中國語文研究 2008 年第 1 期 (總第 25 期)，頁 1-8

粤语历史文献数据库的制作与应用

杨荣祥
北京大学中文系、北京大学汉语语言学研究中心

提要 粤语历史文献数据库共包括十七种十九世纪三十年代到二十世纪四十年代的粤语历史文献。本文简单介绍了十七种文献的内容、体例，回顾了文本录入、校对、数据库设计、入库文本预处理、灌库的历时八年的过程，总结了录入校对复杂文本、为复杂文本建立数据库的经验、教训，详细说明了粤语历史文献数据库的库表结构、检索功能、应用价值和应用方法。

关键词 粤语历史文献、数据库、库表结构、检索系统

粤语历史文献数据库是根据香港研究资助局援助的研究项目《近代粤语的演变——早期广东话话语材料研究》(The Cantonese language: Its past as reconstructed from early colloquial texts)(项目编号：HKUST6055/02H)之研究需要设计研制的。

数据库的制作启动于 1998 年。项目主持人张洪年(香港中文大学)先生经过多年的努力，从欧美等地收集到一批 19 世纪到 20 世纪初刊印的粤语文献，这批文献比较真实地记录了当时的粤语面貌，对研究粤语的历史演变甚至汉语的历史演变都具有重要价值。但由于这批文献原来分散在世界各地，一直没有受到学者的重视。张洪年先生将这些宝贵的材料收集起来，决定和北京大学蒋绍愚教授共同负责组织人员进行系统的研究，这是很有意义的。但是，由于这批材料的文本性质很复杂，仅对单一文献进行研究显示不出其重要价值，而要将各种文献联系起来研究，又必须对各种文本进行统一的处理才能有效地使用，因此决定为这批文献创建一个专门的数据库。

一 文献情况简介

张洪年先生交付北京大学汉语语言学研究中心制作数据库的粤语文献共计有以下十七种：

(1) Cantonese made easy(广东话入门)，第一版本：1883 年刊行。
第一部分为前言，内容包括对广东方言语音、词汇、语法的全面说明，有音类表，附有例字；第二部分为课文，每课内容用两页，第一页左栏为英语，右栏为广东话，第二页左栏为罗马注音，右栏为英语单词对译；第三部分为语法，按词类分章，以英语说明为主，有广东话例句，例句后均附罗马注音。第四部分为附录，是对广东话的附加说明；第五部分为索引。
(2) Cantonese made easy(广东话入门)，第二版本：1888 年刊行。
对第一版本有修订、扩充。
(3) Cantonese made easy(广东话入门)，第三版本：1907 年刊行。
对前两种版本有修订、扩充。语音系统有很大的变化，特别是声调方面记录了大量的变调。

(4) **Cantonese made easy**(广东话入门)，第四版本：1924 年刊行。
对第三种版本有修订扩充。
(5) **Chinese Chrestomathy**(汉语读本之广东方言)，**1841** 年刊行。
第一部分为长达 36 页的前言，内容包括对广东方言语音、词汇、语法的说明，对中国古代典籍的详细说明。其中典籍说明部分 18 页，分为三栏，左栏为英译，中栏为四库全书总目的节选，右栏为罗马字注音；第二部分分习唐话、身体、亲谊、人品、日用、贸易、工艺、工匠、耕农、六艺、数学、地理志、石论、草木、生物、医学、王制十七篇，每篇内又分若干章，以第一篇习唐话为例，其中包含教学用语，三字经，写毛笔字的方法，汉字的结构和书写等内容。大致上正文分三栏，依次为英文、汉语、罗马注音，每页下半部分是详细的英语注释。其他篇各章内容有词汇，有白话，有广东话，有古文。
(6) 英语不求人(A Chinese and English Phrase Book)，1888 年刊行。
为广东人学习英语而编写，按主题编排，每句英文下使用(汉字 粤语)注音。又将英文翻译为粤语，再给粤语句子的汉字注音。
(7) 广东省土话字汇(VOCABULARY OF THE CANTON DIALECT)，1828 年刊行。
PART I，196 頁。以英语字汇为纲，后附广东话字汇及罗马字注音。
PART II，92 頁。以广东土语字汇为纲(按罗马注音排序)，先标注音，次列汉字，若一字汇有不同说法，则附列于后，并用罗马字母注音。
PART III，357 頁。谚语、警言、成语等汇编，按事类编排；每条的排列为：注音(不标调)、汉字、英译。
(8) 《注音广州话》，1935 年刊行。
粤语常用词语和句子，用国语注音字母注音。
(9) **First Year Cantonese**(初级广东话)，1927 年刊行，1941 年第二版。
有注音和英譯。
(10)《新编广东省城白话》，20 世纪 30 年代刊行。
共分三集，无英语翻译和注音。
第一集，1-51 课为按义类排列的白话词汇，52-70 课为广东话语汇及俗语。
第二集，通用白话问答，按主题编排，包括对话和篇章两种。
第三集，醒世故事，均为篇章。
(11)《分類通行廣州話指南》，1930 年刊行。
内容为广州话常用词语、句子，分类编排。無注音，無英譯。
(12) **Beginning Cantonese**(教話指南)，1927 年刊行。
這是一部詞典性質的教材，以單字爲綱，單字注音；單字下列出該字在詞組和單句中的用法。
(13)《麥仕治廣州俗話・書經解義》卷壹，1833 年刊行。
用廣州俗話解說《尚書》；無注音，無英譯；
(14) **Everybody's Cantonese**，1955 年第四版。
有注音；有英譯。
(15)《散語四十章》，1877 年刊行。
广州话常用词语、句子。無注音；無英譯。
(16)《粵語》，1932 年刊行。
学习粤语的课本，共 70 课。無注音；無英譯。
(17) **How to Speak Cantonese**，1888 年刊行。
粤语教材。分三栏。中栏是粤语，左栏是对粤语句子的英文翻译，右栏是对粤语句子中一个一个词的英文对译。没有注音。

这十七种文献，文本的编排体例各不相同，使用文字符号也各不相同。有的文献同时有

粤语汉字、英文、罗马注音；《注音广州话》使用了国语注音字母；另外还有一些民间专业用的奇特符号，如木工用的计数符号等。罗马注音使用的符号也有各种各样的差异。

根据研究的需要，我们完成了以上十七种文献的全部录入工作，1-7 种已经灌库，8-11 种作了精细校对。12-17 种有电子本录入，有注音的文本已经整理出音节表。

二　文本的录入和校对

如前所述，这批粤语文献的文本性质相当复杂，录入并不是一件很容易的事情。刚接到这些文本时，我们把问题想得很简单，马上组织研究生和部分本科生分别将各个文本录入电脑。根据项目组的要求，录入时尽量保持文本原貌，但大量的方言字当时的电脑字库中没有，一些音标和调号电脑的普通字符中也没有。当时我们组织录入者开会，要尽量在电脑中找相似的字符。这样，各个录入者通过不同的办法将普通字库和符号库中没有的文字和其他符号录入文本中，有些甚至是临时造字。这种处理办法给后续的校对工作以及入库处理带来了一系列的麻烦。这些麻烦的根本原因是电子文本的保真要求与电子文本的便利使用之间的矛盾造成的。

保真——尽量保持纸本文件的原貌，是对电子文本的一项高标准要求，也是一项基本的要求。如果是普通的文字文本，只要认真录入、精心校对，是能够达到保真要求的，但对于这批粤语历史文献，要做到完全保真却存在相当大的困难。

第一，原文本是活字排版印刷，原字体与现在的通行字体有差别，比如，“度”，原本上面的“点”作一小“横”，要保真就得新造一个字符，录成现在的字型，就跟原本有差别；“鹅”，原本作左“鳥”右“我”，这在通用字符里没有，扩充字符里也没有。类似这种笔型、结构部件与通用字符有差异的情况很多，有些可以在扩充字符库中找到，有的根本找不到。找不到的，要保真就得新造字；找得到的，有的因为字体不同，根据电脑所安装的软件不同，有些电脑就不能显示，或者即使能够显示，但不能查询。同时，有的录入人员为了保证录入速度，就直接用了一个通行字符，就没有达到保真的的要求。对于这类问题，课题组多次开会都提出来进行了讨论，最后的处理办法是，笔型有异的字，直接用今天通行的字；结构部件位置与通用字有异而大字符集中找不到的字，采用与之大致对应的大字符集中有的字。

第二，原本中有大量的异体字。这与第一点有联系。对于文本中的异体字，我们开始时没有充分的认识，所以有些录入人员全都录成了通用字。可是有些异体字对于研究文字的变异是有价值的，有的还可能反映了语言的变化。针对这种情况，我们在校对时统一要求：凡是大字符集中能够找到的字，一律按原本录入，大字符集中找不到的字，用与之形体最相近的异体字录入，差异太明显的，采用“合成字”。各个文本的异体字使用情形不一样，要保证标准统一，就必须对各个文本的异体字进行整理。这项工作非常艰巨，必须对文本反复阅读、排比，找出异体字，然后拿出统一的处理意见。如果每种文本都要经过整理异体字的工作程序，将要花费大量的时间和人力。对此，我们先选用 CME(1888)作了一个实验，整理文本中的异体字，探询异体字的使用规律和处理办法。其他文本，根据 CME(1888)的经验，在校对中对异体字作统一的处理。

第三，原文本有错讹，是否保真？如原文本将“茶”写作“荼”，将“船”写作“般”，而注音用“茶”、“船”的读音。对录入人员来说，只能照录，但对研究者来说，这将提供错误信息。这样的错讹必须由懂粤语的人校对才能发现。所以，三校以后的校对，我们请了母方言为粤语，又是粤语研究专家的杨必胜(北京联合大学教授)先生来做，将这类错讹全部记录下来，对原文本采用加出校记的办法来处理。这样既保持了原文本的真实面貌，又避免了电子文本给研究者提供错误信息。

第四，注音符号复杂，有许多符号是通用符号中所没有的。录入时，录入人员只好用插入符号的办法来解决。还有一些符号电脑里根本没有，录入人员就用一个近似的符号代替。这样做的结果，一是给校对增加了工作量——校对者认为这个符号不对，改过来，植入电脑

时可能改用了另一个符号。可是校来校去，虽然跟原本接近了，但不符合最后的入库要求——因为字体不统一，入库后根本无法辨认。这个问题曾经困扰了我们很长时间。后来就请杨必胜先生将所有有注音的文本中的注音进行整理，编制出音节表，看各个文本共出现了那些注音符号，然后每个符号作统一字体的处理。统一字体用 Times New Roman，Times New Roman 字体中没有的符号用近似符号。必须这样做，注音材料入库后才能被有效地检索。

注音符号中最难处理的是声调符号，其中又以 CME(1907)和 CME(1924)的变调符号最为复杂。对这些符号的电子文本处理，直到 2004 年着手进行入库时才决定采用现在的办法——统一用 InPanAdd 声调符号，该符号集中没有的符号，规定用某一符号统一代替。所以现在的电子文本中的声调符号是最不“保真”的。必须这样处理，否则电脑无法进行数据处理，因为声调在注音检索中有着决定性的作用。

第五，原本的格式情况复杂。如四种 CME 文本，原本每个页面两栏，两个页面合起来才构成一个整体：一栏粤语汉字，一栏粤语注音，一栏英文句子翻译，一栏英文单词对译。如果分四栏录入，既不能保持原本面貌，页面显示也不方便；如果按原本分两栏连续录入，将无法显示四栏之间的联系，入库后四栏之间的内容无法关联，失去了研究价值。最后我们的电子本采用三栏录入，将两种英文放入一栏，但入库时进行了处理，分成四栏关联，如检索“佢做賊偷我野咯。”这个句子，我们会查到每个汉字的粤语注音，这个句子的英文翻译，每个单词的英文对译。如果仅检索其中的一个字如“佢”，我们也能够找到这些信息。同样，我们检索这个句子中相对应的任何一个项目，如注音“ʦkʻui”、英文“He is a thief, and has stolen things of mine.”、“steal”，也会得到所有相关的信息。有些文本，横排中间偶然加进直排的表格或板块，录入时也不能完全保真，只能改变为统一的板式。

以上谈的是电子文本保真与电子文本便利使用之间的矛盾给录入、校对带的困难以及我们采取的解决办法。下面简单说一下现在所得电子文本的形成过程。

先期的工作就是将所有文本录入电脑，随后进行第一次校对。校对全用打印的纸本，标出校改的地方，再植入电脑。但第一校的植入没有对所使用的字体作明确要求，导致同一文本中使用了多种不同的字体，这对文本的入库是无效的，所以后来的二校、三校、四校、终校我们都用专人负责植入。在校对中提出前文所述遇到的各项问题并寻求解决办法。这期间因为找不到粤语研究专家的帮助使项目进展受到阻碍。后来请到杨必胜先生，但一个人要承担的工作太多：要制作音节表，要校对电子文本，所以时间拉得很长。

课题组首先设计了一个音节表，请杨必胜先生整理出各个文本的音节及声、韵、调系统，再根据这个系统中所出现的每一个标音符号检索文本，检索一个，剔除一个，全部剔除后还有标音符号的话，就只有两种情况了：一种是音节表漏收了，需要补充修改音节表；一种是录入的符号不正确，需要校改为正确的符号。这样保证每个注音符号都与原文本对上之后，用 Times New Roman 字体和 InPanAdd 声调符号进行统一替换，这就做到了各个文本的注音符号具有了一致性，从而保证入库后程序能够有效识别，研究者能够有效检索。

由于文本本身非常复杂，加上请不到专业校对人员，所以对电子文本只好采取多次校对的办法以减少错误率。对前文列出的十七种文本中的前十一种，我们都进行了五次(或五次以上)校对加一次抽查。我们估计，前七种文献的错误率最多在万分之三左右，次四种文献的错误率最多在万分之四左右。

在这批文献的录入校对过程中，我们走了不少弯路，也得出了一些经验教训。文本的录入、校对是制作数据库的基础工作，必须达到尽可能高的准确性和检索的有效性。要做到这一点，首先要对文本的内容、性质、复杂性有深刻的了解；同时，要明确文本的使用目的，一开始录入就要想到今后使用起来是否方便。其次，多人录入不同文本，必须对格式、字体作统一要求，任何插入符号，字体必须与全部文本所用字体一致。第三，对异体字、缺字有统一的处理意见。因为是方言材料，有不少怪字电脑字库中没有，我们最后采用了“合成字”的办法。第四，各个文本在录入、校对时遇到的疑难问题要有详细记录，将这些记录集中起

来，由课题组统一提出处理意见。

三　数据库的设计与材料入库前的预处理

在原始文献录入电脑的基础工作将要完成的时候，课题组提出了一个“粤语文献资料库”需求说明，主要内容包括“资料库系统的总体要求”、“建库前的文本预处理”、“检索需求”。系统的总体要求有：(1)开放性，即数据库必须具有可扩充能力，包括可增加同类型文本入库，增加其他类型文本入库(如新的类型的粤语文献、其他方言的资料、近代汉语文献等等)；具备标注(tagging)扩充能力，即可在本数据库的基础上，对库中资料进行自动或人工标注(如分词、注明词类、语法结构、特殊语音标记等等)。(2)关联性。(3)多平台迁移能力。(4)易操作性。(5)可维护性。“文本的预处理”要求检索结果明确标示每一项目的出处、类别、年代信息等。“检索要求”包括字词检索、语法查询、语音查询。

根据需求说明和对录入文献的性质的分析，我们于 2001 年春提出了数据库建立的基本指导思想，并于 2002 年春在北京大学召开了课题组会议进行讨论。讨论结果：在检索目标方面将“语法查询”和“语音查询”分为两个不同的库实现，“词汇查询”与“语法查询”使用同一个库；先使用关系数据库建立“粤语文献语音检索系统”。“语法检索系统”放在后一步。在“粤语文献语音检索系统”中留下接口，今后再将中古音库和现代广东方言数据库扩充进去。“语法查询”涉及到对文本作自动分词、标注词性以及其他各种语法信息标注等问题，不是本项目的时间和经费所能够完成的，所以只能等以后有条件再做，但本数据库应该能够提供单词、语句、句式等方面的查询。关于语音检索系统，应做到：音节查询——某个音节在单个文本中出现的频率、在所有文本中出现的频率、对应的汉字、对应的英文(如果有英文翻译)、出现的句子、出现的句子的有关文本信息等等；声母、韵母、声调、韵头、韵腹、韵尾的分项查询——能够获取与音节查询同样的各种信息；单字查询——选择任何一个汉字，都能得到该汉字的所有语音信息以及该汉字的所有出处。这样，同音异字、同字异音、同声母的字、同声调的字、同韵母、同韵头、同韵腹、同韵尾的字都可以任意查询。

确定了数据库的设计要求后，随后对一种文本——CME(1888 年版)进行入库前的预处理。由于文本性质复杂，预处理中遇到了很多问题。

首先是入库材料的取舍问题。相对数据库的需求而言，现有语料包含两种不同性质的材料，一是主要信息，包括汉字符串(单字、词、句、篇章)、罗马字注音符号(声、韵、调)、英译；二是附加信息，包括前言、语音系统、词汇现象、语法现象说明、正文下的注释文字、文化知识等。附加信息如果全部入库，数据库的程序将非常复杂，且预处理的工作量将大大增加。考虑到这部分信息对研究者来说，价值不是特别大，而且只要通过文本的阅读就能了解，决定这部分材料不入库。

其次，需要为每一种文献建立声母表、韵母表、声调表和同音字表；比较各种文献的音节表，理清不同文献注音符号的异同和声调系统之间的关系；最后将各种文献的四种表进行合并；提取每一种文献中全部的单字，并列出相对应的罗马字注音、英文翻译和出处，建立汉字、注音、英译三种资料之间的严格对应表。这项工作花费了近半年的时间(仅 CME(1888 年版)一个文本)。

第三，程序对单个汉字的识别很容易，但对音节、英文单词的识别需要另行规定。经与工程师讨论，确定用“声调符号+空格键/回车键”辨认注音符号的音节单位，因为每个音节都一定有声调(其实《广东省土话字汇》PART III 的注音就没有声调)；用“前空格键……后空格键/回车键”辨认英文单词。这个办法对数据库的程序设计来说不困难，但对预处理来说，就遇到以前没有想到的问题：所有文本在校对时，对空格键是没有留意的，因为无论是纸本校对还是电脑直接校对，多一个或少一个空格键都是不容易发现的。对此，只好再花时间对预处理本进行校对。

2003 年年底，终于完成了 CME(1888 年版)的预处理，并实验入库。入库后的结果并不

理想，主要是提供关联项错误太多，音节分析错误太多，有些项目的检索无结果等等。这些问题，有些是程序设计本身有缺陷造成的，有些是语音系统的整理有误造成的，有些是文本校对不精造成的，有些是预处理时关联项对应不准造成的。经过将近一年时间的调试，一方面工程师不断根据检索需求修改程序，一方面对预处理文本进一步校对，到 2004 年 9 月，CME(1888 年版)的数据库已经调整到了比较理想的状态，能够满足各项查询需求。

有了一个文本的经验，其他文本的入库就可以少走弯路了。当然，因为每个文本的类型都不一样，需要在保真的前提下作一致性处理，特别是注音符号的一致性处理，也不是一件很容易的事，因为任何一项处理，都必须保证程序能够识别。经过 2005 年至 2006 年 8 月份一年多时间的努力，终于完成了其他六种入库材料的精细校对和预处理，这期间，预处理要反复多次进行，每一次都要进行灌库实验，发现问题再进行加工。程序也要不断地根据新的文本类型进行修改，所以经常需要工程师、预处理人员、对文本情况熟悉的人员等多方人员聚在一起，逐个解决入库时临时遇到的各种问题。到 2006 年 9 月底，其他六种文献全部入库，并实现了七种入库文献的合并。为了让数据库同时方便内地和港澳台地区的用户使用，我们随后又将程序设计改为 UNICODE 码再进行预处理和入库，至 2006 年 11 月，入库全部完成。现在的数据库所用内码为 UNICODE 码，内地和港澳台用户都可以使用。

四　库表结构和数据库的使用

根据检索需求，库表结构主要包括以下内容：

(1) 以单字为单位将所有的语料经过预处理进入数据库。
(2) 字段包括：单字、对应的罗马注音、对应的英文翻译、频次统计、出处(书名、年代、章节、页码)。
(3) 记录项为每一种文献中全部的单字；如果字同注音不同，则区分为不同的记录；如果字同注音同而出处不同，则不列为新的记录，只在同一记录的“出处”字段内标出，同时由频次统计体现该单字在该文献中的全部出现次数。
(4) 罗马字注音字段进一步区分出声、韵、调作为检索条件，韵母部分又进一步区分韵头、韵腹、韵尾作为检索条件，其实现方式为使用声母表、韵母表、声调表进行匹配检索。
(5) 数据库包括检索系统和数据库使用说明，另附有用于维护系统的程序。
(6) 本数据库为单机版，内码为 UNICODE 码。
(7) 为便于今后扩充规模，本数据库留有多个接口。今后可以加入多种不同类型的文本，包括各种方言材料、汉语历史文献。可以增加高级语法查询，对入库文献进行语法信息标注。可以增加中古音库并与方言语音对应。

下面简单介绍一下数据库的使用。

本数据库可以检索任何一个汉字、字组、句子、搭配字组(如“因为……所以……”;“佢……野咯”——“佢”开头“野咯”结尾的句段等等)、标点符号；可以检索任何一个音节、声母、韵头、韵腹、韵尾、声调以及注音的任意两项、多项的搭配形式(如“上声+声母 p+韵尾 m”)；可以检索英文单词、句子。如果文献是汉字、注音、英文对应，检索任意一项都可以得到对应项，并获得所有出现检索项的例句及其文本信息。

查询界面：包括“按汉字查询”、“查询中文例句”、“查询英文注释”、“按注音查询”(下设“声母、韵头、韵腹、韵尾、声调”五个选项)、“查询的语料版本”(可以选择七个文本中的任意一种、二种至全部)、“查询结果”——包括出现的用例的总数、用例序号、汉字例句、例句注音、英文注释、版本信息等。

按汉字查询：就是按照语料库中记录的汉字字形进行查询，输入要查询的汉字，选择要查询的文本，即可得到所有出现该汉字的例句。GBK 字库中无法寻得的方言字，采取合成字

手段，如{口话}，七个文本中出现的所有使用合成方法的方言字都在《粤方言历史语料检索系统使用说明》中列出。

查询中文例句：指对语料库中感兴趣的中文例句进行查询。在输入框中输入要查询的中文例句即可得到所有的这个句子及其相关信息。还可以在输入的中文例句中使用通配符从而实现模糊查询。如输入“不但%而且”，可查询到所有包含“不但……而且”句式的例句。输入“民[族权]”，可查询到所有包含“民族”或“民权”的例句。输入“北京__大学”，可查询到所有以“北京”开头，并以“大学”结尾，开头和结尾之间包含两个任意字符的例句。等等。

按英文注释查询：指对语料库中感兴趣的英文注释进行查询。在输入框中输入要查询的英文注释即可获得全部条目及其相关信息。还可以在输入的英文注释中使用通配符从而实现模糊查询，通配符的使用方法同上。检索英文单词时，为避免所检索英文单词与其他英文单词中的部分音节重合而导致检索混乱，要在被检索的英文单词前后增添空格“#”，如检索“lock”，应输入“#lock#”。

按注音查询：指对语料库中广东话注音进行查询。选择该种查询方式，可以分别对注音的声母、韵头、韵腹、韵尾、声调以及它们的各种组合进行查询。使用该查询方式时，需要注意：

1） 保证该种查询方式必须被选中，即“按注音查询”复选框必须被设置为选中状态。

2） 声母、韵头、韵腹、韵尾以及声调，这五个注音查询项，在任何一个下拉查询项中选择“*”，则表示系统查询将不考虑该项的值。

3） 声母、韵头、韵腹、韵尾以及声调，这五个注音查询项，在任何一个下拉查询项中选择“ ”(空)，则表示系统需要查询满足该查询项为空的所有注音。

4） 当同时设置多个非“*”查询项，则表示将这些选择项的组合进行查询。例如，按“k+ +a+*”选择，会选取出注音：“kan”、“kat”、“ka”等，而按照“k+ +a+ ”选择，则只会出现注音：“ka”

语料版本的选择：查询系统启动后，系统数据库中所有的语料版本默认地都将显示在查询条件面板的第一排，供查询选择，选择某一个或某几个查询版本仅需在语料版本前点击鼠标将复选框设为选中状态即可，如果选任何版本项，系统将默认地在所有版本语料库中对数据进行查询。

显示查询结果：在“查询条件设置面板”中，查询条件设置完毕后，点击【OK】查询按钮，系统将按照设定的查询条件进行数据检索，查询完毕后，查询结果将显示在“查询结果显示区”中。

保存查询结果：如果需要保存当前视图中的查询结果到本地文件，则进行如下的操作：

1) 菜单“语料查询视图管理”→ “保存查询结果” →选择文件保存的文件目录→ 输入所保存的文件名称→ 点击【保存】。
2) 保存的文件是 rtf 格式，可以使用 word 打开该文件。

五 结语·鸣谢

本数据库历时近八年时间，由于没有经验，又没有一个完全固定的工作队伍，工作进展一直不是很顺利。现在这个结果，可能还存在着种种不足，希望在使用中不断发现问题，及时进行修正。

对于性质复杂的语料，要建立数据库，关键的问题是一开始就要对全部文本的性质有全

盘的认识，要有一个明确的一致性标准，不能中途随意变更一致性标准。

文本的保真只能是相对的，不能是绝对的。为了同时满足电子文本的保真和数据库的使用方便两方面的要求，最好是录入、校对出一分高保真度的电子文本，而在入库前的预处理中，主要考虑程序设计简便、检索方式方便、检索结果关联性强、不同文本的一致性强等因素，对保真文本再作一致性处理。还有一种办法是，考虑到纸本文献不易获得，可以将所有纸本文献制作成 PDF 文件，以供研究者核实原本。

本数据库的制作是在香港中文大学教授张洪年先生(原香港科技大学教授)和北京大学汉语语言学研究中心蒋绍愚教授的领导下、由笔者组织实施完成的。在整个制作过程中，得到了很多人的帮助：中国科学院信息自动化有限公司的梁高中先生为本数据库设计了整套程序；北京联合大学杨必胜先生作为粤语研究专家，为我们整理了所有带注音的文本的音节表，整理了这些文本的语音系统，还对多个文本作了终校。北京大学邵永海教授为数据库的建设作了大量的组织管理工作，并协助工程师设计程序。鲁东大学(原烟台师范学院)文学院姜仁涛副教授带领一批研究生帮助完成了《汉语读本之广东方言》、《英语不求人》、《土话字汇》、《注音广东话》、《初级广东语》《新编广东省城白话》、《分类通行广东话指南》七种文献的最后整理和校对工作。北京大学中文系汉语史专业研究生吴坚同学利用课余时间连续为本项目工作了五年时间，特别是后面两年，协助工程师做了大量的工作。中国社会科学院语言研究所李蓝研究员曾参与对文本性质的研究和音节表的设计，北京大学中文系宋绍年教授参与了项目的管理，北京大学中文系 1997 级本科生李予湘、杜轶、张岩、徐世良、姜南、2001——2004 级硕士研究生李予湘、杜轶、李伟群、运娜、黄高飞、洪琰等近二十位北京大学中文系的同学曾参与过文本的录入、校对工作。对于所有给予本项目提供帮助的人，在此我们一并表示衷心的感谢。

The Establishment and Application of the Database of Historical Cantonese Literature

YANG Rongxiang

Abstract The Database of Historical Cantonese Literature consists of seventeen historical literary works on Cantonese from 1830s to 1940s. This paper briefly introduces contents and styles of these works, and reviews the eight-year process of text input, proofreading, design of database, pretreatment of database text, data-input. The experience in inputting and proofreading complex texts and in establishing database for complex texts has been generalized. The structure, retrieval function, application value and methods of the Database of Historical Cantonese Literature are explicated.

Keywords Historical Cantonese Literature, Database, Structure of Database, Retrieval system