

拼音—汉字自动句转换系统的支撑环境和开发工具

孟洁 陈群秀

智能技术与系统国家重点实验室

清华大学计算机科学与技术系, 北京 100084

E-mail: cqx@s1000e.cs.tsinghua.edu.cn

摘要: 本文讨论了为研制拼音—汉字自动句转换系统而设计的支撑环境和开发工具—字音转换工具、语料抓取对齐工具和音字转换评判工具的实现, 以及其涉及的算法分析, 并探讨了其今后的进一步完善。

关键字: 拼音—汉字句自动转换系统、支撑环境、开发工具、字—音转换工具、多音字、语料抓取对齐工具、分词、拼音—汉字转换评判工具

the Supporting Environment and Tools of Pinyin—Chinese Character Automatic Conversion System Based on Sentence Input

Meng Jie Chen QunXiu

The State Key Laboratory of Intelligent Technology and System

Department of Computer Science and Technology, Tsinghua University, Beijing 100084

Email: cqx@s1000e.cs.tsinghua.edu.cn

ABSTRACT: The paper presents the implementation of Chinese character—Pinyin automatic conversion tool, sentence picking up and aligning tool, and Pinyin—Chinese character conversion evaluation tool, all of which are designed for the research and implementation of Pinyin—Chinese character automatic conversion system based on sentence input. The paper also discusses the improvement and direction for the future work.

Keyword: Pinyin—Chinese character automatic conversion system based on sentence input, supporting environment, developing tools, Chinese character—Pinyin conversion tool, multiphonetic character, sentence picking up and aligning tool, word segmentation, Pinyin—Chinese character conversion evaluation tool

1 引言

随着计算机软硬件技术的飞速发展和计算机信息产业的兴起, 随着电脑汉化的必然进程, 汉字键盘输入技术已成为我国当前中文信息处理和计算机应用的关键技术之一。汉字输入技术的研究也成了中文信息处理中十分活跃和引人注目的领域。

汉字键盘输入是汉字输入的主流。目前, 面对着“万码奔腾”的“春秋战国”时代,

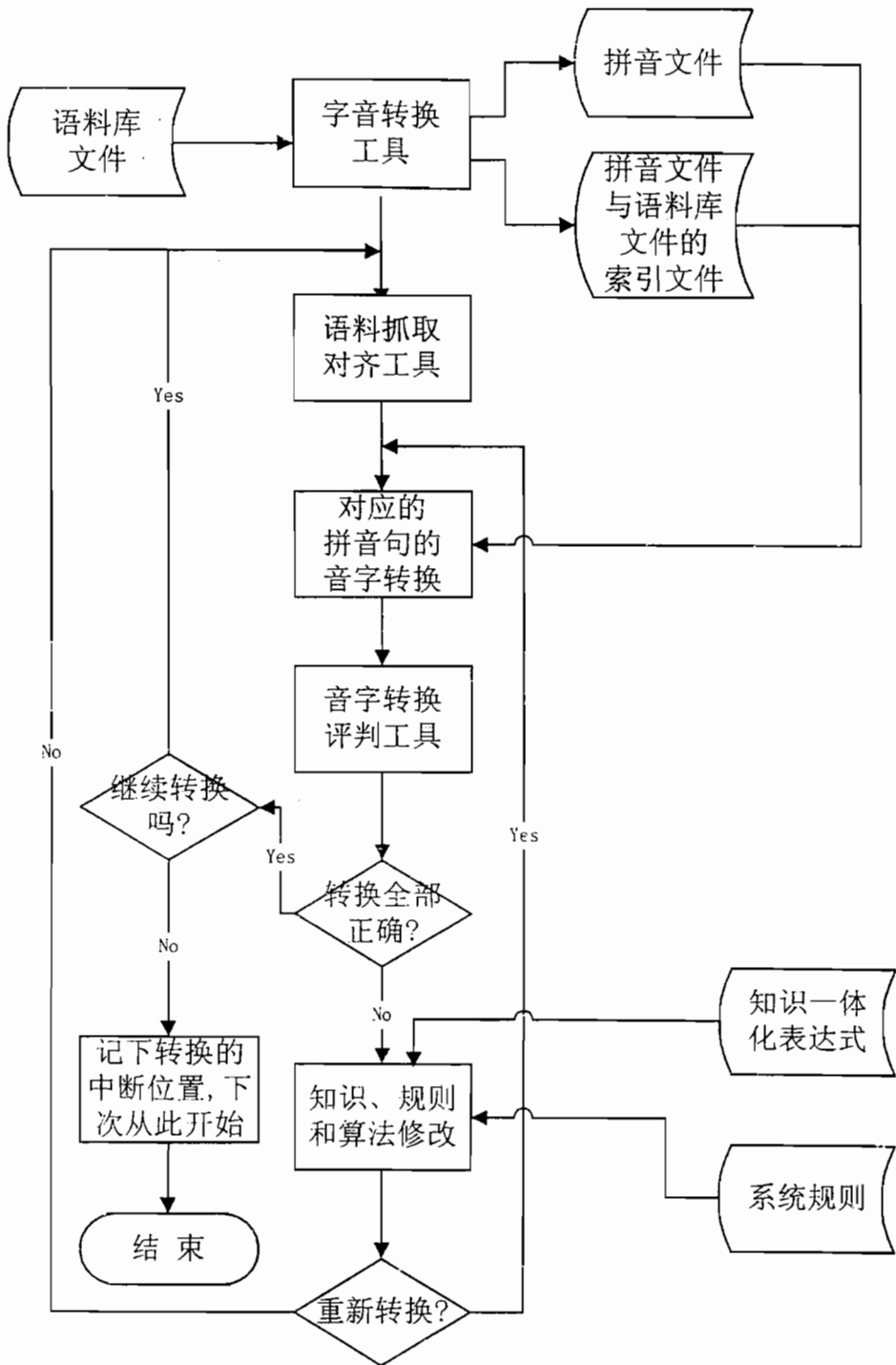
由于众多的汉字编码输入法各有优势和缺点，难以裁定孰优孰劣；政府主管部门未能选出法定的汉字输入法。颇有见地的学者提出两个重要的问题和动向：一是拼音输入法将是今后中文输入法的首选输入法，二是汉字键盘输入法最重要的发展方向是智能化。而这两种动向的结合产生的新动向是智能型拼音—汉字句自动转换系统的兴起，智能型拼音—汉字句自动转换系统的研制势在必行，且意义深远。

我们的“智能型拼音—汉字自动句转换系统”能根据用户已输入的汉语句子对应的拼音串，综合使用相应的汉语的词法、句法、语义和语境知识，动态转换出当前最可能的汉字串，当输入一句拼音后，用户就得到一句最可能的对应汉语句子。在句转换过程中，用户不需要进行屏幕选字；而且由于系统具有自动的学习和改进的功能，因此兼具高效转换和简单易学的特点。系统的转换依据是字、词、短语各语言理解层面的汉语语言知识和大量描述特定词语出现的典型上下文、并带有各部分分数的规则（当一部分匹配成功时，这部分就加上相应的分数从而在同音字、词的竞争中排名上升，容易被选出）。

2 系统的研究路线

我们的“智能型拼音—汉字自动句转换系统”采取如下的研究路线：一是规则方法和统计方法相结合路线；二是既在归纳规则时利用大批语料（句例）作为基础，又在修改完善每条规则时采用大规模语料库提供大量句例进行转换试验和自动评测，找出知识和规则的不足之处进行修改，然后再进行大量句例的转换试验和自动评测，再修改，再测试，直至效果满意；三是进行大规模转换试验时不是用人工录入拼音流，而是采用字音自动转换得到大批的拼音句子流。因此，我们不仅构筑了一个覆盖新闻、军事、经济、中小学教材、文学小说、应用文的大型语料库，而且研制了一个支撑环境和一些开发工具。我们的研究路线请参见图一。

图一中，“音字转换工具”、“语料抓取和对齐工具”和“音字转化评判工具”是本文重点论述的“拼音—汉字自动句转换系统”的支撑环境和开发工具。首先，我们使用“音字转换工具”将语料库中的文件都转换成对应的带有索引的拼音文件。第二步，我们使用“语料抓取对齐工具”在语料库文件中寻找那些包含指定汉字串或拼音串（关键字）的句子。当关键字是拼音串时，我们对照“字音转换工具”得到的拼音文件，并根据索引文件，输出对应的汉字句；当关键字是汉字串时，我们直接查找语料库文件，得到对应的汉字句。我们可以分析这些抓取出来的汉字句例，并用加分规则对关键字的语言环境进行描述。然后，我们将通过索引文件，将这些汉字句例对应的拼音句作为“智能型拼音—汉字自动句转换系统”的输入。然后将音字转换系统的输出和原来对应的语料库中的汉字句例进行对比，计算出转换正确率，并保存转换错误的所有信息，这就是“音字转换评判工具”的功能设计。如果转换结果不令人满意，系统可提供用户进行系统知识、规则和算法的编辑接口（系统中还没有实现算法的编辑接口）。最后，系统可根据测试的需要选择继续进行同音字转换的测试，也可以重新选取关键字，进行语料库句例的抓取和对齐，重新进行一轮转换、评判和修改。关于“拼音—汉字自动句转换系统”的知识一体化表达式，具体论述请参见 ISMT&CLIP 国际会议论文《汉语的一种知识一体化表示方法》^[5]；关于“拼音汉字自动句转换系统”，请参见我们相关的论文。



图一：智能型拼音—汉字自动句转换系统的研究路线图示

3 “拼音—汉字自动句转换系统”的支撑环境和开发工具

3.1 字音转换工具

语料的更新和规整是进行语料库字音转换的一个必要的前期处理。语料的收集工作必须考虑到具体的使用，我们根据音字转换的性质和要求，选取了包括新闻、军事、经济、中小学教材、文学小说、应用文等各种题材的语料。为了便于下一步的语料库文件的字音转换处理，需确保语料库文件的规整性，即无乱字符，半角字符需成对出现。为此，我们设计了一个规整程序，将一些半角字符转成了相应的全角字符，去掉了一些格式字符，将一些不规则的符号（如不常用的标点符号等）转换成相应的全角符号，等。语料的规整大大方便了字音转换处理中的读句过程。

词表的编辑和审核是字音转换的一个必不可少的步骤。首先对一字词进行注音。若是多字词，则将其所有可能的读音标注上；然后，对于二字词、三字词……七字词词表，采取查找一字词词表的方法，对每个词的每个字进行顺序注音，注音的结果可能是：

(1) 一字词中有此字的条目，且注音唯一，不是多音字，则将此唯一拼音标注在此字在此词的对应位置上。

(2) 一字词中有此字的条目，且注音不唯一，是多音字，则将各种可能的注音依次标注上并加上多音字标记‘*’。

(3) 一字词中无此字的条目，则在相应位置上不标注任何信息。

经过如上的处理后，各词表中各词的唯一标注部分已确定下来，后面的工作是在多音字的众多拼音中选择一个在该词中的正确读音，并给没有注音的字人工注音。这些工作以及此后的审核校对都是手工完成的，这就保证了词表的可靠性。词表文件的行格式如下：

词 词频 第一个字的拼音 第二个字的拼音 …… 第 n 个字的拼音

其中，n 的最大值为 7。

语料库字音转换处理的目的是为下一步语料抓取对齐过程提供相应的拼音文件和索引文件，以解决按拼音流抓取时匹配查找的问题。此程序关键是解决多音字的音标注问题。

为了解决多音字的正确选音问题，我们根据具体问题的性质，选择了一种简易实效的分词和排歧策略。稍加分析可以看出，对于字音转换中特定的多音字的选音问题，机械分词方法已基本能胜任。考虑到正向最大匹配法分词策略的错误率为 1/169；且为了能通过一定的排歧方法进一步提高分词命中率，我们采用二次扫描法进行正反向两次最大匹配，再采用基于互信息的排歧策略进行裁定以选择一种较可能的分词结果。采用二次扫描法进行分词产生交集型歧义时，我们计算相应词串的各词词频之积，较大者为最后分词结果。例如字串 $C_1C_2C_3$ 在正、反向最大匹配分词过程中有两种分词结果 C_1C_2/C_3 和 C_1/C_2C_3 ，则引进“互信息”，其定义为

$$I(C_1C_2) = \log_2 [P(C_1, C_2) / (P(C_1) * P(C_2))]$$

表示两字的结合紧密程度。其中 $P(C_1, C_2)$ 表示字 C_1, C_2 的同现概率， $P(C_i)$ 表示字 C_i 出现的概率。则：

$$I(C_1C_2, C_3) = \log_2 [P(C_1C_2, C_3) / (P(C_1C_2) * P(C_3))]$$

$$I(C_1, C_2C_3) = \log_2 [P(C_1, C_2C_3) / (P(C_1) * P(C_2C_3))]$$

其中 $P(C_1C_2, C_3) = P(C_1, C_2C_3)$ ，计算

$$I(C_1C_2, C_3) - I(C_1, C_2C_3) = \log_2 [(P(C_1) * P(C_2C_3)) / (P(C_1C_2) * P(C_3))]$$

可得, 若 $P(C_1) * P(C_2 C_3) > P(C_1 C_2) * P(C_3)$, 则 $I(C_1 C_2, C_3) > I(C_1, C_2 C_3)$, 说明: $C_1 C_2$ 同 C_3 的结合度大于 C_1 同 $C_2 C_3$ 的结合度, 所以 C_2 应同 C_3 组成词, 正确的分词结果为 $C_1 / C_2 C_3$ 。对于词频为零的词, 我们采用简单的平滑数据方法——加上一个小量, 如 0.1。实验证明, 运行情况良好。

为了方便使用, 本处理过程在命令行提供了一些开关参数以供选择, 这使得本处理具有以下功能:

- (1) 批处理: 可一次进行多个语料库文件的字音转换。
- (2) 自设拼音目标文件后缀名: 缺省的拼音文件后缀名为 phn, 但用户可自设文件后缀名。
- (3) 自设新闻标题最大长度: 处理过程中, 需要首先读取语料库文件中的句子。由于语料是逐行输入的, 而新闻标题一般比其他编辑行都短。为了能判断出当前读取行是否包括了一个整句, 需要设立一个标题行的最大字数界限, 字数超过此界限, 则将其作为普通编辑行处理; 否则, 此行为标题行, 一句在此截断。缺省值是 20。
- (4) 自设语料库文件的行最大字符数: 语料库文件的行最大字符数的设定是为了读句时能动态分配暂存整句的内存大小。缺省值是 100。
- (5) 自设词表文件名: 进行分词必须要有词表。此处理过程现提供的缺省词表是 bh1.lst—bh7.lst, 为了保证词表的可修改性和可维护性, 允许用户自定义词表, 只需将词表名列举在相应的词表列表文件中。

命令行格式为:

```
wtos FileName [/b] [/p:PyFileExtName] [/n:NewMaxTitleLen] [/l:MaxSenLen]
[/w:WordListFileName]
```

命令行参数	功能说明
wtos	语料库字音转换的执行文件。
FileName	若为批处理(即/b开关项有效), 则为此批处理文件名, 此批处理文件的行格式为: 待处理的语料库文件[Enter]; 否则, 即为待处理的语料库文件名。
/b	进行批处理, 决定 FileName 是批处理文件名还是语料库文件名。
/p:PyFileExtName	设定相应拼音结果文件后缀名, 若批处理开关项有效, 则批处理过程中的所有待处理语料库文件的拼音结果文件名后缀都是此指定名。否则, 则相应拼音结果文件后缀名取缺省值 phn。
/n:NewMaxTitleLen	设定标题的最大长度。此项的设立, 用来判断当前行是否为标题行。若行字数小于设定的最大字数, 则为标题行而结束读句过程转而进行分词; 否则, 且在当前行未找到句间隔符时, 说明此行为某句的一部分, 需要继续读下一行, 直至读出整句为止。若此项未选中, 则标题的最大长度取缺省值 20。
/l:MaxSenLen	设定语料库文件每行最大字符数。语料库文件的字音转换是逐行进行的, 此项的设立, 用来提供动态分配临时存句内存的大小。若此项未选中, 则语料库文件每行最大字符数取缺省值 100。
/w:WordListFileName	设定词表文件的列表文件。此列表文件的行格式为: 词表文件名[Enter], 且词表文件的排列顺序为一字词词表文件名、二字词词表文件名……七字词词表文件名。词表文件的行格式为: 词 词频 第一个字的拼音 第二个字的拼音 第 n 个字的拼音[Enter]。

3.2 语料抓取和对齐工具

语料的抓取对齐工具的功能是从语料库文件中寻找给定数量的含有相应拼音串或汉字串（关键字）的例句。抓取关键字可为汉字串、有音调或无音调音节流，并将抓取出来的句例的抓取拼音串或汉字串上下对齐，并生成相应的结果文件，以供系统规则的归纳、修改和测试。

基本算法是：根据关键字是字或音，决定打开语料库汉字文件或是经字音转换的拼音结果文件；然后进行匹配查找，将中间结果记录在临时结果文件中，当查找结束（已找到要求数目的句例或是所有文件查找完毕）后，在逐句调整，使得各句的关键字上下对齐。为了防止语料库文件的重复查找，每次查找结束后的文件指针位置都被记录下来，已查过的部分不会再次参与下次查找过程；另外，当用户需要查找的句数较多，单个语料库文件不可能有如此多的例句，则系统再次产生随机数以决定下一次待查找的文件，且记录下来已查找完成的文件，下次的随机数竞争将不会选中已查找过的文件。这样，若系统未找到足够的句例而退出，则表示所有的语料库文件均已查过，没有找到要求数目的例句。

命令行参数	功能说明
grasp	语料库抓取对齐的执行文件。
KeyWord_BatFileName	若为批处理（即/b 开关项有效），则为此批处理文件名。此批处理文件的行格式为：抓取关键字 抓取字数[Enter]。关键字可为汉字串、有音调拼音串和无音调拼音串。关键字为无音调拼音串时，每个音节后需加音节间隔符“%”。
/b	进行批处理，决定 KeyWord_BatFileName 是抓取关键字还是语料库文件名。
/e:PyFileExtName	设定相应拼音文件后缀名，若此项未选中，则拼音文件后缀名取缺省值 phn。
/n:GraspNum	设定缺省抓取句数，若此项未选中，则缺省值为 100。
/t:NewsMaxTitleLen	设定标题的最大长度。若此项未选中，则标题的最大长度取缺省值 20。
/l:MaxSenLen	设定语料库文件每行最大字符数。若此项未选中，则语料库文件每行最大字符数取缺省值 100。
/y:YlListFileName	设定词表文件的列表文件。此列表文件的行格式为：语料库文件名 相应拼音文件名 抓取句数[Enter]。若抓取句数缺省，则此关键字的抓取句数取决于/n 参数项的设定值，若/n 参数项未选中，则此关键字的抓取句数为缺省值 100。进行汉字串抓取时，打开语料库文件；进行拼音串抓取时，打开相应的拼音文件。

它提供以上命令行开关：

- (1) 批处理：可一次抓取多个关键字。
- (2) 自设抓取句数：对每个关键字，可设定相应的要求抓取的句数。缺省值为 100。
- (3) 自设语料文件：可由用户设定语料文件，只需将语料文件名及相应的拼音文件名列举在语料列表文件中。
- (4) 自设新闻语料中的新闻标题的最大长度。
- (5) 自设语料文件的行最大字符数。

命令行格式为：

```
grasp KeyWord_BatFileName [/b] [/e:PyExtName] [/n:GraspNum] [/t:NewsMaxTitleLen]
[/l:MaxSenLen] [/y:YlListFileName]
```

3.3 音转字系统评判工具

音字转换评判程序的功能是将音字自动转换后的汉字串与正确的汉字串（原文）进行对照，判断现有音字转换系统的某个音节或总体的转换正确率。它的基本设计思想是：先将文本通过开发环境中的字音转换程序转换成拼音流，然后以此拼音流模拟用户的拼音序列输入，得到系统拼音—汉字自动转换的输出；将此输出结果与原来包含正确汉字串信息的文本进行对照，标记出出错的地方，并得出转换正确率。

4. 举例

建立两个虚拟语料库文件 t1.tst 和 t2，内容如下：

(1)t1.tst

弟弟拿着个地球仪。

他递给我一把伞。

第三次浪潮正席卷全球。

敌人抵挡不住红军的进攻。

他一副低声下气的样子。

这等于为本地的农民打开了一条生财之路。

地对地导弹发射成功了。

(2)t2

帝国主义到底会不会走向终结？

一滴泪从她脸上滴下来。

王名有五十亩地。

邮递员给他送来一封挂号信。

一只甲壳虫爬上了那朵牡丹花的花蒂。

顾客是上帝，这是商场经营应掌握的真谛。

用整理好的带拼音的词表 bh1.lst—bh7.lst 进行字音转换（命令行：wtos t1.lst[Enter]、wtos t2[Enter]），生成拼音结果文件 t1.phn 和 t2.phn，将 t2.phn 改名为 t2.mph，然后编辑语料库列表文件 ylfile.txt 如下：

```
t1.tst
t2 t2.mph
```

进行如下语料抓取对齐

(i)grasp di4 /n:14

结果文件--_di4.res 内容如下:

[帝]国主义到底会不会走向终结?
王名有五十亩[地]。
邮[递]员给他送来一封挂号信。
一只甲壳虫爬上了那朵牡丹花的花[蒂]。
顾客是上[帝],这是商场经营应掌握的真谛。
顾客是上帝,这是商场经营应掌握的真[谛]。
[弟]弟拿着个地球仪。
弟[弟]拿着个地球仪。
弟弟拿着个[地]球仪。
他[递]给我一把伞。
[第]三次浪潮正席卷全球。
这等于为本[地]的农民打开了一条生财之路。
[地]对地导弹发射成功了。
地对[地]导弹发射成功了。

(2)grasp di% /n:20

结果文件--_di%.res 内容如下:

[弟]弟拿着个地球仪。
弟[弟]拿着个地球仪。
弟弟拿着个[地]球仪。
他[递]给我一把伞。
[第]三次浪潮正席卷全球。
[敌]人抵挡不住红军的进攻。
敌人[抵]挡不住红军的进攻。
他一副[低]声下气的样子。
这等于为本[地]的农民打开了一条生财之路。
[地]对地导弹发射成功了。
地对[地]导弹发射成功了。
[帝]国主义到底会不会走向终结?
一[滴]泪从她脸上滴下来。
一滴泪从她脸上[滴]下来。
王名有五十亩[地]。
邮[递]员给他送来一封挂号信。
一只甲壳虫爬上了那朵牡丹花的花[蒂]。
顾客是上[帝],这是商场经营应掌握的真谛。
顾客是上帝,这是商场经营应掌握的真[谛]。

_di%.res 中共有 19 个例句。当用户要求抓取的句例数大于实际语料库文件中的例句数时,系统给出实际的全部例句。

(3)grasp 弟 /n:2

结果文件--_弟.res 内容如下:

[弟]弟拿着个地球仪。

弟[弟]拿着个地球仪。

当然，这三步语料抓取对齐也可作为批处理同时进行，只需生成批处理文件 yl.bat:

```
di4 14
```

```
di% 20
```

```
弟 2
```

并运行如下命令: grasp yl.bat /b, 即可得到三个结果文件--_di4.res、_di%.res 和_弟.res。

5. 今后的工作及展望

今后我们的工作主要有以下几个方面:

(1) 扩充词表, 并注音、校对: 注音准确、收词丰富的词表是字音转换工具高效工作的基础。原来我们采用的是经过整理和校对的 6 万词的词表。在实际使用过程中, 我们发现, 这个词表收词不够, 特别是没有详尽收录包含多音字的词语。我们准备尽快在原有 6 万词词表的基础上, 加入人名和地名词典, 并注上音, 形成一个规模约为 10 万词的词表。另外, 还要进行一次较细致的校对工作, 特别关注多音字在各个词语中的正确发音, 争取在资源一级上尽可能地杜绝错误。

(2) 规则的修改与语料抓取、转换和实验: 音字转换、语料抓取和对齐、字音转换评测工具的研制目的是为了自动高效地测试出智能型拼音—汉字自动句转换系统的准确率, 从而为规则、知识的修改提供实验依据。目前, 智能型拼音—汉字自动句转换系统已初步实现, 规则集基本建立, 应将重点转移到语料抓取、转换实验和规则修改的系统测试与完善工作上来。在系统测试与规则修改过程中, 系统支撑环境和开发工具也将得到不断的检验和改善。

(3) 系统支撑环境、开发工具和整个系统整合界面的研制: 本系统最终的性能目标之一是使用方便的用户界面。但由于系统的整体开发工作量巨大, 我们分成两步完成。一部分是本文论述的系统的支撑环境和开发工具, 另一部分是拼音—汉字转换系统, 因此有两个不同操作系统下的用户界面, 我们希望能尽快地完成这两部分界面的整合工作, 以方便系统的调试及以后的使用。

参考文献

- [1] 张普: “汉字编码键盘输入纵横谈(1—13)”, 中国青年报, 1995 年连载。
- [2] 许闻廉、陈克健: “‘国音’智慧型输入系统的语义分析‘脉络会意法’”, 《计算语言学研究与应用》, 陈力为主编, 北京语言学院出版社, 1993 年 10 月, 北京。
- [3] 曹剑芬: “谈谈语料库的语样选取问题”, 《计算语言学研究与应用》, 陈力为主编, 北京语言学院出版社, 1993 年 10 月, 北京。
- [4] 郭进: “统计语言模型及汉语音字转换的一些新结果”。
- [5] 孟洁、陈群秀: “汉语的一种知识一体化表示方法”, ISMP&CLIP 国际会议论文, 1999 年 6 月 26 日, 北京。