January 1997

English-Chinese bilingual database and the compilation of dictionary

Gan Ye
Pepperdine University, ganye@msn.com

Nanjing University

南京大学学报(哲学·人文·社会科学) Vol.34 1997年 第1期

英汉双语语料库与英汉词典的编纂

叶 敢* 张柏然

摘 要:本文介绍了国内外语料库及南京大学筹建的英汉双语语料库的情况。鉴于双语检索为该项工作中的主要难点,本文针对词典文本中词条格式的特殊性,提出先对词条进行切分和标引,以便于建立特定的数据结构,为今后的检索、排序和统计作好准备,本文对语料的收集、处理和语料库的结构作了初步设想。限于篇幅,设想的信息流程以方块图显示。

关键词:词典编纂 计算机 双语语料库

一、引言

词典编纂工作有自身长期的经验传统和生产"工艺",这种工艺流程中最繁琐、最费时也是最薄弱的环节是词汇卡片的组织和建立。所建的卡片库往往是为编纂一本词典而日积月累起来的。对卡片的分类、整理、存贮不但需要耗费编辑人员大量的时间和精力,而且具有一定的难度。最典型的是十七卷本的《现代俄罗斯标准语词典》,其卡片系统是从1884年就开始累积的,耗时17年,卡片总数为六百万张。根据这套卡片进行的编纂工作直到1965年才完成。而这种以卡片为载体的语言资料库一旦词典完成就难以重复利用,即使是对原词典进行增订、修订也需一张张重新开始。编辑人员很难对其进行各种统计分析,以了解所掌握的资料情况。另外卡片库是个全封闭系统,很难实现知识共享。全国范围内为编纂词典而形成的编纂群很多,各个编纂群都自有一套这种以卡片为载体的语言资料库,互不沟通,信息重复率高,涵盖面不全。

伴随着信息时代的到来,语言信息的变化速度与词典的编纂速度之间的距离越拉越大。反映出的问题是词典尤其是大型综合性词典无法及时跟上现实生活中语言的变化。要想在短时期内依靠目前的编纂工艺流程,完成对大型综合性词典的增订和修订是无法想象的;另一方面编纂活动中这种大量重复性费时而简单的手工劳动,又恰恰是计算机最擅长的工作。如何尽可能多地利用计算机替代这些繁复的工作,使编辑的时间和精力完全投入到高级的创造性劳动中去是当前面临的新问题与新挑战。

二、国内外情况介绍

本世纪50年代计算机开始被用于自然语言的处理,人们想到的第一件事就是用计算机把一种语言翻译成另一种语言,而对自然语言的理解不仅依赖于语法知识,而且还要运用相

* 南京大学双语词典研究中心工作人员(南京 210093)

关的背景知识，其背景知识又往往来源于上下文中的特定的意义。缺乏大量的真实文本作为研究基础，其研究目标难以实现。"语料库(Computer Corpus)"就是在这个背景下诞生的，它的目的是为语言研究工作者提供一个研究平台——一个建立于计算机硬件之上的大容量的、经一定标注处理的语言资料库并附带一套功能强大的检索统计软件系统，它利用计算机处理速度快、存贮量大又便于查询等特点为语言学家提供大量最新的语言信息资料。

1964 年 W. N. Francis 和 H. Kucera 在美国 Brown 大学建成世界上第一个存储在数字计算机上的语言资料库。该库收集当代美国英语语料，按系统性原则采集 15 类文体样本共 500 个，每个样本不少于 2000 词次。1967 年又在该库的基础上提出一张书面语词表——《Computational Analysis of Present Day American English》，这是语言资料库用于基础研究的一次尝试。于是"语料库"这种依赖于计算机技术而建立的语言资料库系统便在欧美一些国家开始发展起来。目前国外最具代表性的大型英语语料库有以下几个：BROWN 美国标准书面英语语料库，共一百万字；AHI 美国中小学教材语料库，五百万字；LONDON 英国口语语料库，四十五万字；LOB 英国标准书面英语语料库，一百万字；COBUILD 柯林斯一伯明翰大学国际语料库，二千六百万字。其中美国文化遗产出版公司建立 AHI 语料库、英国柯林斯出版公司和伯明翰大学协作建立 COBUILD 语料库都是为词典学和词典编纂工作服务的。对词典编纂，语料库不仅可以作为选词的依据，而且还可提供有关词的意义和用法的大量实例。从库中提取词时可以很容易地连带其出现的语境(某篇、某行)一同显示出来，为自然语言处理提供了有利条件；这种作法在编纂词典工作中十分有用，使得编出的词典无论在释义或举例上都能及时反映出活的语言变化。各种语料入库时进行的查重，减少了语料之间的重复，从而在一定程度上避免了词典编纂工作中存在的那种词典之间辗转相抄的做法。如为了编纂一部美国中小学生适用的词典而建的 AHI 语料库，编者们从语料库中挑选 70,000 项意义和用法，为每项意义和用法摘出实例平均约 10 条，总共摘出实例计 700,000 条，编出的词典共收入词条 35,000 条。该词典自然能反映每个词在英语中的实际应用。1987 年英国柯林斯出版公司出版了一部新的英语词典，取名《柯林斯 COBUILD 英语词典》，其中很多词的释义和举例都根据 COBUILD 语料库的资料进行了及时更新，使其能跟上当今英语的实际用法。编纂该词典成功的经验为语料库的应用提供了范例。

我国从事汉语自然语言处理研究起步较晚，汉字信息处理的三大难题(汉字的输入/输出；汉语语词的自动切分；以及汉语的句法、语义的自动分析)中汉字的输入/输出问题已获得圆满解决；汉字编排软件具有很强的优势，能对各种西文软件包实现汉化，并且围绕词处理问题开展了诸如词频统计、分词规范、通用词表、自动分词及词库设计等研究工作。汉语的信息处理正处于从字处理向词处理过渡阶段。我国在语料库的建设及语料加工技术方面的研究虽只有三、四年的时间，但却在基于语料库的汉语句法分析、颗粒度较小的语言知识获取等方面取得了很大成绩。其中汉语真实文本的词性的自动标注和义项自动标注方面已达到了国际先进水平，建立了一批有代表意义的现代汉语语料库，如北京语言学院 200 万词规模的汉语句法语料库、深圳大学的红楼梦语料库、清华大学按系统性原则采样而建的 5000 万字语料库、上海交通大学科技英语语料库(100 万字)等。然而我国的自然语言处理主要偏重于汉语，大规模真实性文本语料库目前仅限于汉语单语种，双语语料库的建设虽已开始，但只限于科技方面，如西安交通大学的英汉语料库、上海同济大学在德国建立的德汉语料库

等。应用面更大的是对综合性英汉双语真实文本的语言知识获取与处理工作，至今涉及较少。在我国建立超大规模综合性英汉双语语料库以及利用语料库来支撑大规模真实文本处理的实用研究与开发，不仅为当今文化交流之急需，且为今后语言学界各项研究工作及人工智能领域的研究工作打下基础。

南京大学双语词典研究中心正在筹建大型综合性的"语料库"。打算以《综合英汉大词典》(将由商务印书馆于近期出版，总字数 3 千万字)为主，优选其它中外著名的词典及报刊，文艺作品等，完成后总容量将达 2 亿字。还打算以该综合性双语语料库为基础开发一个综合性的大型中英文字处理软件系统。其主要宗旨是为词典编纂服务。该系统建成以后从例句收集、分选、到调用、编辑、校对、修改、添加直到发排全部编纂处理工作都可在计算机终端上完成。本文将提出该系统的设想框图，并就系统内各种功能间的联系作一些初步论述。

三、语料库组成及其功能

在词典编纂行业中所谈的"语料库"是指建立于语料库上的编辑系统之总称，是指一个大型的应用系统，其内在的结构如图 1 所示。

建立该系统的工作可分"输入"、"信息处理"、"输出"三大阶段。"输入"是整个系统的基础，它作为输入资料的前处理部分而存在，输入资料在这里被分类处理、整理、标引、存放，并为以后的"信息处理"提供各种形式的数据；"信息处理"是整个系统的应用部分，也是整个系统的核心。它包括编辑、排序、查询等模块；"输出"则是整个系统与外界的一个"接口"，通过它将处理结果直接或经加工后送往外部设备。

(一)输入阶段：外界输入的资料按内容可分为词典体和文章体。词典体以国内外优秀词典为主，例如《综合英汉大词典》，词典还分为英英、英汉两种。词典体为编纂人员提供查询依据和样本资料，如一些贴切的例句和释义等。文章体则选自国内外报刊上的优秀文章或优秀的中外名著，以供选择例句和摄取最新语言信息。

1). 词典体输入：对内容庞杂的词典文本进行结构分析，确保其逻辑结构并正确地转为具有特定格式的电子数据。全过程在人工控制下进行。处理流程详见图 2。

经处理后该系统中的词条以特定的统一形式存放，其语法按以下条目描述(整个系统中词典库的数据结构框架据此而建)：

《词条》:《词头》[《词性说明》][《屈折变化》][《标签及说明》][《释义项》][《内词条》][《词源》][《同义词辨析》][《注释》]

语料库之所以能成为词典编纂的得力助手，一是因为它集大量语言信息资料于一体(尤其指大型的语料库)；二也是最主要的一点：各种语言信息以一定的格式存放，这样可为以后的调用带来极大的便利。建立这种系统的主要难点也就是信息的获取及处理部分。自然语言的形式多种多样，有声音、文字、动作等。存储手段和媒体也不同，仅以印刷版词典而言，词条的长短不一。一个词条短则不足 5 个字，长则可能 10 万字以上。词典间标识符也不相同，如在有些词典中用 [ ]标识音标，而另一些词典中 [ ]却表示语法说明，词源括注等。考虑到所建语料库主要为词典编纂服务，外界输入的资料以词典为主，故信息的获取及处理部分，也针对词典进行。为确保词典从印刷符号完全转化为电子数据必须经历：

1. 结构分析，以确定词条格式。

2. 逐条对词条按一定格式进行分解。

3. 按特定的数据结构存放数据。

其中对词条的切分，可按词条特有格式及标识符进行，并在机器自动切分基础上，对各数据项进行字号、字体和色彩的定义，以方便工作人员对切分结果进行校对和修正。这一点现今计算机技术已能达到。

2). 文章体输入：为了能向语料库添加更新的语料，整个系统还专设一个文章库存放各种语言论文、例如最新国内外报刊文章、文艺作品等。这种文体输入的资料，经标引、词频统计、新词处理等预处理后存入语料库。处理流程见图 3。

(二)"信息处理"阶段：主要目的是充分利用前阶段所提供的数据为词典编纂服务。次要的附属"产品"是利用库存的大量各类文献资料建立一个百科资料检索系统。按不同目的分成两大模块：文章资料库和词典编纂两者流程图见图 4 和图 5。

由图 5 可见，通过该系统可编纂各种形式和类型的词典。同时还可对语料库的词或通过这一系统产生的词典进行查询和统计，其途径也是多种多样的。其中逻辑"与"处理可提供任意二个以上信息途径的操作。如欲编一本"生物学名词词典"，则可通过"专业"和"词性"的"与"操作实现。由计算机挑选出全部既属生物学专业又属名词词性的词，排序并调出有关例句，构成一本机内词典。

(三)"输出"阶段：是该系统与北大方正系统的一个软接口(该系统输出的文本排版在北大方正下进行)，也是系统主机与各种输出设备的一个软通道。主要任务是将在该系统下生成的文件转换成北大方正系统认可的文件或对北大方正系统不认的符号进行处理。并负责将处理完毕的文件送往排版机、磁带机、磁盘、光盘等输出设备上。流程图因限于篇幅此处从略。

该系统还为管理和维护人员专设两个特殊模块——主编模块和系统维护模块。两模块的用户是该系统的超级用户，其目的不同，前者针对该系统输入、输出、处理各种阶段形成的信息资料的质量而设。介于编辑、用户或外界输入资料与库存资料之间，两者间的数据交流必须通过该模块确认后才能进行。主编可利用它对编辑人员进行管理、制定权限和工作分配、工作检查。后者即系统维护模块，作为整个系统运转管理的手段而存在，管理人员可利用它察看整体数据库的大小、数据结构及数据运行情况，以便决定增减输入资料的类型和数量，新旧数据的更换和数据的保护及用户介面的管理和维护。

四、应用前景

该系统建成后，词典编纂可实现半机械化操作，编辑人员可在终端前完成编辑工作。由于该语料库中的每个词，都经详细标引，词条以链表形式存储，所以当编辑将编纂要求告知计算机，计算机则自动调用所有符合条件的词条并排序，形成一个机内词典，由编辑在系统提供的窗口上对该机内词典进行编辑修改，系统随时提供各种所需信息，编辑完成后经校对、排版，从网上传递给主编供审定，最后输出。整个系统从资料收集、整理、存储到调用编辑排版全套工作皆可在计算机上完成。它摆脱了卡片的制作、积累和查找。以往制卡工作中完全靠经验、语感、直觉勾出所需的词。而"需要的词"这个概念是相对的，十分含糊，是不科学的。现在由计算机对输入资料进行扫描，勾出新词，由编辑确认后收入。这样可避免罕用词

因出现机会少而被漏掉。而某些常用词则可能被无意识略去。(编辑不可能检查卡片中的每个词,以为它们以前已经记录下来了)。如利用该系统编纂一本动词词典,操作步骤可以如下:

1. 从窗口菜单上选择词典编纂中该词性下的动词;

2. 机器调入含动词词性的所有词目,及词目下指定词性下的释义及例句,以字母为序对词目进行排序形成一机内动词词典;

3. 提供一窗口显示刚形成的词典给编辑修改加工,同时随时可为编辑人员提供和调用各类相关信息;

4. 完成初稿,经校对、排版,送主编审定;

5. 审定完毕发版,送出版部门。

可见该系统建成后,利用该系统进行编纂,将大大提高我国英汉双语词典的编纂速度,使最新语言信息迅速进入词典与读者见面。今后我国可在短时间内出版《综合英汉大词典》一系列中小型派生词典和《综合汉英大词典》。双语语料库在缩短编纂周期的同时也为编辑们提出了挑战。首先编辑们需学会使用计算机;其次该系统的使用为编辑们赢得了大量的时间和精力,提供了在出版最新词典的同时又能出最优秀词典的可能。这样一个大型的综合性的语料库软件系统,还可作为一个全新的情报源而存在。它可为各学科提供信息查询服务,包括学术论文及一般科普性文章的查阅或调用,还可进行各类文章的内容分析,为有关部门提供各类统计数据,行情预测和情报咨询服务等。如与世界上一些语言信息系统联网,就可实现各种最新语言信息资料的交流,成为语言学研究领域最得力的工具,为研究人员提供各种可靠的最新统计数据。另外因为语料库中已对其中的每个词进行多途径的详细标引,如运用著名学者 Peter Mark Roge 创造的主题词类别表进行语言学上的主题词标引,为实现机器翻译(尤其是对文学性题材的文章的自动翻译)提供了可能。

注释:

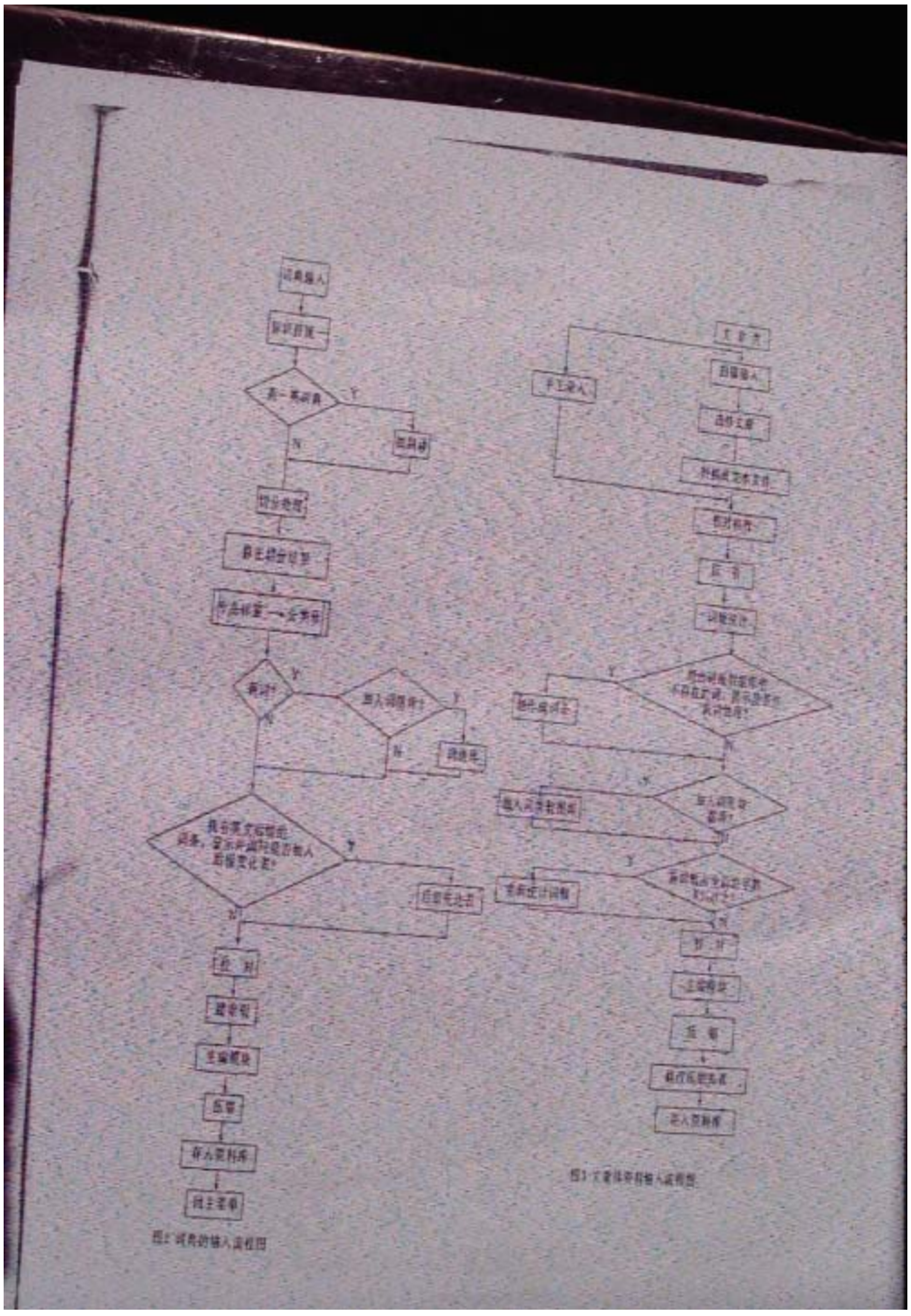
[1]R. R. K. Hartmann, Lexicography with particular reference to English learners' dictionaries, Language Teaching, Vol. 25, No. 3, 107 (1992), Cambridge University Press.

[2]张一舟,计算机与法语词典编纂工作,辞书研究,88年第2期21页。

[3]A. C. 格尔德,再谈词典编纂工作的自动化,辞书研究,90年第1期63页。

(责任编辑 赵 枫)

图1 语言资料库系统框图



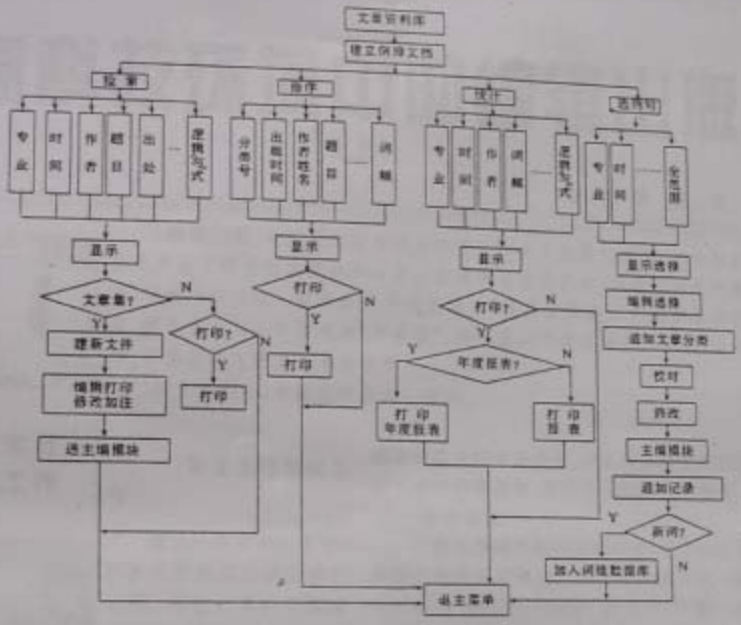


图4 文章资料库流程图

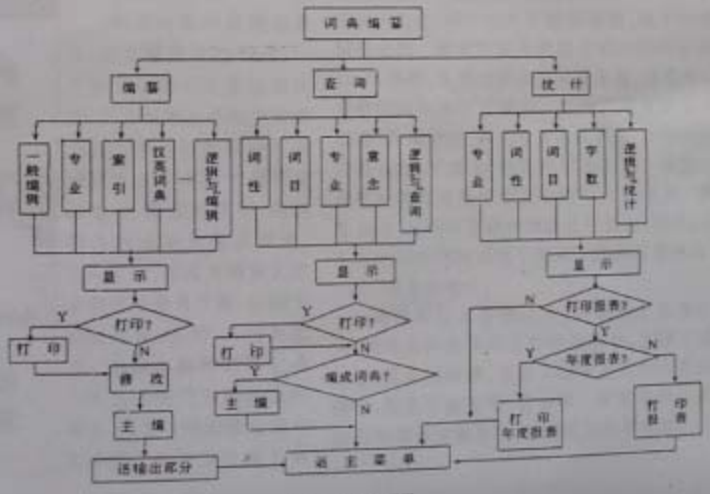


图5 词典编辑流程图