字，这个规模与英国国家语料库 BNC 大体相当。语料库建成后，拟每年增补 5%的新语料。

二 设 计

现代汉语语料库的类型确定后，需要制订一系列的设计原则，以保证建成的语料库符合规定的类型和用途。现代汉语语料库的设计原则包括通用性原则、描述性原则、实用性原则、抽样原则等许多方面。

1.通用性原则

为了保证语料的通用性，需要处理好一般语料与专业语料、普通话语料与方言语料、书面语料与口语语料之间的取舍关系。

现代汉语语料库所收语料应有别于各类专业性语料。随着科学技术的普及，专业用语不断进入通用语言，现代汉语语料库应尽量吸收这部分语料，但从整体上说，该语料库不能大量包容专业语料。

现代汉语语料库所收语料应有别于纯方言性语料。随着社会交际的发展，有些方言词语逐渐进入普通话，各类语料中都有方言词语出现。因此，现代汉语语料库的语料会收入一些方言词语，但是，从整体上说，该库不应包容过多纯方言语料。

现代汉语语料库应以书面语料为主，以口语语料为辅。口语词语因地域不同往往有较大差异，使用也不广泛。大量包容纯口语语料必将影响语料的通用性，只能适当收入一部分能用书面语转述的口语语料，如剧本、相声、谈话录、讲演录等。

2.描述性原则

现代汉语语料库的科学性和应用价值,在很大程度上取决于所收集的语料是否能客观地反映现代汉语在字、词、句法、语义、语用等方面的全貌。所谓描述性选取就是指从现代汉语使用的实际状况出发,客观地选取语料,尽量避免主观干预。描述性选取能使采集的各类语料符合实际的使用频度,使所选的语料既具有广泛性,又具有代表性。然而,代表性作品不一定是优秀作品,优秀作品只是现代汉语语料中的一部分。可见,语料选取的描述性强调的是忠实于语言事实的原貌。

3.实用性原则

现代汉语的时间跨度,从1919年算起,已经有几十年。这几十年中经历了不同的政治历史时期;使用现代汉语的社会成员又处于不同的文化层次;现代汉语各类语料在社会生活中使用情况往往也有较大差别,有些语料的字、词、句、义在社会生活中使用广泛、频繁,有些使用面则相对狭窄或使用不多。这些情况说明,现代汉语语料从时间层次、文化层次和社会使用面层次上说,客观上存在差异。现代汉语语料库所收语料数量虽然不少,但与现代汉语实际使用语料相比,仍然是一个很小的数目。为了使有限的语料反映出现代汉语面貌的主要特征,必须从实用的原则出发,在时间层次、文化层次和社会使用层次上对语料进行不等密度处理:紧紧围绕各个层次上的中心,加大中心部分的选材密度,形成"抓住中心,其他补充"的选材方式。

在时间层次上,选取1919年至今的各个时期语料,但以1977年至今的语料为主。在文化层次上,以具有高中文化程度的人能够阅读的语料为主,其他文化程度为辅。在社会使用面层次上,以使用广泛的语料为主,其他语料为辅;以人文与社会科学为主,自然科学为辅;以学科门类为主,以语料语体为辅,对门类进行补充。

4.抽样原则

选取语料首先需要在实际使用的语料中确定一个范围,称之为母本;被选入库的语料称之为样本。由母本取得样本的过程叫做抽样。为了保证语料库的科学性和应用价值,抽样必须遵循一定的原则:(1)母本的语言材料应具有多样性,以便尽可能完整地涵盖现代汉语的各类语料。母本是抽样的基础,母本的多样性是保证样本具有代表性的前提条件。(2)样本的语言材料应具有相对的完整性。这需要适当确定样本容量。容量太小,语言材料显得支离破碎;容量太大,语言材料可能会过多受作者本人语言风格和主题内容影响。(3)样本总量与母本总量的比例要合理,以保证语料抽样的广泛性。现代汉语语料库的样本总量(即选材数量)与母本总量(即选材所涉及的语料数量)的比例大约为5%。7000万字语料,这意味着选材所涉及的语料数量为14亿汉字。(4)要注重抽样方法的科学性。抽样方法的选用与母本的属性特征有密切关系。现代汉语语料库按门类和语体分类选材,实际上是一种分层随机抽样方法。抽样的随机性是一个基本原则,但并不排除在某些特定情况下进行必要的人工干预,以使语料抽样更加合理、有效。

三 选 材

选材是保证语料库科学性和实用价值的关键环节。各国语料库建设中对选材工作都十分重视,往往需要集中各界专家,经过反复论证,确定选材的原则和方法。现代汉语语料库选材之前,首先在专家充分论证的基础上,于1993年1月制订出《现代汉语语料库选材原则》(以下称

《选材原则》。这份选材原则是严格按照前面叙述的几项设计总原则拟定的，其中各项具体规定和要求充分体现了设计总原则的精神和要求。

1.语料分类

现代汉语语料库的语料可以从门类、语体、来源等不同角度进行分类。按照"以门类为主，以语体为辅"的设计原则，现代汉语语料库由人文与社会科学类、自然科学类和综合类三大部分语料组成。

人文与社会科学类划分为8个大类和30个小类：(1)政法：哲学、政治、宗教、法律；(2)历史：历史、考古、民族；(3)社会：社会学、心理、语言文字、教育、文艺理论、新闻、民俗；(4)经济：工业经济、农业经济、政治经济、财贸经济；(5)艺术：音乐、美术、舞蹈、戏剧；(6)文学：小说、散文、传记、报告文学、科幻、口语；(7)军体：军事、体育；(8)生活。

自然科学类划分为6类：数理、生化、天文地理、海洋气象、农林、医药卫生。

综合类语料由应用文和难于归类的其他语料两部分组成。应用文使用很广泛，主要涉及以下6类：(1)行政公文：请示、报告、批复、命令、指示、布告、纪要、通知等；(2)章程法规：章程、条例、细则、制度、公约、办法、法律条文等；(3)司法文书：诉讼、辩护词、控告信、委托书等；(4)商业文告：说明、广告、调查报告、经济合同等；(5)礼仪辞令：欢迎词、贺电、讣告、唁电、慰问信、祝酒词等；(6)实用文书：请假条、检讨、申请书、请愿书等。

现代汉语语料库的语料来源包括教材、报纸、综合性刊物、专业刊物、图书等。

不同门类、语体和来源的语料，需要确定合适的比例。《选材原则》中规定的选取比例是集中了专家们的意见后确定的。尽管如此，仍难免带有经验成分。为了便于选材操作，事先规定比例是必要的，但在选材过程中，需要根据语料的实际状况，对各类比例进行合理的调整和修正。现代汉语语料库7000万字选材完成后，各类语料的实际比例基本上符合《选材原则》中规定的比例，但都有所调整。

人文与社会科学类语料占语料总量的59.6%，其中包含的8个大类语料在人文与社会科学语料中所占的比例分别为：政法12.7%；历史8.4%；社会14.0%；经济9.8%；艺术6.7%；文学44.9%；军体2.3%；生活1.4%。

自然科学类语料占语料总量的17.24%，其中包含的6类语料在自然科学语料中所占的比例分别为：数理17.2%；生化19.1%；天文地理14.1%；海洋气象9.1%；农林22.8%；医药卫生17.7%。

综合类语料占语料总量的9.36%，其中各类应用文占综合语料的91.1%，其他语料占8.9%。

取材于报纸的语料，难于划分门类和语体，报纸语料的数量和比例没有计入上述各类语料中。报纸语料占语料总量的13.79%。

取材于教材的语料总量有2000万字，已经按学科计入各类语料。

2.语料年限

现代汉语语料库选取1919年至今出版的现代汉语语言材料。由于时间跨度较大，选材年限又划分成五个时期。根据每个时期内社会历史背景和现代汉语的应用特点，确定了不同的选材比例。因此，按选材时间年限说，现代汉语语料库对五四以来各个历史时期的语料采用的是不等密度的选材方式。

人文与社会科学类语料选取分为以下五个时期：

1919－1925 年的语料选取比例为 5%。这个时期的白话文仍有文言痕迹，少量选取对后世影响较大的作品，并注意所选作品在行文上尽量接近现代汉语规范。

1926－1949 年的语料选取比例为 15%。这个时期的白话文逐步脱离文言痕迹，现代汉语日趋成熟。

1950－1965 年的语料选取比例为 25%。这个时期的社会生活发生巨大变化，新词新语大量涌现。

1966－1976 年的语料选取比例为 5%。文化大革命时期出现的许多词语，具有特定历史时期的明显色彩，文革结束后大多不再使用。

1977 年至今的语料选取比例为 50%。改革开放新时期的语料，其内容、题材、体裁都很丰富，代表了现代汉语的最新发展。

人文与社会科学类语料，从历时角度说，采用不等密度的选材方式保证了语料库的实用性。但从共时角度说，语料的选取比例要尽量均匀。如现代汉语语料库中，选自书籍和刊物的人文与社会科学语料，所占比例分别为 60～70%和 30～40%；小说中规定了长篇小说(15 万字以上) 、中篇小说(5－15 万字) 和短篇小说(5 万字以下) 的比例为 1∶2∶2；在内容难易取舍上，规定了专业性强、高中水平的人阅读困难的语料不取，学术性虽强、但能读懂的占 40%，普及性读物占 60%。

自然科学类语料，由于学科发展和技术更新的速度较快，在内容和术语使用上不同时期存在较大差异。有些内容和术语随时间推移，变得陈旧，甚至被淘汰。因此，自然科学类语料的年限确定，原则上以共时性选取为主，年限分布主要在 1977 年至今的范围内。选材范围包括目前比较通用的中小学各科教材，目前比较通用的具有通论性质的大学各科必修基础课程教材，以及涉及自然科学各个门类的科普读物。

综合类语料以各类应用文为主。不同时期的应用文在格式和词语使用上差异明显。现代汉语语料库主要选取 1977 年至今的各类应用文。

报刊语料从 1919 年选起，也分为五个时期，其年限和比例与人文与社会科学类语料相同。其中 1949 年以后的报刊分为两大类：全国性报刊(占 25%) 和省市报刊(占 75%) ，各约 10 种。

3 .语料的描述信息

在选材过程中，需要同时收集和记录语料的有关描述信息，供语料库建立各种描述项，以便根据用户不同需要进行各种方式的检索。现代汉语语料库为每个语料样本建立了 20 个描述项目：总号，分类号，样本名称，类别，作者，写作时间，书刊名称，编著者，出版者，出版日期，期号(版面号) ，版次(初版日期) ，印册数，总页数，开本，选样方式，样本起止页数，样本字数，样本总数，简繁字。

4 .语料样本

现代汉语语料库样本的容量是 2000 字左右，允许有±500 字的伸缩。一个语料样本应取自同一作者的同一篇文章，每个样本必须是连续的语料内容。选用的样本应尽量保持语料原貌，这是目前国际上语料选材实施中普遍主张的做法。现代汉语语料库规定，样本材料原则上只删除样本首尾不足一句的语言片段，即每个样本均从句号(也包括表明句子结束的问号、叹号、省略号等) 为其开始点和结束点。文中内容除删除脚注标志、脚注内容、作者名外，原则上均

应保留。如保留标题、小标题、括注外文等。

四 建 库

语料库是存储大规模真实文本的数据库系统。由于汉字和汉语的自身特点，汉语语料库系统的设计与开发，需要解决一系列特殊的问题。如汉字字符集的选用和库外字的补充；简化字和繁体字混用的语料录入处理；汉语汉字的人机界面设计；汉语分词系统和标注系统的建立，等等。现就系统总体设计、语料录入和核心语料库建立谈谈现代汉语语料库的建库情况。

现代汉语语料库系统功能包括：1．数据维护，包括语料录入、核对、存储、修改、删除及语料描述信息项目管理；2．语料加工，包括语料自动分词，自动标注，语料文本分割、合并、标记处理等；3．用户功能，包括查找、检索、打印、关键词上下文索引等；4．实用系统，可以根据用户要求进行字、词、句法等方面的言语统计及各种应用语言学课题的研究和开发。

现代汉语语料库按照标准化、模块化和系统工程的原则进行设计。汉字字符和汉字库采用国家有关标准，设备和网络选型通用化，系统及子系统设计按照国家有关标准进行；软件采用模块化设计，系统具有较强的适应能力和扩展能力，数据库与应用程序之间相互独立，减少数据格式对应用程序的依赖性。数据操作采取统一的控制方法，保持数据操作的一致性。人机界面设计稳定可靠，无二义性，易于进行一致性检查，简明易懂，记忆量少，符合人机工程学要求和语言文字的应用规范。

现代汉语语料库的语料录入采用流水作业方式，由两条作业线完成语料扫描、校对和入库操作。每条作业线的流程包括：1．语料录入，采用 OCR 技术，进行语料文本的图像扫描，经识别程序，生成数据文件；2．一校，参照语料图像扫描文件，对文字识别后的数据文件进行第一次校对。一校错误率要求低于千分之二以下；3．二校，参照语料原稿，对一校后的数据文件进行校对。二校错误率要求在万分之六以下；4．三校，参照语料原稿，对二校后打印出的文件进行文本校对。三校错误率要求在万分之二以下，以符合国际上有关文本错误率的规定要求。5．语料入库，经三校后的语料数据文件装入现代汉语语料库。每个语料数据文件都有描述信息数据，供检索使用。每个语料数据文件平均具有 0.5K 的检索信息数据。

从上述语料录入流程可以看出，现代汉语语料库的语料录入是按照工程化的要求进行的，这就保证了现代汉语语料库实现语料数据可靠的预期设计目标。

由于 7000 万字的选材量比较大，建库工作分两步走。首先建立核心语料库，然后在核心语料库基础上完成 7000 万字语料的建库工作。核心语料库由 7000 万字的语料中筛选出的 2000 万字语料组成。由于《选材原则》是经几次专家论证确定的，核心语料库的语料筛选工作，在语料分科、年限划分、比例、字数等方面基本上仍依照《选材原则》进行，只是结合核心语料库的用途特点，在语料筛选上突出 1977 年以后的新语料，注意选用内容通俗、通用性强的普及性语料，因而不同年限和门类的语料比例和字数均有小的调整。

五 标 注

标注指由计算机系统自动完成未经加工语料的分析和语言学特征注记工作。标注是语料深加工的重要环节，也是一个语种语料库建设水平的重要标志。1990 年美国计算语言学会召开了关于书面语料资源的专题研讨会，来自文献信息检索、言语识别与合成、机器翻译、多文种

语料库、心理学、语言学和计算语言学等方面的专家在书面语料资源问题上取得的重要共识是，必须从事大规模的书面语料收集工作，并且要有足够数量的经过加工的书面语料，即具有语言学结构特征和其他信息特征标记和标注的语料。

汉语语料库的语料描述标记和语言学特征标注工作近几年已经起步，目前这方面的难点主要集中在汉语词汇研究和规范化两个方面。现代汉语语料库在完成核心语料库建库工作后，专门召集国内这一领域的专家，研讨计算机汉语分词和标注的技术难点和相应规范问题。汉语自动标注必须建立在自动分词的基础上，两者的连带关系十分紧密，这是汉语计算机处理特有的技术环节。

汉语自动分词目前依据的《信息处理用现代汉语分词规范》，普遍反映可操作性差，致使不同的分词系统的分词结果不一致，影响分词语料共享和进一步的计算机加工。因此需要制订一个实例化的词表，弥补《分词规范》的不足，适应计算机自动分词的工程需要。这种"规范＋词表"的思路已成为这一领域专家的共识，并已经立项研制《信息处理用现代汉语词表》。

为了确定该《词表》的收词原则，需要从两方面进行深入的研究：其一是分词颗粒度问题。面向词法分析和面向句法分析的分词颗粒度是不同的。词法分析需要的分词颗粒度小，以满足词频统计、词素分析、词义和词的句法功能分析等方面的需要；句法分析则倾向分词颗粒度大一些，以利于句子成分切分和句法结构分析。为此，专家们建议，这样的词表可以分两步走：先建立面向句法分析的词表，然后再建立面向词法分析的词表。计算机系统需要进行两级切分，形成两个词表，两种词频。其二是如何确定《分词规范》中所述的"结合紧密，使用频繁"这类词的收词范围。这需要对动宾、动补结构的语词进行量化分析，研究这类语词中的构词元素在实际语料中的同现情况，还要研究如果这类语词收得宽一些，对分词歧义可能造成的影响。

汉语自动标注目前需要解决两个问题：一是汉语词类兼类如何处理；二是在词类规范的基础上制订词类标记集。

汉语词类兼类情况十分复杂，特别是名动兼类，在有些情况下即使由人来分辨也很困难。这给自动标注造成很大困难。因此需要对汉语词类兼类情况进行系统的分析研究，得出定量的数据，制订出兼类词和词法分析层次上词类待定词的标注策略。

词类标记集需要在词类规范的基础上适时制订。在信息处理领域，词类划分（特别是词类大类划分）目前已具备条件。现代汉语语料库是面向社会应用的大型通用语料库，其标注的规范性极为重要，在语料加工阶段需要立项研究汉语词类标记集规范。

附　注

①John Sinclair，《Corpus Concordance Collocation》，1991。

②Karin Aijmer & Bengt Altenberg，《Corpus Linguistics》，1991。

③Walker，D·E·，《The Ecology of Language》，1990。

（100010　国家语委中文信息司）

APPLIED LINGUISTICS

No. 3 1996

Main Articles

Abstracts

Study of Corpus for Contemporary Chinese Language —By Liu Lianyuan

The corpus of contemporary Chinese language is a commonly—used corpus. The systematic method was adapted in material selection. This corpus includes 70 millions Chinese characters from 1919 to present days. In this paper, the general design principle, the material selection principle and some standards concerned Chinese corpus are introduced.

Three Relationships for Morality in Writing—By Xing Fuyi

Three relationship for morality in writing are: 1. To make clear the references of other persons to readers; 2. To explain the principles, to be amicable and equal in the academic discussion for diferent viewpoints; 3. To be good at getting undoubted conclusion through research and also to be good at leaving over the questions in order to further explore.

Investigation of Putonghua Usage and Language Viewpoint in Pudong New Region of Shanghai

By Project Group for Study of Language Policy and Language Viewpoint in Pudong New Regiong

In this paper, the method and result of this investigation are introduced, some suggestions for strengthening the popularization of Putonghua and law—giving for language are proposed.