中文音译规范化的自动实现

——以威妥玛-翟里斯式拼音转写为例

张霄军

（英国斯特林大学艺术与人文学院）

摘　要：为了顺应汉语规范化的基本国策和汉语国际推广的实际需求，不仅要在现在和将来的翻译工作中将人名、地名等中文翻译成成规范的汉语拼音，还要将以前用旧方案翻译的相应名称转写成汉语拼音。本文根据威妥玛式拼写法与汉语拼音拼写法的对应关系，总结了自动转写的转写规则，开发了基于规则的威妥玛式拼音自动转写为汉语拼音的软件 WG2PY，并自动转写了林语堂翻译的《浮生六记》中的译音字，取得了很高的正确率，实验数据也具有一定的覆盖率。

关键词：中文音译，《汉语拼音方案》，自动转写，威妥玛-翟里斯式，WG2PY

Rewrite WG Names to PY Names Automatically

Zhang Xiaojun

(University of Stirling, UK)

Abstract: It needs to translate the present Chinese personal names, place names and other proper names into other languages in *Pinyin* style, and also, to rewrite names translated in other styles in the past to *Pinyin* in order to temporize the national language plan and meet the requirement of broadcast Chinese. This paper summaries the rewriting rules according to the respective relationship between WG and PY, designs a rule-based software named WG2PY to rewrite the WG names in novel Six Chapters of A Floating Life into PY ones.

Keywords: Chinese transliteration，*Chinese Pinyin Plan*，auto-rewriting，WG，WG2PY

1 背景介绍

在 1982 年国际标准化组织（ISO）决定采用《汉语拼音方案》作为国际标准的汉语罗马字母拼写法之前，在国内外的翻译、通讯、交通等诸领域，汉语人名英译采用的旧的汉语罗马字母拼写法很不统一，比较流行的有威妥玛式、国语罗马字、拉丁化新文字等[1]。为了顺应汉语规范化的基本国策和专名音译的大趋势，不仅要在现在和将来的汉-英翻译工作中将汉语人名翻译成成规范的汉语拼音，还要将以前用旧方案翻译的人名转写成汉语拼音。但人工转写往往费时费力且很容易出错，转写的差错造成翻译质量低下的例子屡见不鲜。如我国某著名外语出版社在将杨宪益和戴乃迭合译的《红楼梦》英文版中原用威妥玛式拼音拼写的人名转写成汉语拼音重版时，由于采用手工转写，竟然出现了“乱点鸳鸯谱”的英文译文，闹出了不少笑话[2]。

此外，由于历史原因，香港和澳门地区的新语文政策也刚推行不久，本地人对《汉语拼音方案》的认同还不一致，中文音译时并没有完全按照《汉语拼音方案》。台湾地区由于众中所周知的原因，迟

收稿日期：2016-02-26　修回日期：2016-07-23

作者简介：张霄军（1978—），男，博士，英国斯特林大学(University of Stirling)讲师，博士生导师。研究方向为翻译技术、计算语言学。通信方式：xiaojun.zhang@stir.ac.uk。

迟未推行汉语拼音方案，甚至故意在中文音译上大做文章，别有用心[3]，那中文音译的不一致现象和混乱程度也就可想而知了。

然而，中文译音的汉语拼音化的基本国策是既定的，这个趋势谁也逆转不了。事实上，1958 年，我国第一届全国人民代表大会通过《汉语拼音方案》后不久，国际标准化组织（ISO）就已经决定首先在地名音译上采用《汉语拼音方案》[4]。之后，1979 年 6 月 15 日联合国秘书处发出通知，以“汉语拼音”的拼法作为各种拉丁字母文字中转写中国人名和地名的国际标准[5]，到了 1982 年，国际标准化组织决定采用《汉语拼音方案》作为汉语罗马字母拼写法的国际标准。然而，由于各种原因，世界各国对此反应不一。法国等欧洲国家率先响应，各国的出版物和图书馆很快就采用了汉语拼音。但是美国迟迟没有采用，直到 1998 年，美国国会图书馆才决定改用拼音，并准备以三年时间，花费几千万美元，把馆藏 70 万部中文图书的目录全部改成拼音[6]。近年来，随着汉语国际推广力度的增加、各国孔子学院的设立以及语言信息标准化的需要，汉语拼音的作用越来越明显。因此，对于历史所遗留的非汉语拼音式的中文音译罗马字母拼写式，都存在一个转写的问题。然而，诚如前面所述，人工转写往往费时费力且很容易出错。

本文根据威妥玛-翟里斯式拼写法（WG，见下文）与汉语拼音拼写法的对应关系，总结了自动转写的转写规则，开发了基于规则的威妥玛式拼音自动转写为汉语拼音的软件 WG2PY，并自动转写了林语堂翻译的《浮生六记》（*Six Chapters of A Floating Life*）[7]中的译音字，取得了很高的正确率，实验数据也具有一定的覆盖率。

2 自动转写实现

2.1 定义

拼音字——一组以某种汉字标音体系中的音素为标准而组成的音素序列（音节）。如 hsin 为威妥玛-翟里斯标音体系（WG，见下文）中的一个拼音字；xin 为汉语拼音标音体系（PY，见下文）中的一个拼音字。

拼音词——一组以某种汉字标音体系中的音节为标准而组成的词级音素序列。如 ch’ünfangp’u 为威妥玛-翟里斯标音体系（WG，见下文）中的一个拼音词；qunfangpu（“群芳谱”）为汉语拼音标音体系（PY，见下文）中的一个拼音词。

WG——威妥玛-翟里斯汉字标音体系。指以英国人威妥玛（T. F. Wade）所创建的汉字标音体系为基础后经翟里斯（Giles）修订的汉字标音体系，用该体系标音的拼音字和拼音词常见于 1979 年前出版的各种典籍英译著作或者未采用汉语拼音方案的外国人所写的英文著述，用以音译中国人名、地名、机构名等中的汉字。

PY——汉语拼音汉字标音体系。指以汉语拼音方案为蓝本的汉字标音体系，用该体系标音的拼音字和拼音词常见于 1979 年后出版的各种典籍英译著作或者已采用汉语拼音方案的外国人所写的英文著述，用以音译中国人名、地名、机构名等中的汉字。

同音字——WG 中不同音节对应相同 PY 中某一音节的拼音字互称“同音字”，如 WG 中 che 对应于 PY 中的 zhe，WG 中的 cheh 也对应于 PY 中的 zhe，则 che 和 cheh 互称同音字。

2.2 自动转写软件 WG2PY

功能：将威妥玛-翟里斯汉字标音体系的拼音字自动转写为汉语拼音汉字标音体系的拼音字。

处理模块及算法：程序由切字处理、转写处理和后处理（输出）三个模块构成。

切字处理模块：从给定的拼音词中根据“WG2PY 拼音字切字底表” 采用正向最大匹配法切分出

正确的拼音字（包括连字符处理，所有格撇号处理等）。

（1） 输入一个待转换的 WG 串 S1，和已转换的 PY 串 S2.

（2） 如果 S1 为空串，转 6；

（3） 从 S1 的左边复制一个子串 W 作为候选词，W 尽可能长，但长度不超过 Max_WG（我们设定为 10）；

（4） 如果在切字底表中找到 W，则将 W 转换为相应的拼音串，并将其加到 S2 的右边。并且从 S1 的左边去掉 W，转 2

（5） 去掉 W 中最后一个 WG 串，转（4）

（6） 结束。

转写处理模块：根据基础规则、转写规则和补充规则对切好的拼音字进行转写，基础规则见“WG2PY 基础规则库”，转写规则见“WG2PY 转写规则库”，补充规则见“WG2PY 补充规则库”。

后处理（输出）模块：包括连字符、大小写、撇号的处理等。

（1）大小写的转写规则：如果一个输入的 WG 串的首字母是小字母，那么它的转换后的 PY 串的首字母也转换为小写字母。反之亦然。

（2）连字符处理规则：如果一具 WG 串中是连字符“-”加上元音字母如（a,e,o），那么转写成 PY 串则转写为撇号“‘”加上元音字母如（a,e,o），其它情况直接去掉连字符“-”，在 PY 串中不保留任何痕迹。

（3）所有格撇号处理规则：只需要把“’s”放到切字底表中，采用正向最大匹配法，即可做正确切分。

3 实验内容与结果

3.1 WG2PY 拼音字切字底表

“WG2PY 拼音字切字底表”是采用正向最大匹配法切字的基础。根据 WG 与 PY 的声、韵母对应关系，我们参照威妥玛《语言自迩集》[8]中的“音节总表（*Sound Table*）”和“北京话音节表（*The Peking Syllabary*）”拼出了 420 个基本拼音字，后根据 WG 的“轻音”规则（见下文）拼出了 277 个同音字，这 420 个拼音字和 277 个同音字共同构成了“WG2PY 拼音字切字底表”。

3.2 WG2PY 规则

WG2PY 规则由基础规则、转写规则和补充规则组成，分别建立“WG2PY 基础规则库”、“WG2PY 转写规则库”和“WG2PY 补充规则库”。

“WG2PY 基础规则库”中共有规则 49 条，分为“Consonants”（21 条）、“Basic Vowels”（8 条）、“Basic Retroflex Syllables”（4 条）、“Basic Sibilant Syllables”（4 条）、“Semi-vowel Initials”（3 条）和“Basic Finals”（9 条）。基本对应于《汉语拼音方案》中所有的声母表和韵母表。

“WG2PY 转写规则库”中共有规则 420 条，对应于“WG2PY 拼音字切字底表”中的 420 个基本拼音字（音节）。

“WG2PY 补充规则库”中目前现有规则 109 条，分为“轻音”和“固化”两部分：轻音是指 WG 中 ng 和 h 的发音规则，其中和 ng 相关的规则有 10 条，和 h 相关的规则有 59 条；固化是指不符合上述所有转写规则但又在外文音译中已经固定下来的中文译音，如 Peking->Peking（“北京”），chow->zhou（“州”，用于地名），king->jing（“京”，用于地名），kiang->jiang（“江”，用于地名）等，目前已收录 40 条规则。

3.3 测试语料

这里的语料不是指原文全文或整句，而只是 WG 拼音词（字）和 PY 拼音词（字）。这些拼音词（字）的获取可以从 1979 年以前出版的各种典籍英译著作或者未采用汉语拼音方案的外国人所写的英文著述中获得。本实验中我们选用清朝人沈复所著、林语堂先生英译的小说《浮生六记》，我们从中获取 WG 拼音词 322 条，涉及拼音字 235 个（字型而非字例）。

3.4 实验结果

我们以转写正确率作为实验结果的评价指标，正确率计算公式如下：

正确率=转写正确的拼音字字数/全部测试拼音字字数×100%

经 WG2PY 转写后生成的 235 个 PY 拼音字中，转写正确的拼音字字数为 207 个，全部测试拼音字字数为 235 个，转写正确率为 88.09%。

同时，我们以覆盖率作为实验内容的有效性评价指标，覆盖率计算公式如下：

覆盖率=全部测试拼音字字数/拼音字表中的拼音字总数

“拼音字表中的拼音字总数”是指“WG2PY 拼音字切字底表”中的拼音字个数，为 420+277 个，因此覆盖率为 33.72%。

4 实验结果分析

1．实验结果中覆盖率偏低。这是由于在我们目前的科研条件和科研环境下，1979 年以前出版的各种典籍英译著作或者未采用汉语拼音方案的外国人所写的英文著述较难获取，因此从中提取测试拼音字的难度较大。

2．转写错误分析。实验中转写错误的拼音字共有 28 例，其错误原因可以分为以下三类：

第一类，撇号处理。WG2PY 的转写过程中的撇号出现有三种情况——（1）WG 中的送气符，如 Ch’ao，这种送气符会出现在声母 p, k, t, ch, ts 和 tz 与跟在它们后面的元音之间，即 p’, k’, t’, ch’, ts’和 tz’；在 PY 中无送气符号。（2）PY 中的隔音符，如 Xi’an，这种隔音符会出现在汉语拼音 a, o 和 e 开头的音节连接在其它音节后面从而使音节的界限发生了混淆的时候；在 WG 中的隔音符是以连字符-的形式出现的，如 yü-an。转写时 Ch’ao 要转写成 Zhao，yü-an 要转写成 yu’an。但实际文本中送气符的使用较为混乱，实际语料中经常会出现该用送气符的时候没有用，而不该使用送气符的时候却用了的情况*。撇号出现的第三种情况比较棘手，即（3）拼音字后接名词所有格的符号时，如 Wang Hsüchou’s。理论上讲，任何音节后面都可以跟名词所有格，当然也包括声母 p, t, k, ch, ts 和 tz。当所有格的撇号出现在这六个声母之后时就会和送气符相混淆**。如 Wang Hsüchou’s 转写结果应为 Wang Xuzhou’s。实验中出现因送气符而产生的转写错误拼音字例有 5 例。

第二类：ü 和 u 的处理：WG 中 ü 和 u 的使用也较为混乱，实际语料中经常会出现该用 u 的时候用了 ü，而该用 ü 的时候却用了 u 的情况，以后者居多（是不是因为输入时键盘上没有直接的 ü 的输入键的缘故？）。而 PY 中 ü 和 u 的情况也比较特殊，详见《汉语拼音方案》。实验中出现因 ü 和 u 混用而产生的转写错误拼音字例有 2 例。

第三类：译者的错误，这主要体现在译者本身的汉语发音水平上。译者在将中文音译为 WG 时并

*事实上，送气符和隔音符的符号并不相同，前者为“‘”而后两者为“’”。但可能由于“‘”在计算机录入时需切换到全拼状态等原因，在正式印刷品中 WG 的送气符也用“’”表示，所以造成了混乱。

**所幸的是查“汉字拼音字切字底表”中的 420 条音节，这六个声母出现在音节末尾的情况并没有出现。因此当音节以’s 结尾时，我们就判定其为名词所有格形式，只做切字而不做转写。

没有通用的普通话，而当时的北京官话还没有普及到现在的普通话这样的程度，因此，译者本身不可避免地带有自己的口音和方言，他在翻译作品中中国人人名、地名时就会依据自己的口音来进行翻译。林语堂先生是福建人，因此在他的口音中有明显的闽方言的特征，如将“zhai（斋）”读成“zai（灾）”，将“bai（白）”读成“bo（伯）”等。因此在《浮生六记》中他将“李白”音译成“Li Po”，将“赵省斋”音译为“Chao Shengtsai”。那么，我们的 WG2PY 在转写时就只能根据规则将“Li Po”转写成“Li Bo”，将“Chao Shengtsai”转写成“Zhao Shengzai”，造成了转写错误。实验中出现因译者口音而产生的转写错误拼音字例有 21 例。

5 结论

由上述分析可见，WG2PY 在进行 WG 转写时除了译者的口音因素之外，送气符时用时不用和 ü、u 的混用是造成转写错误的主要原因。这给我们提出了两点新的思路：（1）通过分析转写错误中译者的口音因素，我们可以进行译者的方言研究；（2）除去译者因素，我们的转写错误率只有 7/235=3%，也就是说转写正确率可达 97%。如此的正确率一方面使我们对该转写系统抱有很大的应用期望，另一方面也促使我们尽快想办法消除这 3%错误率。

我们只是实现了 WG 到 PY 的自动转写，下一步工作是实现其它拉丁字母化拼音法如国语罗马字、拉丁化新文字等到汉语拼音的自动转写以及它们的一体化。同时，面向大数据的真实文本中 WG 拼音词的自动识别与提取也应成为后续研究的一大内容。

参考文献：

[1] 吴鸿适. 关于科学技术名词术语翻译规范化的问题[J]. 中国翻译，1998（3）：27-31.

[2] 洪涛. 评汉英经典文库本《红楼梦》英译的疏失错误[J]. 红楼梦学刊，2006（4）：236-248.

[3] 许长安. 台湾地区的语文政策及其争论[J].现代语文，2006（4）：28-32.

[4] 李宇明. 中华文化迈向国际新步伐——写在中文罗马字母拼写法国际标准（ISO 7098:2015）修订出版之时. 光明日报. 2016-05-01（7）.

[5] 戴金旺，“拉丁字母”和“罗马字母”[J].科技术语研究，2006（1）:44-46.

[6] 周有光. 21 世纪的华语和华文[M]. 北京：三联书店，2002:1-3.

[7] [清]沈复，林语堂英译，浮生六记. 北京：外语教学与研究出版社，1999.

[8] [英]威妥玛，张卫东译，语言自迩集——19世纪中期的北京话[M]. 北京：北京大学出版社，2002