《全衡》网上中文输入系统的词典建设

张小衡 张群显
香港理工大学中文及双语学系
ctxzhang@polyu.edu.hk

摘要：《全衡》是第一个面向香港和国际实际需要的功能较为全面的网上中文输入系统，其核心部件是词典。一般的输入法依据的是简单的“汉字-输入码”对照表，语言知识较贫乏；《全衡》使用的是一个拥有五万多词条的词典，每一词条讲述一个词语，信息包括该词语的简体字、繁体字、汉语拼音、粤语拼音、仓颉码、速成码等。由其中任何一项入手，借助于系统中的检索程序可以方便地查找其它各项信息。这不仅有力地支持了汉字输入，对于汉语学习也很有好处。本文将专题讨论《全衡》的词典设计、建立与编辑。

关键词：网上中文输入、词典编辑、汉语拼音、粤语拼音、简体字、繁体字

Dictionary Development for the AllBalanced Web-Based Chinese Character Input System

Zhang Xiaoheng and Cheung Kwan-hin
Department of Chinese & Bilingual Studies, Hong Kong Polytechnic University
ctxzhang@polyu.edu.hk

ABSTRACT: AllBalanced is the first Web-based Chinese character input system with substantial functions to meet the needs of Hong Kong in particular and the needs of the international society in general. The primary knowledgebase of the system is a dictionary of over 50,000 word entries encoded in Unicode. The contents of an entry include the traditional characters of the word, the simplified characters, the Hanyu Pinyin, the Jyutping, the Changjie code and the Sucheng code. The present paper discusses the development of the dictionary.

Key Words: Web-based Chinese character input , Dictionary editing, Hanyu Pinyin, Jyutping, Simplified Chinese character, Traditional Chinese character

1 引言

《全衡》(AllBalanced)是由香港理工大学中文及双语学系网上中文输入研究小组研制的一个中文输入系统，其设计原则是以香港的各种需要为重点，同时照顾其它各地的需要，全面考虑，以达最佳平衡。

《全衡》在 WWW 上工作,提供粤语拼音和汉语拼音等四种输入法，每种输入法都支持直接输入繁体字和简体字。通过《全衡》输入法查找到的每一个字词都可以方便地查询其繁体字、简体字、汉语拼音、粤语拼音、仓颉码、速成码等有用信息，支持中文学习。据了解,《全衡》是第一个较全面考虑香港和国际需要的网上中文输入系统。它的总体概况已另有文章介绍[1]，本文详细介绍和深入讨论系统的核心部件–词典，内容包括该词典的设计、建立与编辑。

2 要求与设计

《全衡》的词典是整个汉字输入系统的主要知识源。研制词典的指导思想是：面向香港，面向国际，注意语文规范，支持多种输入法，兼顾汉语学习。 内地的中文输入软件主要是为简体字和普通话服务，台湾的中文输入软件主要是为繁体字和具有台湾色彩的普通话服务，面向香港的中文输入软件不仅要照顾到普通话、繁体字和简体字，而且要重视粤语，还要考虑到在本地影响较大的速成和仓颉输入法。因此我们把词典设计为带有六个域的数据表，信息包括每个词[1]的简体字、繁体字、汉语拼音、粤语拼音、仓颉码、速成码等，如图一所示。其中的附加列“词次”用于存放各个词的使用频率。为方便汉字输入，汉语拼音的标调方式采用数字式，用 1、2、3、4 加在音节末尾表示第几声，轻声用 5 表示，字母 ü 用 v 表示。

词次	繁体字	粤语拼音	简体字	汉语拼音	速成码	仓颉码

图一　词典的设计格式

[1] 为方便起见，如无特别说明本文说的词包括词典中的单字，单词和一些固定词组。

因为《全衡》将在 WWW 的环境中运作，且同时支持繁体字和简体字，所以词典的内码规定为 Unicode。软件在网上工作时对用户操作的反应速度会受到比较大的负面影响，但是汉字输入系统的信息检索功能却要求有相当高的实时性。面对这一限制，词典的规模不宜太大，暂时设为五万左右。这种规模的词典还有一个好处，就是可以在 MS Excel 上编辑，不需用到数据库管理系统，这对于非计算机专业的语言工作者来说会带来不少参与上的方便。

3 词典的建立与初步处理

《全衡》的词典是以《粤语拼音字表》[2]，香港理工大学三地中文语料库词频表和 Windows98(内地中文版)全拼输入法的词表为基础按图一所示的格式建立起来的，其初始数据是通过对上述三个资料源的内容进行取舍加工而获得的。

3.1《粤语拼音字表》的处理

《粤语拼音字表》是香港语言学学会粤语拼音字表编写小组编写的。该书从香港的实际出发共收 10675 个繁体字和未简化字，粤语标音采用香港语言学学会《粤语拼音方案》[2]。该方案的特点包括：能同时发挥坊间各种粤语拼音的功能；所用字符全可用标准键盘直接打出，方便电脑输入和处理；以粤语语音系统分析为根据，系统性较强；有别于一般方案，本方案可拼粤语任何音节；在语音和字母的对号方面同时参考国际音标、汉语拼音方案和香港一般的拼音习惯。因此《粤语拼音字表》是一个面向香港的有利于汉字输入和汉语学习的字表。

把《粤语拼音字表》的汉字全部收进《全衡》的词典，要通过若干步骤的处理。首先，由于字表的电子版本是用大五码（BIG5）2表示的， 所以需要转换为 Unicode。字表里共有带调粤拼、汉字、大五码和仓颉码四项内容。首先删除大五码项，把其它三项数据放入《全衡》的词典中。接着，用 Word2000 的繁简字体转换功能得到相应的简体字，利用南极星文书处理软件的拼音标注功能为每个汉字产生汉语拼音，但标调方式改为数字式。每个字的速成码是通过取该字的仓颉码的首尾两个字母得来的，可通过 Excel 的字符串处理函数来自动生成。电脑产生的汉语拼音和简体字都可能有误，需要人工修订，这将在第四节中详细讨论。

3.2 全拼输入法词表的处理

---

[2]大五码是台湾和香港等繁体字地区所常用的内码和交换码。

全拼输入法是 Windows98(内地中文版)上的一个内置汉语拼音输入法，由北京中易电子公司和微软共同研制，Windows2000 的英文版、内地中文版和台湾版也都配有这种输入法。全拼输入法的码表共收汉语字词 56713 个，以“简体字-汉语拼音”的形式编写，拼音不管声调。

我们首先将原用 GB 内码表示的全拼输入法码表文本文件转换为 Unicode 的 Excel 文件，然后将码表中的单字条目删除。不采用该码表的单字条目是因为其中绝大多数的字同时出现在《粤语拼音字表》中，已经收进《全衡》的词典，其余都是一些非常用的字，而且会给以后的粤拼标注带来困难，因为粤拼标注是以《粤语拼音字表》为依据的。

留下的多字词条目的繁体字形式是通过南极星的简-繁字体转换功能得来的（用 Word2000 也可以进行转换）。至于粤语拼音、仓颉码和速成码的自动产生却没有现成的软件工具可用，因此我们自己编写了一个简单的代码标注程序，它能根据一个给定的“单字-代码”对照表来为一个汉字词表产生代码，如果遇到一字多码的情况，则每个可能的代码都写出来待人工选定。例如“长远”的无声调粤语拼音代码是“coeng/zoeng jyun”。接着我们就用这个代码标注程序和从词典已有内容中抽取的各种单字-代码对照表给从全拼输入法码表中选取的多字词产生粤拼码（无调号）、仓颉码和速成码。最后再把这个包含六项内容的词表加到词典中去。

3.3 三地语料库词频表的处理

这个词频表是从香港理工大学的六百万字现代汉语书面语语料库中生成的，内容包含语料库的所有用词（共 61150 个）及每个词的使用次数。语料库的抽样语料来自 1989~1992 年中国内地、香港和台湾三地的十种影响较大的报刊：香港的成报、明报和信报，内地的人民日报、北京晚报、新明晚报和羊城晚报，以及台湾的中国时报、中央日报和联合报。

三地语料库词频表原本是一个关系数据库表，采用大五码和繁体字。处理时，先把该文件转变为采用 Unicode 的 Excel 文件。接着把词频表中表示字词使用次数的数字复制到词典中相同字词的词次列中，然后将这些重复的字词和其它单字从语料库词频表中删除。经上述处理后，词频表还有约四万词条，所以我们又把使用次数低于 3 的（因而也不可能在三地报纸中同时出现的）低频词删去。其中次数为 1 的有两万余条，次数为 2 的六千多条。接着通过南极星为原有的繁体字词语产生相应的简体字形式和无调汉语拼音，然后用自制代码标注程序和相应的单字-代码对照表给这些词语产生粤拼码、仓颉码和速成码。再把最后结果加到词典中。

经上述处理，初步建立起来的词典含有约五万四千词条，符合原计划。

4 编辑整理工作

词典建立起来以后，里面还有不少欠妥的地方需要编辑修订，这主要由人工逐条检查处理，工作相当艰巨费时。检查的内容主要在简体字、繁体字、汉语拼音和粤语拼音这几项，主要的依据是《新华字典》[3]和《现代汉语词典》[4]，同时参考一些其它文献[5-10]。以下分几个方面来介绍和讨论。

4.1 更正错误词条

词典中的错误有来自单字词条的，也有来自多字词条的。

4.1.1 单字词条的修订

单字条目中的繁体字和粤语拼音两项内容来自《粤语拼音字表》，比较可靠，因此检查的重点是简体字和汉语拼音。汉语拼音方面错误比较多，尤其是多音字的情况。由于当计算机为单字标音时，无上下文可依，所以每个多音字都用同一个音来处理，例如："长"字本有 chang2 和 zhang3 两种读音，但南极星将它的汉语拼音一律标成 chang2。修改时还要同时注意音字的匹配正确，例如"长"字普通话读 chang2 时应该对应粤语拼音 coeng4，读 zhang3 时应该对应粤音 zoeng2。多音字如出现同一粤语拼音音节对应多个汉语拼音音节， 则增加新行，反之也然，使得两种拼音的每一种合法对应都占一行。例如"的"字在粤语中只有一种读音，原来在词典中只占一行，但在普通话中"的"有三种可能的读音，因此需增加两行，如图二所示：

词次	繁体字	粤语拼音	简体字	汉语拼音	速成码	仓颉码
160967	的	dik1	的	de5	hi	hapi
160967	的	dik1	的	di2	hi	hapi
160967	的	dik1	的	di4	hi	hapi

图二 多音词的处理

简体字是根据繁体字转换生成的，由于"一繁对多简"的情况很少，因此错误也很少。错误者如：繁体字"乾"对应于简体字"乾"(qian2)和"干"(gan1)，但是都被转换成"干"。

有些转换错误的简体字，其正确形式在 Unicode 字集中找不到，因此暂用原繁体字代替。例如"顏"的简体字暂用"顏"本身代替。这样处理比造字可取，因为自己造的字在 Web 的其它电脑上是显示不出来的。好在这种缺字的情况为数很少，而且一般都是一些使用频率极低的单字。

4.1.2 多字词条的修订

多字词条的修订既涉及到原有数据又涉及到电脑产生的数据。在原数据方面，首先是由于全拼输入法的词语的汉语拼音都限制在 12 个字母以内，所以不少词的拼音需要补全。例如“百思不得其解 baisibudeqij(ie)”，“百闻不如一见 baiwenburuyi(jian)”、“辩证唯物主义 bianzhengwei(wuzhuyi)”。这样处理不仅有利于规范拼音输入，也有利于汉语学习。

有些拼音在前十二个字母中也有错误。例如全拼输入法原词表中“首都机场”的拼音写成 shoud*ou*jicha，“雄才大略”的拼音是 xiongcaidal*u*。应该分别改为 shoudujichang 和 xiongcaidalve。

多字词的汉字表达也有一些错误，例如原来的全拼码表中有一个词为“指出稹”应改为“指出”。

至于电脑自动产生的数据，用自制工具产生粤语拼音时对于多音字的处理是几个音都给出，修订时应该将不正确的去掉。如粤拼 sing zoeng/coeng（成长），zoeng/coeng gong（长江）分别选定 sing zoeng 和 coeng gong。多字词中由南极星和 Word 产生的简体字、繁体字和汉语拼音也存在少量错误需人工更正。

4.2 删除次要的词条

这主要是为了节省电脑存贮空间和遵守语言规范，删除的对象是那些使用效率低的和规范性差的词语。例如有些短语非常松散，按照语文规范的划词原则[11]很难作为词条收入词典中，例如：“谁也没有想到”、“从表面上看”、“从长远来看”、“第八个五年计划”等等。另外，有些短语的部件词语可以在本词典中较快找到，这样的短语也可以考虑省去。例如:删去“出厂价格”，因为“出厂”和“价格”都是已有的低重码词。类似的例子有“科普读物”（科普、读物）、“课堂教学”（课堂、教学）、犯罪集团（犯罪、集团）等。

为了进一步减少词典在计算机中占用的内存量，汉语拼音和粤语拼音等输入码最大长度限定为 20 个字符，汉字词语长度限定在 6 个字以内。虽然超过这种限制的词语只有几十条，但有些相当长（例如：中华全国妇女联合会），将它们删除之后，词典表格中的好几个列所需的宽度能有效地减少，因此当词典在计算机中通过数组或数据库的结构来表示时，就可以大大节省资源。

4.3 增加重要的词条

增加的词语要求是基本的和实用的。这些词语大多是在使用《全衡》的试验版时发现欠缺的。新增的词语包括以下几类：

A. 常用词

例如：共同语、双语、多语、字词、字音、字符集、异体字、主语、谓语、宾语、补语、定语、状语、句法、词法、语义、语用、词表、国标码、大五码、重码、笔顺、引号、顿号、副词、介词、标点符号、空格、仓颉、圣经、耶稣、道教、朱镕基、江泽民。

B. 新词

例如：万维网、互联网、浏览器、主页、首页、网上、网页、网址、网站、全衡。

C. 香港地方词

例如：九龙、新界、红磡、沙田、罗湖、点钟、强积金、硬碟、粤拼、行政长官、董建华、特首、董特首。

D. 常用短语

有一些常用短语结构紧密且汉字量少，分成更小的单位来分析和输入时有所不便，因此将它们当作词条加入词典中，以提高输入效率。 例如：看起来、看在眼里、可不可能、可不可以、可不是、老实说、那还用说、那就是说、颇受、颇感、人人、深表、身负、身居、深有、誓不、无可、無所、自相 、来讲。

4.4 综合处理

有些词条的处理不是单纯的增、删或改，而是综合处理。例如，原词典中有“带头作用”，但是没有“带头”，因此增加单词“带头”， 删去“带头作用”。因为词语“带头”比“带头作用”用处更广，而且后者可以通过词典中的“带头”和“作用” 方便输入。 类似的例子还有：

凭票供应→凭票，

华北地区→华北，

经研究决定→经研究，

不包分配 →不包，

兩國關係、兩國人民、兩國之間→兩國

相同之处→之处

新兴产业→新兴

种子选手→ 选手，

最好成绩→最好

三地语料库词表中有一些词的写法是错误的。处理时要看该词的正确形式是否也在词典中，如果没有则将其写法改正，如果有则应该将错词的词次加到正确的词上然后将错词删去。例如繁体字词：（制造、製造）→製造，（面粉、麵粉）→麵粉，（头發、頭髮）→頭髮。发现这种词重复现象的有效方法是把词典分别按照简体字和繁体字排序后再检查，因为重复词往往在某种字体的写法上是一致的。

经研制小组半年多来的努力，全衡的词典日趋成熟，而且已经开始在系统上试用。图三是词典现状的一个片段。(词典中多字词的汉语拼音和粤语拼音还未标声调，仓颉码也未填妥。其实仓颉输入法逐字输入也能达到相当高的效率。)

词次	繁体字	粤语拼音	简体字	汉语拼音	速成码	仓颉码
1132	甚至	samzi	甚至	shenzhi	tvmg	
1129	了解	liugaai	了解	liaojie	nnnq	
1129	主席	zyuzik	主席	zhuxi	ygib	
1128	長	zoeng2	长	zhang3	sv	smv
1128	長	coeng4	长	chang2	sv	smv
1128	組	zou2	组	zu3	vm	vfbm
1128	宣布	syunbou	宣布	xuanbu	jmkb	
1123	水	seoi2	水	shui3	e	e

图三：《全衡》词典片断

5 结论与讨论

《全衡》是第一个较全面考虑香港和国际的需要的网上中文输入系统，其中心知识源是词典，虽然该词典的词条数目不算很大，但词条的信息却相当丰富，包括简体字、繁体字、汉语拼音、粤语拼音、仓颉码、速成码和词频等方面的内容。显然这种词典对于汉字输入和汉语学习都有较高的实用价值。但是编辑工作也相当艰巨，应该说我们只是做了一些初步的工作，词典中还存在许多不够完善的地方需要处理。

一个急需解决的问题是，由于三地语料库没有拼音标注，所以由此产生的词频信息对于一字多音的情形不加分别，只是笼统地给出该字在语料库中出现的总次数。例如在词频表中，位居版首的“的”的词次为 160967，但没有说明其中“的 (de5)”、的 (di2)”、“的 (di4)” 各占多少次。这对于拼音输入法会有不良的影响，因为我们不希望打入汉语拼音“di2”或“di4”后得到的第一个候选字是“的”。为此，我们准备根据《现代汉语字频统计表》[12]和《现代汉语频率词典》[13]来填补这个缺陷，使得词典中的每一个“字-音”组合都有合理的词频数字。其实对于汉字输入而言，我们并非一定要精确的词次，只要能表示词频的层次关系就行了。

另一个不够完善的地方是：现在词典中的单字拼音是带调的，但是多字词的拼音并没有标调。为了使用户能灵活选用标调、不标调或部分标调的拼音输入码，词典中所有的词的汉拼和粤拼都应该标调。多字词的仓颉码也应填妥。

此外，目前字典中的单字条目只有一万来个。应该将 Unicode 中的繁简汉字都收进来。另外，还要大大增加香港常用字和常用词。

当然待完成的工作远远不只这些，由于编幅有限未能一一说明。我们也衷心希望同行的学者专家们提出宝贵意见。

虽然我们的词典还很不成熟，但在它支持下的《全衡》试验系统却相当令人满意，关于该系统的功能特性和技术实现将另文介绍讨论。

参考文献

[1] Zhang, X. AllBalanced: A Web-Based Chinese Character Input System to Meet Hong Kong's Needs. Proceedings of the 19th International Conference on Computer Processing of Oriental Languages (ICCPOL'2001), Seoul, Korea. May 14-16, 2001. pp 333-338

[2] 香港语言学学会粤语拼音字表编写小组.《粤语拼音字表》.香港：香港语言学学会, 1997.

[3] 中国社会科学院语言研究所.《新华字典》.北京：商务印书馆，1998.

[4] 中国社会科学院语言研究所.《现代汉语词典》.北京：商务印书馆，1996.

[5] 上海词书出版社.《词海》.上海:上海词书出版社, 1999.

[6] 上海大词典出版社.《汉语大词典》.上海：上海大词典出版社, 1997.

[7] 商务印书馆.《普通话-粤音 商务新字典》香港：商务印书馆，1991.

[8] 商务印书馆.《普通话-粤音 商务新词典》香港：商务印书馆，1990.

[9] 语文出版社.《汉语拼音词汇》. 北京：语文出版社，1989.

[10]杨子来.《标准中文输入码大字典》.香港：聚贤馆文化有限公司，1996.

[11] 国家标准局.汉语拼音正词法基本规则.见：语文出版社.《语言文字规范手册》.北京：语文出版社.1997.

[12] 中国国家标准局,国家语言文字工作委员会.《现代汉语字频统计表》.北京：语文出版社, 1987.

[13] 北京语言学院语言教学研究所.《现代汉语频率词典》.北京：北京语言学院出版社，1986.

鸣谢：本课题得到香港理工大学的两次资助，香港语言学学会提供《粤语拼音字表》,北京中易电子公司提供汉语拼音码表，香港理工大学中国语文教学中心提供三地语料库词次统计表。特此鸣谢。