一个支持人工校对的中文简繁体转换工具*

张小衡
香港理工大学 中文及双语学系，香港 九龙
E-mail: ctxzhang@polyu.edu.hk

摘　要：中文简繁体自动转换是当今社会的一大需求。由于计算机技术还远远不能保证译文的 100%准确，因而为了提供高质量译文人工校对是不可或缺的。本文报道了一个既能作中文简繁体自动转换，又支持人工校对的应用软件工具。该软件的自动转换正确率达到 99.80%，而且使得每一个能看懂繁体字中文的人都可以胜任校对工作。如果校对员精通繁简体字及其对应关系则工作效率会更高。软件设计的另一个特点是重视现行通用规范简体字和繁体字之间的转换和缩小两岸的文字差别。

关键词：繁体字；简体字；中文简繁体转换；人工校对

A Simplified-traditional Chinese Conversion Tool with a Supporting Environment for Human Proofreading

Zhang Xiaoheng
Dept. of Chinese & Bilingual Studies, Hong Kong Polytechnic University, Hong Kong
E-mail: ctxzhang@polyu.edu.hk

Abstract: Computer translation or conversion between simplified and traditional Chinese is widely employed in our society. However the present computer technology is unable to guarantee the correctness of its translation, which means human proofreading is often found indispensible. We have developed a software tool which looks after both requirements: It can translate a text from simplified to traditional Chinese at a correctness rate of 99.80%, and enables every traditional Chinese reader to effectively proofread the computer's output, while a knowledge of both simplified and traditional Chinese will further increase the efficiency. Another important feature of the software lies in its attention to the standardized popular simplified and traditional Chinese and to the unification of the two Chinese written systems.

Keywords: traditional Chinese characters; simplified Chinese characters; simplified-traditional Chinese conversion; proofreading

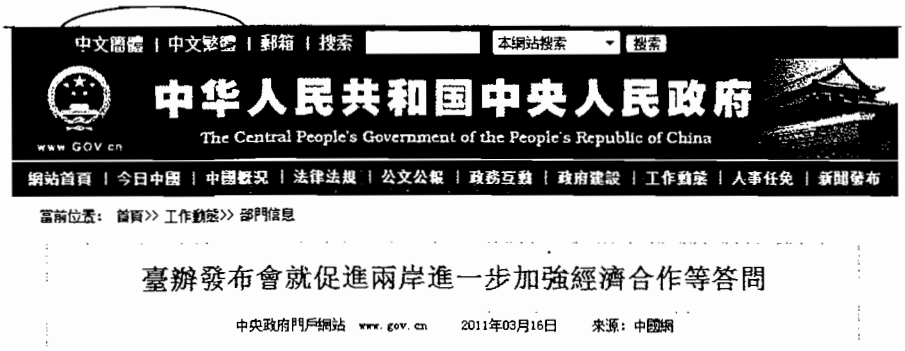
1　中文简繁体自动转换及其人工校对的重要性

由于历史原因(傅永和, 2005; 史定国, 2004)，现代汉语中同时使用着两个文字系统：中国内地的规范汉字是简体字，而港澳台地区则常以传统的繁体字为正体，国际上也广泛存在简体字和繁体字并用的情况。所以，许多中文文献都需要同时提供简体字和繁体字版本。例如中国的中央政府网站就这样做了，如图 1 所示。

不言而喻，简体字和繁体字中文之间的翻译已经成为社会的一大需求。计算机在这里无疑起到了举足轻重的作用。然而，同传统的语言机器翻译一样，简体字-繁体字中文自动翻译（也叫“转换”）不能保证译文的 100%准确。例如，在上图中央政府网站上关于国务院台湾事务办公室新闻发布会的报道中，简体字原文的“准”字共出现 7 次，其中就有 4 次被转换为错误的繁体字。例如将“批准”转为“批準”。根据最新版的《现代汉语词典》（中国社会科学院语言研究所词典编辑室，2005），“批准”的“准”在繁体字系统中仍然是“准”，与“標準”的“準”不同形。在台湾“中央研究院現代漢語語料庫”(繁体中文)中，“批准”出现 54 次，“批準”为 0 次。可见，将 “批准”转化为“批準”是错误的。上述繁体字网页可能是从简体字原文经计算机自动转换得来的。如果译文在上网发表之前经人工校对的话这种机器留下的“笔误”就不会出现在我

* 本研究受香港理工大学研究基金支助，课题编号: 1-ZV7U, A-PK14。

们中央政府的网站上了。



(http://big5.gov.cn/gate/big5/www.gov.cn/gzdt/2011-03/16/content_1825820.htm)

图 1　中国中央政府网站使用简体字和繁体字两个中文版本

我们做了一个实验，利用 MS Word 2007 将简体中文句子“他见王后步行十里路去发卡。”翻译为繁体形式。机器输出的结果是“他見王后步行十裏路去髮卡。”。根据原文的意思，MS Word 的繁体译文至少有两处错误：一是将表示长度单位的“里”误写成方位词“裏”，另一个错误是将动词“分發”的“發”错写成 “頭髮”的“髮”。此外还有一个可能的错误。这是由句子本身的歧义性所决定的。在没有上下文的情况下，简体句子“他见王后步行十里路去发卡。”有两种解释，一是“他见王(以)后，步行十里路去发卡。”，相应的繁体译文是“他見王後步行十里路去發卡。”；二是“他见王后(娘娘)步行十里路去发卡。”，相应的繁体译文是“他見王后步行十里路去發卡。”。只有原文的作者能说明是那种意思，其他人无法确定，而对于计算机来说简直是无从着手。

既然简体字-繁体字中文自动转换无法保证 100%准确，那么为了确保译文的质量人工校对是必不可少的。然而目前计算机还没有给人工校对工作提供专门的支持。

2　中文简繁体自动转换译文校对所面临的问题及其解决方法

汉字简繁体自动转换译文校对主要包括两方面的工作：1) 识别转换错误的字，2) 更正转换错误的字。下面我们从这两方面来讨论目前汉字简繁体自动转换译文校对所遇到的问题及其解决方法。

2.1　错转字识别方面的问题

据笔者了解，目前的简繁体汉字转换工具并没有在译文中留下任何有助于人工校对的标注。于是，为了确保全文正确，校对人员不得不从文章的开头到末尾逐字检查。这等于预先假设每个字都可能有错，无区别无重点，工作效率自然低下。

对于计算机简繁体中文转换来说，只有一简对多繁的情形才可能出错。而一简对多繁的简繁体字也只有一百来对，占 7000 现代汉语通用字的 2%以下。笔者曾在真实文本上作抽样统计（下文第 4 节将详细介绍），发现一简对多繁的字的动态使用频率也不到 4%。如果我们让计算机把这些可能转错的字标示出来，那么人工校对的文字范围就可以缩小到原来的 4%以下，这将大大提高译文校对的速度和质量。

在判断一个字是否转换正确时，目前社会上通用的计算机软件也未能为校对员提供帮助信息。这就要求校对员精通繁体字和简体字及其对应关系并能根据上下文作出正确的判断，否则就得频繁翻查相关的工具书。我们可以考虑让计算机提供必要的参考信息，方便校对人员在有需要的时候查看，以助判断一个字是否转换正确。

2.2 错转字更正方面的问题

要更正一个错别字，首先要找到正确的字，然后以后者取代前者。在确定正确的繁体字方面目前的简繁体中文转换软件也没有提供什么帮助，这意味着校对员为此又得频繁翻查相关的工具书。在这里我们也可以让计算机提供必要的参考信息。而且帮助判断错别字的信息和帮助更正错别字的信息可以结合在一起，进一步提高译文校对的效率。

发现错别字并找到正确的写法后，下一步是用正确的字去取代译文中的错别字。这一工作目前也是手工完成的，校对员需要先确定和删除错字，然后在原错别字的位置输入正确的字。这样做既麻烦费时又可能因不小心而产生新的笔误，例如删除不该删的字或输入另一个错别字或将字输入到一个错误的地方。此外还要求较对员具备计算机繁体字输入的能力。其实错别字的具体删和改，也可以让计算机去做。较对员只需要指点一下就可以了。至于如何在简繁体汉字转换软件中实现这些功能，我们将在下一节介绍。

3 系统设计与实现

严格来说，中文的简繁体转换应该涵盖从简到繁和从繁到简两个方向。我们现阶段主要研究从简到繁的转换。因为汉字中一简对多繁的情形比一繁对多简多得多。从简到繁的转换技术挑战性比较高，这方面处理好了，今后增加从繁到简的转换功能应该比较容易。所以本文主要讨论从简到繁的中文翻译。

3.1 建立简化字-繁体字对照字典

这个简化字-繁体字对照字典是系统的核心知识库，它不仅要支持简繁体中文的自动转换，还要为译文的人工校对提供有用的参考信息。同时还要重视现代中文繁简体地区的语言规范和习惯，支持大中华书同文(沈克成，2008)的发展，方便广大民众的语言交际。

汉字转换字典只需要收集简繁体异形的字和一简对多繁的字，因为其余的是一简对一繁的繁简同形的传承字，不需要转换。我们使用的资料来源主要有：

- 《简化字总表》(国家语委, 1997)
- 《香港、大陸、台灣 - 跨地區、跨年代現代漢語常用字頻率統計》(香港中文大学，http://humanum.arts.cuhk.edu.hk/Lexis/chifreq/)
- 《字頻總表》(台湾“教育部”, http://www.edu.tw/files/site_content/m0001/pin/biau1.htm?open)
- 《現代漢語平衡語料庫》(台湾“中央研究院”, http://db1x.sinica.edu.tw/kiwi/mkiwi/)
- 《两岸现代汉语常用词典》(北京语言大学，中华语文研习所合编, 2003)
- 《现代汉语词典》(中国社会科学院语言研究所词典编辑室，2005)

简化字-繁体字对照字典的制作包括以下几方面工作：

（1）首先是在 MS Excel 上建立一个包含整个《简化字总表》的简化字-繁体字对照表。为方便计算机处理，我们将一简对多繁的字条化解为多个简繁对应条目。例如 将“发(發，髮)”分解成“发(發)”和“发(髮)”。

（2）将台湾教育部《字頻總表》的每个繁体字与《简化字总表》中的每个简体字比较，如果同形则将该传承字加入字典。例如，《简化字总表》中原有 “后(後)”，通过上述处理添加了“后→(后)”。

（3）根据《字頻總表》和《香港、大陸、台灣 - 跨地區、跨年代現代漢語常用字頻率統計》给每一个字条中的繁体字标注三地的字频。然后将字典中的一简对多繁字条按照《字頻總表》的次序排列，以便选用高频字。我们采用《字頻總表》的次序，是因为该表收字较多，而且对于繁体字信息处理来说权威性较高。当然，也可以根据具体需要选用其他的字频。

（4）修正和优化内容。例如，在《简化字总表》中有简繁体字对应“么[麼]”。但是在香港中文大学的香港和台湾繁体字语料中都没有出现“麼”，而“麼”出现两千多次。台湾平衡语料库中也没有“麼”，但“麼”出现超过5000次。台湾的《国语标准字体母稿》也选用字型“麼”，不用“麼”。因此，根据现代繁体中文的实际应用情况和繁体字地区的官方标准，我们将简繁体对照字典中的“么[麼]”修改为“么[麼]”。又如，根据《两岸现代汉语常用词典》，在繁体中文中，可以用“台”代替“臺”。在平衡语料库中，“臺灣”出现2554次，“台灣”出现5000次以上。可见“台”比“臺”更常用。因此，我们删除了字条“台[臺]”，即将其合并到“台[台]”上。这样处理还有利于海峡两岸书同文的发展。

（5）为了方便繁体字译文的人工校对，我们给每一个一简对多繁的繁体字标注音义和例词等有助繁体字选用的信息。例如，简体字“后”对应繁体字“後”和“后”，当作方位詞用或指後代的人时应该转换为繁体字“後”，例如“前後, 先後”和“無後”；而当其意思是君主的妻子或古代的君主时，对应繁体字“后”，例如“皇后”。又如“发”，读 fa4 时，意思是毛发，对应繁体字“髮”，例如“頭髮,毛髮”；读 fa1 时，意思不是毛发，对应繁体字“發”，例如“發展, 發生, 發現, 發表, 開發,…”。让校对人员在转换软件上直接参考这些内容，可节省大量翻查工具书的时间。经过上述处理后，对照字典中共有2,291个简繁体字对应条目，其中一简对一繁(繁简异形)的有2,147条，一简对多繁的有144条。表1是一小段样本。

表1　简繁体字对照字典示意样本

Jianti	Fanti	Freq_TWE	Freq_ML	Freq_TW	Freq_HK	Sound_Sense	Examples
万	萬	513	326	312	276	wan4 (1)十千, 很多; (2)姓氏	一萬; 萬先生
万	万		4203		4050	mo4 用於姓氏“万俟”	万俟 (Mo4qi2)
丑	醜	2194	1997	1833	1671	難看, 叫人厭惡的	醜陋, 出醜
丑	丑	2848	3049	2986	3291	(1)地支的第二位; (2)小花臉	丑時, 小丑
丰	豐	951	738	1096	583	豐富, 大	豐收, 豐功偉績
丰	丰	4069		3092	4053	美好的	丰采,丰姿

字条内容包括简体字形、繁体字形、台湾教育部《字頻總表》序号、香港中文大学的大陆、台灣、香港字频序号、繁体字的读音和意思、词语举例等。这些内容既可用于计算机简繁体字的自动转换和一简对多繁的高频先见处理，又可帮助译文校对员发现和更正错别字。

词典建设是一项艰巨而重要的工作。由于篇幅所限，本文只能作简单的介绍，详细情况将另文讨论。

3.2　系统设计和工作原理

在对照字典的基础上我们设计实现了一个支持人工校对的从简体字到繁体字的中文转换工具j2f (Jian to Fan)。为方便使用，我们将j2f设计成一个WWW网站，用XHTML和JavaScript等技术编写。j2f的用户界面如图2所示。

软件操作十分简便。用户只要把简体字原文复制到上方的“A 简体字原文”文字区之中，然后点击从简体字区指向“B 繁體字譯文”区的按钮，j2f就会开始转换翻译，并将译文写入B区。如果只需要翻译原文区中的一部分文字，则需要在点击翻译按钮之前选择好这部分内容。

j2f的工作原理很简单。首先看“简体字原文”文字区中选择了部分行文没有，如果有则将待转换(翻译)的原文设置为该部分文字，否则设置为整个简体字区的内容。接着取待转换的简体字原文的第一个字符。如果该字符没有在对照字典中出现，则一定是一个传承字或其他不需要转换的

字符；如果该字符在对照字典中出现，则还要检查该简体字是否对应多个繁体字。如果只对应一个繁体字，则将该简体字转换为对应的繁体字就可以了。如果对应多个繁体字，则将该简体字转换为与其相对应的几个繁体字中使用频率最高的那个。一般来说，这个字最有可能是正确的选择。但是在一定的上下文中，正确的转换对象也可能是其他的对应繁体字。因此，为了方便人工译文校对，我们将初步选用的那个使用频率较高的繁体字设置成网页上的一个特殊连接，当用户点击该字时，计算机将根据对照字典的内容显示与原简体字对应的所有繁体字及其各自的选用条件，供校对人员参考。当校对员点选一个不同的繁体字选项时，计算机就会自动地将原来(机器)选用的繁体字更正为用户新指定的繁体字。处理完一个字符后，计算机会取简体字原文中的下一个字符根据上述方法加以处理，以此类推，直至全文转换完毕。最后输出译文。

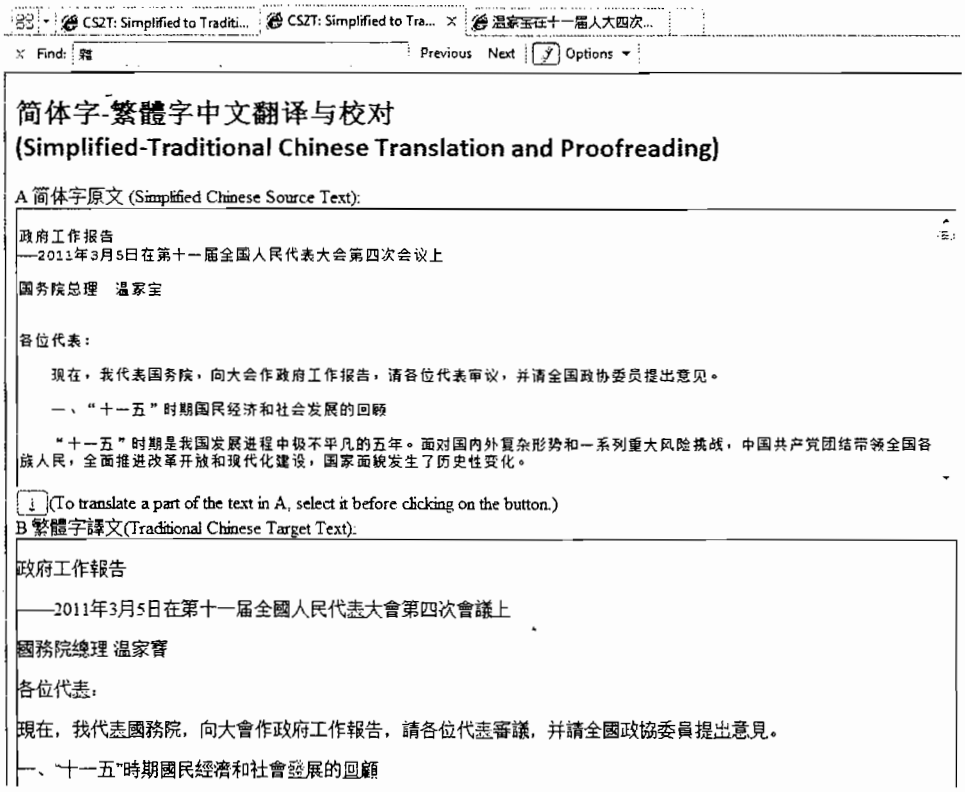


图2 j2f的用户界面

4 试验结果

为了测试j2f的性能，我们利用该软件尝试处理了一些中文文本。其中最有代表性的是今年的国务院总理《政府工作报告》（温家宝，2011）。因为这是一个举世瞩目的文献，内容覆盖面广，时代性强，而且语言规范。下文介绍政府工作报告的简繁体中文转换和校对情况。

4.1 操作方法

我们先从中国中央政府网址（http://www.gov.cn/2011lh/content_1825233.htm）把《政府工作报告》的简体字原文全文复制到j2f首页上的“简体字原文”区中，然后点击从简体字原文区指向繁體字譯文区的箭头按钮，j2f就会开始简繁体中文翻译，并将译文写入繁體字譯文区。翻译完毕后，用户界面的状态如图 2 所示。这时，用户就可以开始校对位于屏幕下方的繁体字计算机译文了。译文

中以链接形式出现的字是一简对多繁的高频繁体字，例如，“表”，“出”，“發”等。在简繁转换的层面上，只有这些字需要校对。当校对员发现某个一简对多繁的繁体字译文有误或需要进一步检查其正确性时，只要点击该字的连接，就会看到每种相关的简繁字对应及其选用条件，如图3所示。

图3　点击一简对多繁的繁体字译文，就会看到每种简繁对应及其选用条件

如果需要更正原来的选用字，只要点击正确的简繁对应选项，计算机就会自动将原来的选用字替换为新指定的字，如图4所示。用户不需要手工删除原来的错别字和输入正确的字，也不需要翻查其他的工具书。

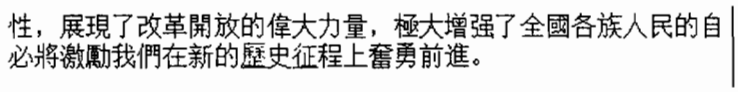


图4　计算机按用户指示自动将译文中的“徵”字更正为“征”

4.2　试验结果

我们用上述方法在j2f上对《政府工作报告》全文作了从简体字到繁体字的自动转换，并完成人工校对。试验结果和有关统计简介如下。

《政府工作报告》简体字原文共有21,018个字符(包括汉字，标点符号和空格等)，其中6,339个简体字在对照字典中出现，包括一简对一繁简繁异形的字和一简对多繁的字。一简对多繁的字共有823个，占全文字符总数21,018的3.92%。只有这一小部分字的繁体字译文需要校对。由于这些字在译文中都以链接形式出现，与其他字符颜色不同，还加了下划线，所以很容易辨识。

在人工校对的过程中，我们共发现和更正了42个计算机转错的字。也就是说，计算机转换一简对多繁的字的错误率是(42/823 * 100 %) = 5.10%。相对于整篇《政府工作报告》的字符数来说，自动简繁中文转换的错误率是 (42/21018*100%) = 0.200%，正确率为 1- 0.200% = 99.80%。

5　结论与讨论

计算机自动简繁体中文转换的正确率早已经达到99%以上（沈达阳，孙茂松，1996；辛春生，孙玉芳，2000），但是一直无法做到100%正确。在可见的将来也很难做到(Halpern, J, and Kerman, J, 1999). 因此，为了保证译文质量，人工校对必不可少。本文报道了一个既能作中文简繁体转换又支持人工校对的软件工具。自动转换的正确率达到99.8%，而且允许校对人员只要能看懂繁体译文就可以胜任工作。如果校对员精通繁简体字及其对应关系则工作效率会更高。值得注意的是，99.8%的正确率是在字处理层次上赢得的。在此良好基础上我们将引入词处理和上下文语言分析技术，可望把这一数字提高到99.9%以上。

我们的软件还有广泛的空间可进一步发展。首先是词典内容的优化，在此基础上，除了进一步提高从简体到繁体自动转换的正确率和改善对人工译文校对的支持之外，还要增加繁体字转简体字的功能。除了译文校对之外，还要支持对原文分析的校对，使得任何一个能读懂简体中文或繁体中文的人都能胜任校对工作。例如，假如有一个简体中文的作者想要将自己的文章翻译为繁体中文。即使他不懂繁体字，我们也可以让计算机帮助他完成这一工作。方法是，用计算机给简体字文本中的每个一简对多繁的简体字标注一个最可能正确的繁体字，例如“...必将激励我们在新的历(歷)史征(徵)程上奋勇前进。”其中“历”和“征”对应多个繁体字，需要人工校对。点击“历

(歷)”就可以看到用简体字书写的有关简繁对应及其选用条件，例如“(1) 历→歷: , e.g.历史,经历; (2) 历→曆: , e.g.日历, 农历”。这些信息可以帮助简体字用户验证到简体字“历史”的“历”确实对应繁体字“歷”，计算机的繁体字标注“历(歷)”是正确的。点击“征(徵)”也可以看到有用的信息，例如“(1) 征→徵: , e.g.征税,征求;特征,…; (2) 征→征: (军队)远行;征伐, e.g.征途,征讨”。这些信息会帮助用户发现“征程”的“征”在繁体中文中也写成“征”，应该把原来计算机的繁体字标注“征(徵)”更正为“征(征)”。通过计算机辅助校对将每一个对应多个繁体字的简体字都标注上与上下文一致的正确的繁体字之后，为整篇简体字原文自动生成100%正确的繁体字译文就是易如反掌的事了。当然，这还只是初步设想，有待深入探索。

还有一点要说明，从方便现代社会语言沟通这一目的出发，我们更关心的应该是中国内地和台港澳通用规范字之间的转换。而不是纯粹的笔画多的繁体字和笔画少的简体字之间的转换，或现代汉字和古代汉字的转换。例如台湾的“台”有一种繁体字写法“臺”。但是在台灣中央研究院的《現代漢語平衡語料庫》中查得：“臺灣”使用 2554 次；“台灣”出现 5000 次以上（免费使用该语料库最多只能显示 5000 条）。可见“台灣”是更为通用的写法，而且在繁体译文中选用“台灣”还有利于缩小两岸的文字差别，发展统一大业。

我们还没有看到有关繁简体中文转换译文校对辅助软件的报道。即使有这样的专门软件，上文介绍的翻译-校对二合一方法依然有优势。因为对于自动翻译的结果情形（何处有把握，何处可能有错）翻译软件本身“最了解”，所以与校对者的合作最有基础。还可以通过互动学习，让机器翻译从人工校对中得益。这种方法还可以应用于拼音标注和方言机器翻译(张小衡，1999)等其他许多需要人工校对的语言信息处理，发展前途相当广泛。

参 考 文 献

[1] 北京语言大学, 中华语文研习所 (合编, 2003). 两岸现代汉语常用词典. 北京: 北京语言大学出版社.

[2] 傅永和 (2005). 汉字简化五十年回顾. 中国语文, 2005 年 06 期.

[3] 国家语委 (1997). 简化字总表. In 语文出版社(编), 语言文字规范手册. 北京：语文出版社.

[4] 沈达阳，孙茂松 (1996). 汉字简繁体智能化转换系统. 中文信息, 1996 年第 6 期.

[5] 沈克成 (2008). 书同文: 现代汉字论稿. 上海:上海锦绣文章出版社.

[6] 史定国 (主编, 2004). 简化字研究. 北京: 商务印书馆.

[7] 温家宝 (2011). 政府工作报告——2011 年 3 月 5 日在第十一届全国人民代表大会第四次会议上. 北京: http://www.gov.cn/2011lh/content_1825233.htm.

[8] 辛春生，孙玉芳 (2000). 简繁汉字转换系统的设计与实现. 软件学报. 2000 年第 11 期.

[9] 张小衡(1999). 粤-普机器翻译中的词处理. 中文信息学报), No. 3, Vol. 13 (1999), pp. 40-47.

[10] 中国社会科学院语言研究所词典编辑室 (2005). 现代汉语词典. 北京: 商务印书馆.

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