《全衡》词典的设计与建设①

张小衡 张群显

(香港理工大学中文及双语学系 香港)

摘要:《全衡》是第一个较全面考虑香港和国际的需求的网上汉字输入系统,其核心部件是词典。《全衡》使用的是一部拥有六万余词条的词典,每一词条讲述一个词语,信息包括该词语的简体字形式、繁体字形式、汉语拼音表达式、粤语拼音表达式、仓颉输入法代码、速成输入法代码等。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帮助。本文简要介绍《全衡》的词典建设。

关键词:网上汉字输入;词典编辑;汉语拼音;粤语拼音;简体字;繁体字

中图分类号: TP391.1

Design and Development of the AllBalanced Dictionary

ZHANG Xiao-heng CHEUNG Kwan-hin

(Department of Chinese & Bilingual Studies Hong Kong Polytechnic University Hong Kong)

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 societies in general. The primary knowledgebase of the system is a dictionary of over 60,000 Chinese word entries encoded in Unicode. The contents of each word entry include the traditional characters of the word, the simplified characters, the Hanyu Pinyin expression, the Jyutping expression, the Changjie code and the Sucheng code. The present paper presents a brief introduction to the design and development of the dictionary.

Keywords: Web-based Chinese character input; Dictionary editing; Hanyu Pinyin; Jyutping; Simplified Chinese character; Traditional Chinese character

一、引言

《全衡》(AllBalanced)是由香港理工大学中文及双语学系网上汉字输入研究小组研制的一个汉字输入系统,其设计原则是:以香港的利益为重点,同时照顾其它各地的需求,全面考虑,以达最佳平衡。

《全衡》在 WWW 上工作并直接采用 Unicode 编码,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。下图是这些功能的运用实例与说明。

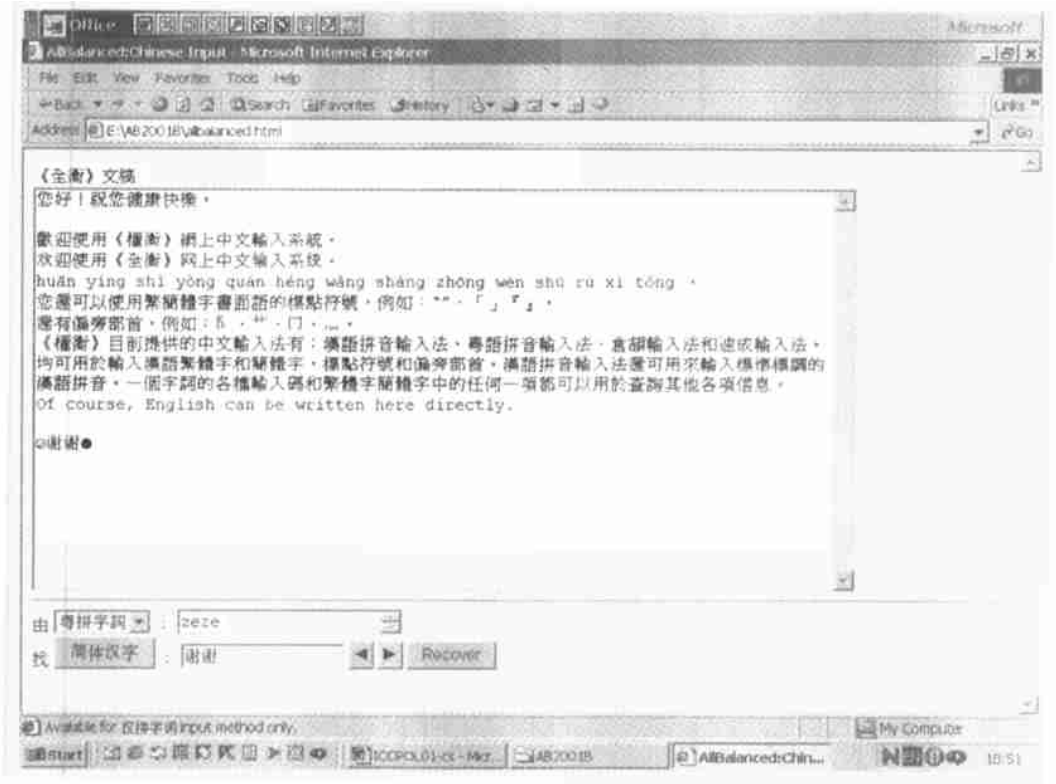
关于《全衡》的特性与功能已另有专文叙述[1],本文将简要介绍和讨论系统的核心部件

① 收稿日期:2001－11－14

本文得到香港理工大学研究资金(GS968;1－9827)支持

作者张小衡,男,1958 年生,博士,助理教授,主要研究领域为计算语言学和计算机辅助教学. 张群显,博士,副教授,主要研究领域为语言学和中文信息处理.

——词典，内容包括该词典的设计、建立、编辑与现状等方面。



《全衡》多功能运用实例与说明图

二、需求分析与词典设计

《全衡》词典是整个汉字输入系统的主要知识源。研制该词典的指导思想是：面向香港，面向国际，注意语文规范，既支持汉字输入，又兼顾汉语学习。内地的汉字输入软件主要是为简体字和普通话服务，台湾地区的汉字输入软件主要是为繁体字和（具有台湾地区色彩的）普通话服务，而面向香港的汉字输入软件不仅要照顾到繁体字、简体字和普通话，而且要重视粤语，还要考虑到在本地影响较大的仓颉输入法及其简化版速成输入法。因此，我们把词典设计为一个数据表，含七个域（列），分别用于记录汉语词① 的使用频率、简体字形式、繁体字形式、汉语拼音、粤语拼音、仓颉码和速成码七项信息。

为方便汉字输入，汉语拼音采用数字标调方式，用 1、2、3、4、5 加在音节末尾分别表示第一至第四声和轻声，字母 ü 用 v 表示，ê 用 e^表示。

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。

另一个问题是，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。通过多次试验，我们发现词条总量为五至六万较为合适，这样既能满足一般写作词汇量的

① 为方便起见，如无特别说明本文说的“词”包括词共中的单字，单词和一些固定短语。

需要，又能保证可接受的速度。这种规模的词典还有一个好处，即可在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。

三、词典的建立与初步数据处理

《全衡》的词典是在 MS Excel 2000 上按上述的设计要求建立起来的。资料主要来自《粤语拼音字表》[2] 和香港理工大学三地汉语语料库词频表。更确切地说，《全衡》词典的初始内容是通过对上述两个资料源的数据进行取舍和加工而获得的。

3.1 《粤语拼音字表》的数据处理

《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，共收 10675 个繁体字（包括未简化字），字条内容包括繁体字、粤拼、大五码 和仓颉码四项。粤语标音采用香港语言学学会的《粤语拼音方案》[2]。

我们把《粤语拼音字表》的所有汉字及其粤拼和仓颉码全部收进《全衡》的词典。接着，用 Word 2000 的繁－简字体转换功能为原有的繁体字产生相应的简体字，利用南极星文书处理软件[3] 的拼音标注功能为每个汉字产生汉语拼音（采用数字标调方式）。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（如果仓颉码为单字母，则取一个字母），这可通过 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，这将在第四节中讨论。

3.2 三地语料库词频表的数据处理

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，包含语料库的所有用词（共 61150 个，用繁体字形式）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、香港和台湾地区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，以及台湾地区的中国时报、中央日报和联合报。为满足词典的设计规模，只收录词频表中词次为二以上的多字词（包括二字词）及其词频数字。

接着通过 Word 和南极星为这些繁体字词条产生相应的简体字和带调汉语拼音。至于粤语拼音、仓颉码和速成码的自动产生却没有现成的工具软件可依，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”对照表来为一个词表产生代码，如果遇到一字多码的情况，则将每个可能的代码都写出来待人工选定。例如“长江”的粤语拼音代码暂时写成“coeng4/zoeng2 gong1”。利用这个代码标注程序和从《全衡》词典已有数据中抽取的各种单字－代码对照表，我们给从词频表中选取的多字词产生了粤拼码、仓颉码和速成码。

经上述各项自动化或半自动化处理后，我们建立起了一个拥有约四万九千词条的初始词典，每一词条都带有简体字、繁体字、汉语拼音、粤语拼音、仓颉码、速成码和使用频率（不在三地语料库词频表中出现的字词的频率暂定为 0）等七项信息。

四、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，这主要由人工逐词条检查处理。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》[4]、《现代汉语词典》[5]、《普通话－粤音 商务新字典》[6] 和《普通话－粤音 商务新词典》[7]。

4.1 更正错误词条

词典中的错误有来自单字词词条的，也有来自多字词词条的。单字词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，但南极星却将它的汉语拼音一律标成 chang2。修改时还要注意普－粤字音匹配正确，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。需要时，则增加记录，使得每一个字词的粤－普读音的每一种合法对应都占一行。

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”对应简体字“乾”(qian2)和“干”(gan1)，但是都被转换成“干”。少数转换错误的简体字的正确形式在 Unicode 字集中找不到，因此暂用原繁体字代替。例如“[illegible]”的简体字暂用“[illegible]”本身代替。这样处理比造字可取，因为自己造的字在 Web 上的其它电脑是显示不出来的。

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，来自三地语料库词频表的多字词有些有误，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。例如繁体词：(制造、[illegible]造)→[illegible]造，(面粉、[illegible]粉)→[illegible]粉，(头发、头[illegible])→头[illegible]。

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 sau2 zoeng2/ coeng4(首长)，zoeng2/ coeng4 gong1(长江)分别改为 sau2 coeng4 和 coeng4 gong1。多字词中由南极星和 Word 产生的简体字和汉语拼音也存在少量错误需人工更正。例如，南极星虽然能为“重庆”标上正确的汉语拼音“chong2qing4”，但是“重生”的汉语拼音却写成“zhong4sheng1”。

4.2　删除次要词条

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，按照语文规范的分词原则[8]很难作为词条收入词典中，例如：“从表面上看”。另外，有些短语按其部件词切分时并不影响意思表达，且这些部件词都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”，因为这个短语可从“出厂”和“价格”来理解，而且两个部件词都是已有的低重码词。

此外，为了提高系统的工作效率，我们将词长限定在 5 个字以内。

4.3　增加重要词条

增加的词语要求是基本的和实用的。这些词语大多是我们在使用《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、字符集、词表、国标码、重码、笔顺、空格、主语、朱[illegible]基。

B.新词　例如：万维网、互联网、浏览器、主页、首页、网页、网址、网站、全衡。

C.香港地方词　例如：九龙、新界、红[illegible]、点钟、强积金、硬碟、粤拼、董建华、特首。

此外，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，可以提高输入效率。例如：看起来、可否、颇感、深表、身负、深有、誓不、自相、来讲、网上、上网、很好 等等。

4.4　其它处理

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“相同之处”，但是没有“之处”，因此增加“之处”，删去“相同之处”。因为词语“之处”比“相同之处”用处更广，而且后者可以通过词典中原有的“相同”和新增的“之处”方便地输入。类似的例子还有：最好成绩→最好，两国关系－两国人民－两国之间→两国等。

五、现状与结论

经研制小组近两年来的艰辛努力,《全衡》词典日趋成熟。下表是词典现状的数据片段。

《全衡》词典片断

词频	繁体字	粤语拼音	简体字	汉语拼音	速成码	仓颉码
10	行进	hang4zeon3	行进	xing2jin4	hn一yg	hommn一yog
10	行进	haang4zeon3	行进	xing2jin4	hn一yg	hommn一yog
377	行业	hong4jip6	行业	hang2ye4	hn一td	hommn一tctd
11	行当	hong4dong1	行当	hang2dang5	hn一fw	hommn一fbrw

由上图可见,现在的词典与旧版[9]相比有几个明显的改进:a).同单字词一样,多字词的汉语拼音和粤语拼音都标上声调;b).多字词的速成输入码中的字码之间用连字号分开,以便用户查阅;c).多字词的仓颉码也填上,使得《权衡》的所有输入法都能支持以字和词为单位的中文输入。另一个重要的进展是单字的词频数字改用《现代汉语字频统计表》[10]的数据,使得每个多音字的每种"字一音"组合都有合理的频序标示。此外,单纯来自全拼输入法词表的词语已基本删除。

《全衡》目前词条(记录)数目达60056,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,但其编辑工作是长期而艰巨的。现有内容虽然已经过两次人工检查,但其中一定还存在不少欠妥之处需要更正。

我们还打算增加一些有用的内容。例如,目前词典中所收的单字只有一万来个,应该将Unicode中的20902个简繁体汉字都收进来。另外,还要增加香港常用字和常用词。除主词典外,还应该为用户提供方便自行编辑使用的用户词典。我们还打算给词条内容增加一个"英语译文"项,以便英一汉词语互查。这样既能为熟悉英语的人提供一种辅助性汉字输入法,同时也为一般的中英双语学习带来一点方便。

总之,语言是活的,《全衡》的词典内容也应该根据时代的发展不断丰富和更新。我们衷心希望同行的学者专家们提出宝贵意见。

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》,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。

参　考　文　献

[1] 张小衡.《全衡》网上中文输入系统——功能与特性.自然语言理解与机器翻译.北京:清华大学出版社.2001,428一436

[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] NJStar Software Corp.:http://www.njstar.com/

[4] 新华字典.北京:商务印书馆,1998

[5] 现代汉语词典.北京:商务印书馆,1996

[6] 普通话一粤音 商务新字典.香港:商务印书馆,1991

[7] 普通话一粤音 商务新词典.香港:商务印书馆,1990

[8] 汉语拼音正词法基本规则.语言文字规范手册.北京:语文出版社.1997

[9] 张小衡、张群显.《全衡》网上中文输入系统的词典建设.自然语言理解与机器翻译.北京:清华大学出版社.2001,419一427

[10] 现代汉语字频统计表.北京:语文出版社,1987