The Pitfalls and Complexities of Chinese to Chinese Conversion

汉字简繁转换的复杂性和陷阱

漢字簡繁轉換的複雜性和陷阱

春遍雀来（Jack Halpern）

日中韓辭典研究所 所长

華留萬陽貳（Jouni Kerman）

日中韓辭典刊行會软件开发总工程师

目录

（株）日中韓辭典研究所

The CJK Dictionary Institute, Inc.

〒352-0001　日本国埼玉県新座市東北2-34-14　小峰ビル 3･4F

电话：048-473-3508　　Fax：048-486-5032

E-mail：jack@cjki.org　　网址: http://www.kanji.org

0. 摘要

汉语有两种书面形式：中华人民共和国和新加坡使用的**简体中文**，和台湾、香港、澳门以及大多数海外华人使用的**繁体中文**。但是存在一种常见的误解,认为这两个体系之间具有直接的对应关系，相互转换只需要从一个字符集对应到另一个字符集就可以了，例如从国标码(GB2312-80)转换到大五码(Big5)。

虽然很多代码转换工具实现了这种转换，但事实却是截然相反的。这两种体系在不同级别上都存在重要的差异，不论是字符集，编码，拼写法(字的选择)，词汇(词的选择)，还是语义(词义)，都有着显著的差别。

随着东亚在世界经济里的地位日益重要,地方化公司和翻译公司都有着对中文简繁体转换的迫切需求，但也必须克服以下障碍：(1)现有的转换工具产生的结果不能令人满意；(2)缺乏发展好的转换工具所需的知识；(3)无法得到高质量数据的字典；(4)手工转换费用太高。

1996年，**日中韓辭典研究所**(The CJK Dictionary Institute, Inc.)开始深入调查这些问题，并建立了一个总括了中文简繁互转的数据库(**300**万条，且仍在发展中)，其目的是要使转换软件的准确性接近百分之百.

这篇论文解释了涉及的复杂问题，并展示这项基于Unicode的新技术将如何大大减少中文地方化和翻译项目的时间和费用。

1. 序

1.1 历史背景

汉字在它几千年的历史中经历了许多变迁。很多书法风格，异体字，和字体设计都有逐步的演变。有些完整的，复杂的字体被提升为“正字”，而那些令人眼花缭乱、泛滥成灾的变体则往往被降级为“俗字”。

在中华人民共和国于1949年成立后不久，新政权就发起了一场积极的运动，贯彻大规模的书面语改革。在五十年代，毛泽东和周恩来提出了简化汉字是一项应该优先完成的任务。1952年成立的语言改革委员会开始深入研究这一问题，并从事编纂简化字表的任务。

这些行动导致了许多书面语的改革，最重要的有：建立了一套标准化的罗马字系统(拼音)，限制日常用字的数量，以及大大地简化了数以千计的字形。一度，这项运动的目的是完全废除汉字，以罗马字母表代之，但后来还是倾向于使用简化字形而放弃了这项政策。

随后几年出版了几种简化字表，其中最著名的是1964年出版的“权威”**简化字总表**，之后又重新发行了几次并作了次要的修改。最新版本是1986年出版的，收录了2244个简体字[简体字总表 1986]。

台湾、香港和多数海外华人没有实行简化。尤其是台湾,还在严格地遵循着繁体的形式。台湾教育部出版了几种字符表，例如有4808个字的“常用國字標準字體表”，作为正确字形的标准。

1.2 简体与繁体中文

虽然简体与繁体中文的最大区别在于字形，我们将会看到两者之间还有字符集、编码方式和词汇选择方面的差异。

从实用角度来说，**简体中文**一词通常指满足以下条件的中文文本：

1. **字形**：简体中文必须是用简体的字形书写的(除非不存在简体的形式)。
2. **字符集**：简体中文通常使用国标码字符集，或其扩充版本，国家标准扩展码(GBK)。
3. **编码**：简体中文通常将国标码编为EUC-CN或用于互联网传送数据的HZ的文本。
4. **词汇用法**：词汇的选择采用中国大陆的用法。

与此类似，**繁体中文**一词一般指满足以下条件的中文文本：

1. **字形**：繁体中文必须是用繁体的字形书写的。
2. **字符集**：繁体中文通常使用大五码字符集。
3. **编码**：繁体中文通常编为大五码。
4. **词汇用法**：词汇的选择采用台湾或香港的用法。

以上只有第一条是必要条件。“简体”中文的定义决定它不能用繁体的字形书写，除非一个繁体字形不存在对应的简体形式。同样，“繁体”中文除了某些次要的例外情况(如某些专有名词)之外必须不能以简体字形书写。字符集和编码方式的限制要小一些，下面1.4节会讨论这一点。

词汇的用法上也有一些变化。例如台湾文本可能会包括某些中华人民共和国式的词汇，而新加坡的文本可能会采用台湾而不是大陆的计算机术语。尽管如此，总的来说简体中文和繁体中文两词的用法如上文所述。

1.3 问题本质

中华人民共和国的语言改革对书面汉语产生了重大影响。从处理中文数据的角度出发，最相关的问题有以下几个：

1. 许多字形经过了较大的简化，以至无法辨认它们的繁体形式。例如，繁体中文中的**徵**变为简体中文中的**征**。
2. 在很多情况下，一个简体字与多个繁体字对应(相反的情况较为少见)，例如简体中文的**征**与繁体中文的**徵**和**征**对应。根据上下文意思，通常只有一个是对的。
3. 有时一个简体字与多个繁体字对应，根据上下文意思，*每个*对应的繁体都可能是对的。
4. 简体中文使用的国标码标准与繁体中文使用的大五码标准互不相容，因此双方都产生了无数的漏字。

上述的第二条是中文简繁转换的关键问题，也是这篇文章的重点。在对此的讨论中采用的“经典”例子是繁体字**發**和**髮**。从词源学来看它们是两个不同的字，被合并成了一个简体字**发**。下表展示了这个以及其它一个简体字对应到多个繁体字的例子。

表1：简繁一对多的对应

简体源字	繁体标字	语义	繁体例子
发 fā	發	Emit	出發 start off
发 fà	髮	Hair	頭髮 hair

干 gān	乾	Dry	乾燥 dry
干 gàn	幹	Trunk	精幹 able, strong
干 gān	干	Intervene	干涉 interfere with
干 gàn	榦	tree trunk	楨榦 central figure
面 miàn	麵	Noodles	湯麵 noodle soup
面 miàn	面	Face	面具 mask
后 hòu	後	After	後天 day after tomorrow
后 hòu	后	Queen	王后 queen

如上所示，成功地把这些简体字转换为对应的繁体字取决于它们的上下文，尤其是它们所在的词。转换往往不能仅从一个码点对应到另一个码点，而是必须建立在更大的语言单位上，比如词。

除上表之外，数以百计的其它简体字也与多个繁体字对应，产生了语义不清的以一对多的对应，只有上下文能决定它们的关系。在这篇文章里，这些对应被称为**多字体**的对应，因为一个简体字——或*书写单位*——可能会与多个的繁体字对应，而相反情况也成立。

1.4 字符集和编码

这篇文章主旨不是对中文字符集和编码方法进行深入的讨论。小林剑(Ken Lunde)的重要著作 *CJKV Information Processing* 有对此的讨论。[Lunde 1999]这一节只简单地概括一些重要问题，因为我们的主要目的是论述更高一级的语言学问题。

简体中文通常使用国标码字符集，或其扩充版本国家标准扩展码，并通常被编为EUC-CN。在互联网上传送数据时，它常常被编为HZ，或是更早的zW。繁体中文通常被编为大五码，有时也被编为基于台湾国家标准(Chinese National Standard) CNS 11643-1992字符集上的EUC-TW。

在日本，有些文字处理系统通过JIS X 0208:1997字符集及其附加部分处理中文字符。同样，也可以把中文编为韩国的KS X 1001:1992字符集。但是，这两种情况都没有足够的简体或繁体中文字供日常中文之用。此外还有用来编辑中文的字符集CCCII(仍在使用的台湾早期标准)，可见情况的复杂程度。

从简繁码转换的角度出发，一个重要问题是国标码和大五码互不相容。前者包括了6763个字，而后者有13053个字。国标码大约三分之一的字是大五码里没有的简体字。这一点导致了双方的许多漏字现象，如下表所示。

表2：国标码和大五码的不相容性

汉字	国标码 (EUC)	大五码	Unicode
頭	*	C059	982D
發	*	B56F	767C
計	*	AD70	8A08

头	CDB7	*	5934
发	B7A2	*	53D1
计	BCC6	*	8BA1
干	B8C9	A47A	5E72
里	C0EF	A8BD	91CC

简繁互转中的困难并不仅限于国标码和大五码字符集。其实，大五码只包括了繁体字的一个子集。出乎意料的是，国标码也不包括某些简体字，如下表所示。

表3：国标码和大五码中没有的简繁对应

简体 Unicode	简体源字	繁体标字	繁体 Unicode
7EBB	纟	紆	7D35
8BEA	诔	譚	8B78
8D51	贠	贔	8D14
94D4	铚	錣	930F
9613	阌	鬮	95E0
98CF	飏	颺	98BA
9978	饸	餄	9904
9A89	骘	驫	9A6B
9C97	鲡	鯔	9C02
9E40	鸦	鴉	9D50

国际标准ISO-2022:1994[ISO 1994]试图建立一个紧缩字编码系统来处理这些不相容的问题，用逸出顺序机构表示字符集之间的转换，但这并没有完全解决这一问题。

国际标准字符集Unicode/ISO 10646解决了许多与简繁码互转有关的问题。[Unicode 1996]因为Unicode是这两种标准的超大集，在允许Unicode的系统里可以表现所有的大五码和国标码的码点，并在同一个文件中展示它们。这大大简化了简繁在码点一级的互转。尽管还有一些问题需要处理(例如现有版本排除了许多字[Meyer 1998])，Unicode有效地解决了大五码和国标码字符集之间不相容而导致的问题。

2. 转换的四级

自动把简体中文转换为繁体中文的过程(在一定程度上，从繁体中文到简体中文也是如此)潜在着许多复杂问题和常见错误。这个转换是从容易引起无数错误的一级码对转换开始，直到会参考语义和句法的四级语境转换，通过这从浅入深的四级转换方法进行处理，以期达到近乎完美的效果。下表描述了每一级。

表4：转换的四级

一级	**码对的**	字对字，*码*基础上的替换

二级	**字对的**	词对词，*词*基础上的转换
三级	**词对的**	词对词，*词汇*基础上的转换
四级	**语境的**	词对词，*语境*基础上的翻译

2.1 一级：码对转换

2.1.1 基本概念

最简单但也是最不可靠的简繁或繁简转换的方法是在码点对码点的基础上进行转换；就是说，在硬编码的，一对一的对应表里找到源点，然后用另一个字符集(例如大五码0xB0EA的繁体**國**)的标码点取代这个字符集的一个源码点(例如国标码(EUC)0xB9FA的简体**国**)，进行简单的替换。

这种转换可被描述为字对字，*码*基础上的替换，又称**码对**转换，因为参与转换过程的单位仅限于单个码点。也就是说，文本没有被分解为更高级的语言单位，而是作为互不相关的多字节字的一序列编码值被进行处理。

以下是一个一对一的编码对应例表。

表5：编码对应表

简体源字	国标码(EUC)	繁体标字	大五码	省略的候选项
出	B3F6	出	A558	齣
发	B7A2	發	B56F	髮
干	B8C9	幹	A47A	乾 干 榦
暗	B0B5	暗	B774	闇
里	C0EF	裡	B8CC	里 裏
征	D5F7	徵	BC78	征
门	C3C5	門	AAF9	
汤	CCC0	湯	B4F6	

由于这种表把每个源字只对应到一个标字，其它有可能的候选项就被忽略了(见“省略的候选项”一栏)，经常导致错误的转换。

例如，一个简体的字符串“头发”不是作为一个单位处理，而是被逐字转换。由于简体的**头**只与繁体的**頭**对应，转换是成功的。但是，由于简体的**发**与繁体的**髮**(用于头发)和繁体的**發**(用于发射)对应，转换就可能失败。就是说，一种经常出现的情况是，如果表把**发**对应到**發**，结果将是无意义的**頭發**：“头”+“发射”。另一方面，如果表把**发**对应到**髮**，**头发**会被正确地转换为**頭髮**，但其它的常见词汇，如简体的**出发**，会被转换为无意义的**出髮**：“出去”+“头发”。

如果一个复合词的每个语素都与多于一个的字对应的话(多字体的复合词)，这些问题就更加复杂了，因为这样排列的数目会以几何级数增长，如下表所示。

表6：简繁多字体的复合词

简体源字	词义	正确繁体	其它繁体候选项
特征	characteristic	特徵	特征
出发	start off	出發	出髮 齣髮 齣發
干燥	dry	乾燥	干燥 幹燥 榦燥
暗里	secretly	暗裡	暗里 闇里 闇裡 暗裏 闇裏
千里	long distance	千里	韆里 千裡 韆裡 千裏 韆裏
秋千	a swing	鞦韆	秋千 秋韆 鞦千

很明显，当存在几个候选项供挑选时，一对一的码对转换很有可能产生错误的结合。这表明在没有(显著的)人为干预时不能靠码对转换提供准确的结果。

2.1.2 转换过程

有三种不同的，越来越复杂的方式进行码对转换：

1. **简单化的转换**：指基于一对一的对应表的系统，在几个选项中选择标码点时没有充分考虑它的出现频率。简单化的转换经常产生不令人满意的结果，需要很大的人为编辑的努力。不幸的是，很多转换手段采取这种方法。它唯一的优点是使用简单，花费不多。

2. **基于频率的转换**：指建立在一对一对应表上的一种系统，其中标码点是几个选项中的*第一个*，从按出现频率排列的表中被选择出来。2.1.1节里的表5是一个基于频率的对应的例表。

   尽管这种方法经常产生正确的结果，在许多情况里，多标对应里的第二个(或第三个)选项本身也是高频率的，这时它就有可能失败。比如**发**这个例子，与繁体的**發**和**髮**都对应。

   我们调查了几个基于频率的系统，发现了很多错误和遗漏。建立一个基于频率的码对转换程序的最大困难是迄今为止还不存在建立在可靠统计上的准确全面的对应表，需要进行广泛的研究。附录C给出了一个知名转换程序里的错误对应的例子，并与日中韓辭典研究所发展扩充的对应表作了比较。

3. **基于候选项的转换**：指建立在一对多的对应表上的系统，候选项按出现频率排列。在一对多对应的情况下，用户得到一串候选项，或是直接出现在用户界面(UI)上，或是一个括号里的表。

几个声称支持繁体中文的主要中文电子字典和文字处理程序似乎是建立在简单化的方法上的。有些中文输入系统结合了(1)和(2)。第三种方法很少见，用于我们内部的码对转换程序之一。

概括地说，码对转换有以下缺点：

1. 如果使用简单化的转换，通常会产生不满意的结果。

2. 即使使用较复杂的转换(如上述的(2)和(3))，也还可能需要大量的人为干预，如需选择候选项和/或进行事后编辑。

3．彻底地忽略了词汇用法上的区别(详见下文)。

2.2 二级：字对转换

2.2.1 基本概念

简繁转换的更复杂的下一级可被描述为词到词，*词*基础上的转换。我们称之为**字对转换**，因为参与转换过程的单位是拼字单位：也就是说，在字典和对应表里被作为单个条目处理的字或有意义的字的结合。

在此文中，我们称之为**词的单位**。词的单位代表有意义的语言单位，例如单字词(自由形式)，象词缀这样的语素(黏附语素)，多字复合词(自由和黏附)，甚至更大的单位，比如成语词组。为简短起见，如果不会造成混乱的话，我们有时会使用*词*作为*词的单位*的同义词。

2.2.2 转换过程

字对转换通过四个步骤在词的单位的基础上进行。

1．把源句或词组分词为词的单位。

2．在拼字(词的单位)的对应表里查找词的单位。

3．产生标词的单位。

4．在需要的编码里输出标词的单位。

例如，简体词组**梳头发**先被分词为**梳**这个词的单位(单字自由语素)和**头发**(两字复合词)，把每个单位都在对应表里查找一遍，然后被转换为标字符串**梳頭髮**。要点在于**头发***没有*被分解，而是被作为单个词的单位处理。(实际上，这个例子由于**梳頭**也是一个正当的词的单位而更加复杂了。)

以下是一个拼字(词的单位)的对应例表。附录B给出了一个更详细的表。

表7：拼字对应表

简体词的单位	繁体词的单位	拼音	词义
头发	頭髮	tóufa	Hair
特征	特徵	tèzhēng	Characteristic
出发	出發	chūfā	Start off
干燥	乾燥	gānzào	Dry
暗里	暗裡	ànlǐ	Secretly
千里	千里	qiānlǐ	long distance
秋千	鞦韆	qiūqiān	a swing

值得注意的是，在码对转换和字对转换中，结果都必须和源有**拼字上的对应**。就是说，源和标都不过是同样的底层*词位*的拼字变体(见下2.3.1节)。这意味着每个源字都必须和标字一致，或是精确的一对一对应。

例如，在把简体的**计算机**转换为繁体的**計算機**时，**计**与**計**对应，**算**与**算**对应(同样的文字)，

机和**機**有一对一的对应。和词对转换(三级)不同的是，没有把简体的**计算机**“翻译”为繁体的**電腦**的企图。

2.3 三级：词对转换

2.3.1 基本概念

只要源词和标词有拼字上的对应，如简体的**头发**和繁体的**頭髮**，字对转换是有效的。然而，台湾，香港，和中华人民共和国有时在创造技术术语时采取了不同的途径。结果是在很多情况下简体和繁体对同一概念有完全不同的词。也许最有名的例子就是*计算机*了，在简体里通常叫做**计算机**，而在繁体里是**電脑**。

简繁互转更复杂的下一级是把这些不同之处考虑进去，从一个“翻译”出另一个，也可被形容为词到词的，*词汇*基础上的转换。我们称之为**词对转换**，因为参与转换过程的单位是语义单位，或*词位*。

一个**词位**是词汇的基本单位，例如单字词，词缀，或复合词。在这篇文章里，它也代表更大的单位，例如成语词组。为了实用的目的，它和字对转换里用的词的单位类似，但*词位*用在这里强调这个转换过程的语义上的本质。

在某种意义上，把一个词位转换为另一个和翻译两种语言有相似之处，但我们称之为词对转换而不是“翻译”，因为它局限于一门标准语言的几种互相有紧密关系的变体的词和词组，而且不象普通的双语翻译那样变动词的顺序。

2.3.2 转换过程

让我们用简体字符串**信息处理**作例子。它先被分词为词位**信息**和**处理**，在词位对应表里查找每个词位，然后转换为标字符串**資訊處理**。

值得注意的是，**信息**和**資訊**在拼字上是*不*对应的；就是说，他们本身是不同的词位，而不只是同一个词位的拼字变体。这和美式英语的“汽油”(gasoline)及英式英语的“汽油”(petrol)之间的差别是类似的。

另一方面，**处理**和**處理**之间的区别和美式英语的“颜色”(color)和英式英语的“颜色”(colour)相似，是同一个词位的拼字变体。一定不能太刻板地理解这个与英语的类比，因为英语和中文书面系统是根本不同的。

词对转换和字对转换有类似之处，但在两个方面有重要的区别：

1. 对应表必须把一个词位在语义一级上对应到另一个。比如，简体**计算机**必须被对应到它的繁体的词位的同义词**電腦**，*不是*它拼字的对应词**計算機**。

2. 分词的算法必须复杂到可以确认专有名词的地步，因为标字的选择有可能取决于某词位是否是专有名词(见下2.3.3节)。

下面是一个词位对应的例表。

表8：词位对应表

英语	简体词位	简体拼音	繁体词位	繁体拼音
Bit	位	wèi	位元	wèiyuán
Byte	字节	zìjié	位元組	wèiyuánzǔ

CD-ROM	光盘	guāngpán	光碟	guāngdié
Computer	计算机	Jìsuànjī	電腦	diànnǎo
Database	数据库	Shùjùkù	資料庫	zīliàokù
File	文件	Wénjiàn	檔案	dàng'àn
Information	信息	Xìnxī	資訊	zīxùn
Internet	因特网	Yīntèwǎng	網際網路	wǎngjì-wǎnglù
Software	软件	Ruǎnjiàn	軟體	ruǎntǐ
Week	星期	xīngqī	禮拜	lǐbaì

可以看到，上表把一种中文词位的语义的内容对应到另一种，在这方面与双语词汇的结构是一样的。

2.3.3 专有名词

词对转换的另一个方面是对专有名词的处理。专有名词简繁互换在分词过程和编纂对应表时都造成特殊的问题。一个主要的困难是许多非中文的(甚至一些中文的)专有名词在拼字上不对应。在这种情况下，码对转换程序和字对转换程序都会不可避免地产生错误的结果。

转换专有名词时的主要问题有：

1. **分词**：分词的算法必须复杂到可以确认专有名词的地步，因为标字的选择有可能取决于某词位是否是专有名词。

2. **非中文名字**：在有些非中文的专有名词里，简体和繁体中文用字不同。例如，简体的**肯尼迪**是“Kennedy”的音译，与繁体的**甘迺迪**对应。注意**肯**和**尼**与**甘**和**迺**不对应。

3. **二维对应**：有时一个源必须沿着二维对应到标：普通的词汇和专有名词。例如，简体**周**在一般词里对应到繁体的**周**或**週**(甚至**賙**)，但在人名中只对应到**周**。

下面是拼字上不对应的非中文名字的对应例表。

表9：非中文名字的词位对应表

英语	简体源	正确繁体	错误繁体
Berlin Wall	柏林墙	柏林圍牆	柏林牆
Chad	乍得	查德	乍得
Georgia	佐治亚	喬治亞	佐治亞
Kennedy	肯尼迪	甘迺迪	肯尼迪
Wisconsin	威士康星	威士康辛	威士康星

这种例子还有很多。这些区别不仅本身非常有趣，还有实际意义的后果。就是说，忽视它们的码对和字对转换程序会产生上面“错误繁体”一栏里列出的不令人满意的结果。

下面是如上(3)条里解释的二维对应的例子：

表10：二维对应

简体源	拼音	繁体(人名)	繁体(词)
周	zhōu	周	周 週 賙
发	fā	發	發 髮
才	cái	才	才 纔

这意味着简体的**发**作人名时必须总被转换为繁体的**發**，不可被转换为繁体的**髮**。这是相当困难的，因为分词程序必须复杂到可以区分作词用和作专有名词用的字。这是一个复杂的问题，本身就值得写一篇文章来论述。

2.4 语境转换

2.4.1 基本概念

简繁互转的最高级可以被形容为词到词，*语境*基础上的翻译。我们称此为**语境转换**，因为必须分析语义和句法的语境才能正确地把语义不清、一词多义的词位对应到多个标词位。

我们已经看到，字对转换程序和码对转换程序比起来的一大好处是它们处理词的单位，而不是单个码点。这样，简体的**特征**就被正确地转化为繁体的**特徵**(而不是错误的**特征**)。与此类似，词对转换程序处理词位。例如，简体**光盘**被转换为词位对应的繁体**光碟**，*不是*和它拼字相应但是错误的**光盤**。

在大多数情况下这是有效的，但有些特殊情况下一词多义的简体词位对应到多个繁体词位，取决于语境，*每一个*都有可能是对的。我们把这些称为**语义不清的多字体复合词**。

一词多义的简体复合词一对多的对应在拼字和词位级上都会出现。简体**文件**是个合适的例子。作“文件”一义时，它与自己对应，也就是繁体的**文件**；但作“数据档案”时，它与繁体的**檔案**对应。这种情况也可能发生在繁简转换的时候. 比如，繁体**資料**与简体**资料**作“材料，方法”时对应，但在作“数据”时和简体的**数据**对应。

2.4.2 转换过程

据我们所知，能自动转换语义不清的多字体复合词的转换程序还不存在。这需要类似于双语机器翻译使用的高级技术。这样的系统通常可以把文本流分解成词组，确认它们的句法功能，把词组分词为词位，确认它们的词类，并进行语义分析以确定使用语义不清的多字体复合词的特别意义。

日中韓辭典研究所现正在发展一个能部分解决这一难题的“伪语境的”转换系统。它不做句法和语义的分析，但通过一个允许用户起交互作用的半自动过程来达到高准确度。为了达到这一目标我们正在：

1. 为语义不清的多字体复合词建立一个一对多的数据库。
2. 发展一个用户界面，以使用户从候选项的表中手动选择。

以下是为拼字和词位级上语义不清的多字体复合词设立的对应例表。

表11：语义不清的多字体复合词

简体源	繁体选项1	繁体选项2

编制	編制 organize; establish	編製 make by knitting
制作	制作 creation (music etc.)	製作 manufacture
白干	白幹 do in vain	白干 strong liquor
阴干	陰乾 let pickles dry	陰干 even numbers
文件	檔案 (data) file	文件 document

2.4.3 最高级的转换程序

我们的最终目的是发展一个能达到近乎完美的转换准确性的语境转换程序。这样的转换程序至少要能做到以下几点：

1. 在句法和语义的基础上进行复杂的分段。
2. 确认专有名词和其它语态。
3. 包括全面的，建立在频率基础上的一对多的编码对应表。
4. 包括全面的拼字的和词位的一对多对应表。
5. 包括全面的二维的一对多的专有名词的对应表。
6. 自动转换多字体的词位，包括语义不清的多字体复合词。
7. 用批处理方式或与用户互动的方式操作。

下面的简体句无疑会使甚至最复杂的转换程序感到困惑：

> 发！请发这封传真可以吗？发点了点头发了传真。
>
> Hey, Fa! Could you please send this fax?
> Fa nodded his head and sent the fax.

今天最先进的转换程序最好也只能做到：

> 發！請發這封傳真可以嗎？發點了點頭髮了傳真。

说中文的人会感到好笑。转换程序把简体的独立词**头**和**发**和复合词**头发**混淆起来了。理想的语境转换程序应该能认出偶然相邻的独立词，并能产生正确的结果：

> 發！請發這封傳真可以嗎？發點了點頭發了傳真。

有讽刺意味的是，正*因为*一个简单化的码对转换程序无法辨识词的单位，它也许能在这个情况里给出正确的结果，但却是因为错误的原因！应该承认的是，这个例子很复杂。但是它是一个很自然的中文句子，清楚地证明了中文简繁转换的常见错误和复杂情况。

3. 讨论和分析

3.1 简繁转换的样本

下列是一个简繁词位(三级)的转换。

普通话简体字

根据《计算机周报》的报道，佐治亚软件研究所所长威廉肯尼迪氏和广东大学的信息处理研究所所长周东丰教授在香港举办了关于“因特网的现状”及“信息高速公路的未来”的发表会，并且对于明年两研究所将合并开发的因特网信息数据库进行了讨论。

臺灣的國語繁體字

根據《計算機週報》的報導，喬治亞軟體研究所所長威廉甘迺迪氏和廣東大學的資訊處理研究所所長周東豐教授在香港舉辦了關於“網際網路的現狀”及“資訊高速公路的未來”的發表會，並且對於明年兩研究所將合併開發的網際網路資訊資料庫進行了討論。

英文译文

According to the *Computer Weekly,* the director of the Georgia Software Research Institute William Kennedy, and the director of Canton University's Information Processing Institute Professor Dongfeng Zhou, held a press conference in Hong Kong on the topics “The Internet Today” and “The Future of the Information Superhighway.” They also discussed the plans of both institutes to build a “Database of Internet Information.”

上面一段是繁简词对转换的例子。它有几个有趣的特点，证明达到近乎完美的转换必须克服的主要挑战。下面我们来研究与前三级每级转换过程相关的问题。

3.2 码对转换问题

让我们先考虑一下如果用普通码对转换程序转换以上段落会出现什么情况。我们使用了某中国大学发展的很受欢迎的文字处理程序，得到了以下(很不令人满意的)结果：

根據《[計算機]{周報}》的[報道]，[佐治亜][軟件]研究所所長威廉[肯尼迪]氏和廣東大學的[信息]處理研究所所長周{東丰}教授在香港舉辦了{關于}“[因特網]的現狀”及“[信息]高速公路的未來”的發表會，{并且}{對于}明年兩研究所將{合并}開發的[因特網][信息][數據庫]進行了討論。

上面这段简短的文字包括六个在括号里的拼字错误，和11个方括号里出现的词位错误。105个字里有29个，即百分之28，被转换错了。它在转换所有词位时都出现了错误。现在我们先忽略词位错误(比如把**计算机**转换成**計算機**)。下表展示了拼字错误(“繁体结果”)，正确的繁体对应和其它的候选项。

表12：简繁转换结果

简体源	繁体结果	正确的繁体	正确	其它候选项
所长	所長	所長	是	
大学	大學	大學	是	
香港	香港	香港	是	
未来	未來	未來	是	
发表	發表	發表	是	發表 髮表 發錶 髮錶
东丰	東丰	東豐	否	東丰
周报	周報	週報	否	周報 賙報
并且	并且	並且	否	併且 并且
合并	合并	合併	否	合并 合並
关于	關于	關於	否	關于
对于	對于	對於	否	對于

只对应到一个繁体字的简体字组成的复合词只有一个繁体候选项，所以转换的准确率达到百分之百。有些包括多字体字的复合词，例如简体**发**(与繁体的**發**和**髮**对应)，有时被正确地转换过来，比如从**发表**到**發表**。但在其它情况下，例如简体**周**(与繁体**周**，**週**和**賙**对应)，它们经常不能被正确地转换，正如把**周报**转换为**周報**，还有在其它的五个例子里也是这样。

上述分析证明了码对转换是多么不可靠。

3.3 字对转换问题

没有正确地转换简体的**周报**，**并且**和其它词的问题可以通过使用二级字对转换解决。分词程序认出这些复合词是词的单位，在拼字对应表里查找它们，然后明确地把它们转化为正确的繁体对应。

下面是一个在拼字一级上把简体词的单位对应到繁体词的单位的例表。

表13：拼字对应

简体源	繁体标	拼音	英语
大学	大學	Dàxué	University
举办	举辦	Jǔbàn	Conduct, hold
所长	所長	Suǒzhǎng	Chief
处理	處理	Chǔlǐ	Processing
东丰	東豐	Dōngfēng	Donfgeng (a name)
周报	週報	Zhōubào	weekly publication
并且	並且	Bìngqiě	Moreover
合并	合併	Hébìng	Merge
关于	關於	Guānyú	about, concerning
对于	對於	Duìyú	Regarding

使用这种表保证了在词的单位一级上正确的转换，也避免了一对一码对转换程序内在的问题。

3.4 词对转换问题

我们已经看到，码对和字对转换程序不能处理简体**计算机**和繁体**電腦**这样的词位区别，因为同样的概念有不同的词位。还有许多非中文的专有名词在音译时用字不同。例如，简体的**佐治亚**，是“Georgia”的音译，应该对应到繁体的**喬治亞**，而*不是*它的拼字对应**佐治亞**。

如下表“正确”一栏所示，所有简体和繁体拼字不对应的词位和专有名词都没能被正确地转换。

表14：词位对应

英语	简体词位	简体拼音	繁体词位	繁体拼音	正确
Computer	计算机	Jìsuànjī	電腦	diànnǎo	否
Database	数据库	Shùjùkù	資料庫	zīliàokù	否
Georgia	佐治亚	Zuǒzhìyà	喬治亞	qiáozhìyà	否
Information	信息	xīnxī	資訊	zīxùn	否
Internet	因特网	yīntèwǎng	網際網路	wǎngjì-wǎnglù	否
Kennedy	肯尼迪	kěnnídí	甘迺迪	gānnǎidí	否
Report	报道	bàodào	報導	bàodǎo	否
Software	软件	ruǎnjiàn	軟體	ruǎntǐ	否

上述分析表明使用词位对应表对达到转换的高准确度是至关重要的。

3.5 繁简转换

一对多的对应问题并不局限于简繁转换。实际上，大多数简繁转换中遇到的困难在繁简转换中也存在。但是，拼字一级上一对多的对应在繁简转换中要少得多。

尽管如此，我们找到了数十个对应到两个简体字的繁体字，如下表所示。

表15：繁简一对多对应

繁体源	简体标	意义	简体例子
著 zhe	着	Particle	沿着
著 zhù	著	Writings	著作
乾 gān	干	Dry	干燥
乾 qián	乾	Male	乾坤
徵 zhēng	征	go on journey	长征
徵 zhǐ	徵	Ancient note	宫商角徵羽
於 yú	于	at, in	关于
於 yú	於	Yu (a surname)	於先生

有些字，例如繁体的**著**对应到简体的**著**和**着**，频繁出现在数以百计的复合词里，所以繁简转换不象开始看上去那么无足轻重。

值得指出的是，繁简对应不总是可逆的。比如，简体的**后**对应到繁体的**後**和繁体的**后**，而繁体的姓**後**只与简体的**後**对应。这意味着简繁对应表必须和繁简对应表分开保持。

3.6 问题到底有多严重？

问题的程度到底是怎样的？让我们看看统计数字。几个调查(例如[Xiandai 1986])证明最常用的2000个简体字占当代简体素材中出现的所有字的百分之97。其中，有238个简体字(几乎百分之12)是多字体的；就是说，它们与两个或多个繁体字对应。这个百分比是相当大的，也是简繁准确转换的主要困难之一。

在另一个方向的繁简转换，问题的程度要小得多，但我们发现，基于1亿7千万的繁体字素材(Huang 1994)上最常用的2000个大五码字中有20个与多个简体字对应。

但这些数字只表现了问题的一面，因为它们是建立在单字的基础上的。要正确地体会问题的严重性，我们必须研究所有包括多字体字的词的单位。

在我们现有的,简繁体各有100多万字条的全面的中文词汇数据库基础上[Halpern 1994, 1998]，据初步计算表明，大约97000个最常用的简体词的单位中有20000多有至少一个多字体的字，导致了一对多的简繁对应。这一比率达到了惊人的百分之21。类似的繁简对应的计算在大约87000个最常用的繁体词的单位中产生了3025个多字体的字，占全体的百分之3.5。这些数字证明仅仅从一个码点转换到另一个码点，尤其是简繁的方向，会导致不令人满意的结果。

由于许多高频率的多字体字是数以百计，甚至数以千计的复合词的组成部分，错误的转换会经常出现，除非一对多对应能(1)把字节串分词为语义上有意义的单位(词的单位或词位)，(2)分析语境以决定几个候选项中的正确选择，使意义明白无误。

4. 转换的新技术

4.1 项目概述

1996年，以日中韩计算辞书学[Halpern 1994, 1998]为专攻的**日中韓辭典研究所**,着手发展了一个中文简繁体转换系统.其最终目的是为了能得到近乎完美的转换结果。这是一项重大举措，需要投入大量的人力,物力。

为了达到这一目的，我们进行了以下研究和发展活动：

1. 深入研究所有和中文简繁转换有关的技术和语言问题。
2. 为前三级建立了简繁相互对应表。
3. 展了中文分词技术。

为了达到转换的高准确度，我们的对应表很全面，包括大约100万以上普通词汇的词位，技术术语，和专有名词。它们还包括一些其它特征，比如拼音读法，语法信息，语态，和语义分类编码。

4.2 系统组成部分

以下是对转换系统，尤其是我们的对应表的主要组成部分的概述：

1. **编码对应表**：我们的简繁编码互应表非常全面。它们不局限于国标码和大五码字符集，而是包括所有Unicode的码点。在一对多的情况下，候选项按频率排列，作为它的基础的数据是从一个庞大的1亿七千万字的素材以及我们繁体字专家组几年的研究中得出的。例见附录A。

2. **字对对应表**：为数以万计的多字体复合词建立准确的字对对应表需要很多手工劳动。我们的繁体字专家组检查和复查了每个字。例见附录B。

3. **词对对应表**：建立准确的词位对应表更加困难，因为简体和繁体字之间没有词对对应，而且(似乎)不存在显示简体繁体区别的词典。每个词都得单独检查，还要考虑到词义不清的多字体复合词带来的额外难题(见2.4.2节)。例见2.3.2节。

4. **专有名词对应表**：专有名词，特别是人名和地名，都经过了特殊处理。我们的中文和非中文的对应表现有约180万个专有名词。与词位表不同的是，这些表由于需要二维对应而特别的复杂。细节及例子见2.3.3节。

5. **转换引擎**：转换引擎的主要构成部分有：(1)复杂的**中文分词程序**，把文本流分词为词的单位并确认它们的语法功能；(2)**转换模块**，在对应表里查找词的单位并产生标的编码输出。

4.3 结论

中文简繁转换对地方化、翻译和出版业，及想要进入东亚市场的软件发展公司来说都变得日益重要。但是，我们看到问题是复杂的，建立对应表和发展分词技术需要很大努力。

日中韓辭典研究所占据了得天独厚的位置，向软件发展公司提供高品质的中文词汇资源和可靠的转换技术，消除了昂贵的手工劳动，显著地降低了费用。我们坚信，我们在这方面正在进行的研究和发展努力必将使我们接近建立完美的转换程序这一很难达到的目标。

鸣谢

对以下阅览了此文并提供了建设性批评建议的人士，在此我们表示衷心的感激。按字母表排列，包括：Glenn Adams, James Breen, Carl Hoffman, Timothy Huang, Ken Lunde, Dirk Meyer, 钱溯宁, Tsuguya Sasaki, David Westbrook, and Christian Wittern。评论组的几位成员都是中日韩信息处理领域的知名权威。感谢程似锦翻译本文。

同时特别向详细阅览了此文并提出了许多宝贵建议的 Glenn Adams 和 James Breen 致谢。

参考材料

[Halpern 1990] **Halpern, Jack** (1990): “New Japanese-English Character Dictionary: A Semantic

Approach to Kanji Lexicography" *Euralex '90 Proceedings.* Actas del IV Congreso Internacional, 157-166. Benalmádena (Málaga): Bibliograf.

[Halpern 1990] **Halpern, Jack** (1990): *New Japanese-English Character Dictionary* (Sixth Printing). Tokyo: Kenkyusha.

[Halpern 1994] **Halpern, Jack, Nomura Masaaki,** and **Fukada Atsushi** (1994): "Building a Comprehensive Chinese Character Database," *Euralex '94 Proceedings.* International Congress on Lexicography in Amsterdam.

[Halpern 1998] **Halpern, Jack** (1998): "Building A Comprehensive Database for the Compilation of Integrated Kanji Dictionaries and Tools," 43rd International Conference of Orientalists in Tokyo.

[Halpern 1999] **Halpern, Jack** (1999): *The Kodansha Kanji Learner's Dictionary*. Tokyo: Kodansha International.

[Huang 1994] **Huang, Shih Kun** (1994): *Chinese Usenet Postings.* Department of Computer Science and Information Engineering, National Chiao-Tung University, Taiwan (http://www.csie.nctu.edu.tw/).

[ISO 1994]: *ISO 2022:1994 Information Technology -- Character Code Structure and Techniques.*

[Lunde 1999] **Lunde, Ken** 1999: *CJKV Information Processing.* Sebastopol: O'Reilly & Associates.

[Meyer 1998] **Meyer, Dirk** (1998): "Dealing With Hong Kong Specific Characters," *Multilingual Computing & Technology,* Vol. 9 No. 3. Multilingual Computing, Inc.

[Unicode 1996]: *The Unicode Standard, Version 2.0.* Reading: Addison-Wesley.

[Xiandai 1986] 现代汉语频率词典 xiàndaì hànyu pínlǜ cídiǎn (1986). Beijing: Beijing Language Institute.

[Zongbiao 1986]: 国家语言文字工作委员会 (1986): 简化字总表 jiǎnhuàzì zǒngbiǎo (Second Edition): 语文出版社.

附录

附录A：码对转换对应表

表A-1:简繁编码对应表

国标码	简体源	繁体标	大五码
B0B5	暗	暗 闇	B774 EEEE
B2C5	才	才 纔	A47E C5D7
B3D4	吃	吃 喫	A659 B3F0
B5D6	抵	抵 牴 觝	A9E8 ACBB DBD3
B6AC	冬	冬 鼕	A556 C35D
B7E1	丰	豐 丰 風	C2D7 A4A5 ADB7
B8F6	个	個 箇	ADD3 BAE7
C0DB	累	累 纍	B2D6 F5EC
C3B9	霉	霉 黴	BE60 C5F0
CAAC	尸	屍 尸	ABCD A472
D5F7	征	徵 征	BC78 A9BA
DAD6	谥	謚 謚	EBAC EEB0
F3BD	蠼	蠼 蠷	F96E F8BE

表A-2：繁简编码对应表

大五码	繁体源	简体标	国标码(EUC)
AB5D	侷	局	BED6
ADB7	風	风 丰	B7E7 B7E1
B054	訊	讯	D1B6
B0A2	陝	陕	C9C2
B0AE	乾	干 乾	B8C9 C7AC
B16A	強	强	C7BF
B3CA	傘	伞	C9A1
B3F2	圍	围	CEA7
B6C4	傭	佣	D3B6
BAE0	箋	笺	BCE3
BBB1	跼	局	BED6
BC78	徵	征 徵	D5F7 E1E7
BECA	彊	强	C7BF
BFFD	錄	录	C2BC

附录B

表B：字对转换对应表

简体源	繁体标
暗杀	暗殺
暗码	暗碼
暗里	暗裡
暗昧	闇昧
幽暗	幽闇
霉菌	黴菌
霉雨	霉雨
霉菌	黴菌
特征	特徵
象征	象徵
秋征	鞦徵
长征	長征
出征	出征
累进	累進
系累	繫纍
丰姿	丰姿
丰韵	風韻

附录C

表C：某很受欢迎的转换程序的错误对应

国标码(EUC)	简体源	错误繁体	正确繁体
B7E1	丰	丰 豐	豐 丰 風
C3B4	么	么	麼 么
D4C6	云	云 雲	雲 云
CAB2	什	什	甚 什
B6AC	冬	冬	冬 鼕
BCB8	几	几 幾	幾 几
BACF	合	合	合 閤
BAF3	后	后	後 后
B8B4	复	復 複	復 複 覆 复
CAAC	尸	尸	屍 尸
B8C9	干	干 幹	幹 乾 干 榦
D5F7	征	征	徵 征
B5D6	抵	抵	抵 牴 觝
BDDC	杰	杰 傑	傑 杰
B4D6	粗	粗	粗 麤
B7B6	范	范	范 範
B8B2	覆	覆	覆
C3B9	霉	霉	霉 黴

作者介绍

JACK HALPERN 春遍雀來（ハルペン　ジャック）

株式会社 日中韓辭典研究所 所长
汉英字典刊行会 总编
日本昭和女子大学近代文化研究所 研究员

春遍雀來于1946年生于德国，在六个国家住过，通晓十二国语言。他在以色列基布兹居住时对汉字产生了浓厚的兴趣，在1973年到了日本，在十六年间编纂了**新汉英字典**[Halpern 1990]。他是职业词典编纂家、作家，经常作日本文化方面的演讲，在日语国际演讲比赛中夺得过头奖，还创建了国际独轮车协会。

春遍雀來目前兼任**汉英字典刊行会**(KDPS，一个专攻编纂汉字字典的非盈利组织）的总编和**日中韓辭典研究所**(CJKI)的所长，专门从事中日韩词典编纂业，并发展了全面的中日韩数据库(DESK)。他还编写了世界上第一个中日韩字符的Unicode字典。

下面是春遍雀来在中日韩词典编纂领域的主要著作。

Halpern, Jack (1982): “Linguistic Analysis of the Function of Kanji in Modern Japanese,” 27th International Conference of Orientalists in Tokyo.

Halpern, Jack (1985): “Function of Kanji in Modern Japanese,” *Transactions of the International Conference of Orientalists in Japan.* The Tō hō Gakkai (The Institute of Eastern Culture). 27th International Conference of Orientalists in Japan in Tokyo.

Halpern, Jack (1985): “Kenkyusha’s New Japanese-English Character Dictionary,” *Calico Journal*, December 1985.

Halpern, Jack (1987): 漢字の再発見 *Kanji no Saihakken* ‘Rediscovering Chinese Characters’. Tokyo: Shodensha.

Halpern, Jack (1990): *New Japanese-English Character Dictionary* (Sixth Printing). Tokyo: Kenkyusha.

Halpern, Jack (1990): “New Japanese-English Character Dictionary: A Semantic Approach to Kanji Lexicography,” *Euralex ’90 Proceedings.* Actas del IV Congreso Internacional, 157-166. Benalmádena (Málaga): Bibliograf.

Halpern, Jack (1993): *NTC’s New Japanese-English Character Dictionary.* Chicago: National Textbook Company.

Halpern, Jack, Nomura Masaaki, and **Fukada Atsushi** (1994): “Building a Comprehensive Chinese Character Database,” *Euralex ’94 Proceedings.* International Congress on Lexicography in Amsterdam.

Halpern, Jack (1995): *New Japanese-English Character Dictionary, Electronic Book Edition.* Tokyo: Nichigai Associates.

Halpern, Jack (1998): “Building A Comprehensive Database for the Compilation of Integrated Kanji Dictionaries and Tools,” 43rd International Conference of Orientalists in Tokyo.

Halpern, Jack (1999): *The Kodansha Kanji Learner’s Dictionary.* Tokyo: Kodansha

International.

Halpern, Jack and **Kerman, Jouni** (1999): “The Pitfalls and Complexities of Chinese to Chinese Conversion,” Fourteenth International Unicode Conference in Boston.

Halpern, Jack: *Dictionary of Unified CJK Characters -- for the Unicode Standard*. Forthcoming.

JOUNI KERMAN 華留萬陽貳（ケルマン　ヨゥニ）

日中韓辭典刊行會软件发展总工程师
昭和女子大学研究员

華留萬陽貳于1967年生于芬兰，从十几岁开始就对语言，计算机编程，和经济学产生了广泛的兴趣。除了他的母语芬兰语以外，他还学习过英语、瑞典语、法语、德语、意大利语、日语、普通话和粤语。1992年日本教育部授予他文部省奖学金以赞助他进一步进修日语。1996年他从赫尔辛基商业管理学院毕业并获得硕士学位。

華留萬陽貳在1996年获得昭和女子大学研究基金，加入汉字字典刊行会并为*讲谈社日英学习字典*发展一个页面组合系统[Halpern 1999]。项目结束后，他还参与了为日中韓辭典刊行會发展中日韩数据处理系统和设计数据库的课题。