Reversing the Palladius Mapping of Chinese Names in Russian Text

Katherine M. Young
N-Space Analysis, LLC
305 Winding Trail
Xenia, OH 45385
nspaceanalysis@earthlink.net

Jeremy Gwinnup, Joshua Reinhart
SRA International, Inc.
5000 Springfield St., Suite 200
Dayton, OH 45431
Jeremy_Gwinnup@sra.com
Joshua_Reinhart@sra.com

Abstract

We present the Reverse Palladius (RevP) program which assists the linguist in correcting the transliteration of Mandarin Chinese names during Russian to English machine translation. When writing Chinese names, Russian writers follow a traditional Palladius mapping that uses the Cyrillic characters in a unique way. Standard transliteration from Russian into English then produces errors that require hand-correction. RevP uses rules to reverse the Palladius mapping, yielding the correct forms.

When the linguist applies the RevP program to a Chinese name within a document, the name is corrected throughout the document, and the name mappings are saved to a dictionary; these mappings can then be applied to names in new documents, either via the RevP interface or through compilation of the dictionary as a Systran user dictionary.

The RevP program saves time by removing the need for post-editing of Chinese names, and improves consistency in the translation of these names. The user dictionary becomes more useful over time, further reducing the time required for translation of new documents. [1]

1 Introduction

When writing Mandarin Chinese names, Russian writers use a traditional mapping that was created in the 1800s by the Russian monk Palladius (Pyotr Ivanovich Kafarov). This Palladius mapping represents the Chinese sounds in a systematic way that is unfortunately different from the traditional sound values of the Cyrillic characters. For example, the Cyrillic character ж typically represents /zh/, but Palladius used it to represent /r/, while the combination чж /ch zh/ was then used for /zh/. Some examples are given in Table 1; sound values are indicated in Hanyu pinyin.

Cyrillic	Sound	Palladius	Context
ж	zh	r	
ч	ch	ch	
чж	ch zh	zh	
с	s	s	before back vowel
с	s	x	before non-back vowel
н	n	ng	syllable-final
нь	n[2]	n	syllable-final

Table 1. Some Palladius Sound Mappings Compared to Typical Cyrillic Sound Mappings

This separate mapping causes errors for machine translation of Russian to English, in which unknown words may be written out according to the normal sound mappings. For example, the name Чжай Чжиган ‘Zhai Zhigang’ comes out of Systran as Chzhay Chzhigan, as shown in Table 2.

[1] This work is sponsored by the Air Force Research Laboratory under Air Force contract FA8650-09-D-6939. Opinions and recommendations are those of the authors and not necessarily endorsed by the United States Government.

[2] The characters нь typically indicate a palatalized [n].

Chinese	翟志刚
Cyrillic (via Palladius)	Чжай Чжиган
Russian>English translit	Chzhay Chzhigan
Correct pinyin	Zhai Zhigang

Table 2. Error Caused By Automatic Transliteration of a Chinese Name

The linguist must then identify and hand-correct the Chinese names in a post-editing process. While there are programs [3] to automatically convert Chinese to Cyrillic following the (forward) Palladius mapping, we are not aware of any existing programs that produce the reverse mapping. The Reverse Palladius (RevP) program addresses this transliteration need.

2 The Reverse Palladius Program

The RevP program was developed by the Air Force Research Laboratory's Speech and Communication Research, Engineering, Analysis, and Modeling (SCREAM) Laboratory to support Russian-to-English translation efforts. The RevP program provides the linguist the ability to pre-translate Chinese names within a Russian text according to a set of rules that reverse the Palladius mapping, deriving the correct pinyin form. The Palladius rules are sensitive to syllable boundaries, so we first syllabify the Russian words, and then apply the transliteration rules. We considered using a list-based conversion, but found that a rule-based system was more robust in dealing with spelling errors in the Russian.

Additionally, we found that Russian text may contain Chinese words in Palladius form that have subsequently been inflected according to the Russian inflectional system. These inflectional endings need to be removed in order to recover the original Chinese names, so we incorporate a Russian stemmer as well. Examples of inflected Palladius forms are given in Table 3. The user must determine if stemming is appropriate. For example, we want to remove the inflectional ending /-a/ from the name Bominga, but not from the name Baohua, as shown in Table 4.

[3] See, for example, http://mandarinspot.com/annotate and http://www.russki-mat.net/trans3.html.

Inflected	Stemmed	RevP	Meaning
Бомина	Бомин	Boming	genitive
Бомином	Бомин	Boming	instrumental
Бохайского	Бохай	Bohai	adj + genitive[4]

Table 3. Chinese Names with Russian Inflectional Endings

Cyrillic	Stemmed	RevP	Correct pinyin
Баохуа	Баоху	Baohu	Baohua

Table 4. Over-Aggressive Stemming of Potential Russian Inflectional Endings

In general, the RevP program relies on a combination of automatic processing and human decision-making: The user selects a word for examination, RevP automatically provides transliteration variants, the user selects a transliteration, and RevP records the information for re-use in other documents.

2.1 The Document Correction Interface

The RevP program is implemented in Java and utilizes the Eclipse SWT widget library to leverage the host operating system's font display and rendering capabilities. This is particularly useful when displaying languages in non-Latin character sets, as there has been greater internationalization support added to modern operating systems. RevP can process documents in text or Microsoft Word format. The Apache POI library supplies Microsoft Word document support, allowing for the extraction of text from the document. Unfortunately, only the text is extracted; advanced formatting such as columns and pictures are currently not available.

After launching the program, the user opens a Russian document and selects a word to transliterate. The interface provides a menu of possible transliterations. We include the reverse Palladius form (P), a form that is stemmed and then subject to reverse Palladius rules (Ps), a traditional Cyrillic transliteration (Cp), a preview of how Systran would translate the original word (S), and an option for manual entry if none of these are satisfactory. See Figure 1.

[4] From a phrase meaning "the waters of the Bohai sea".

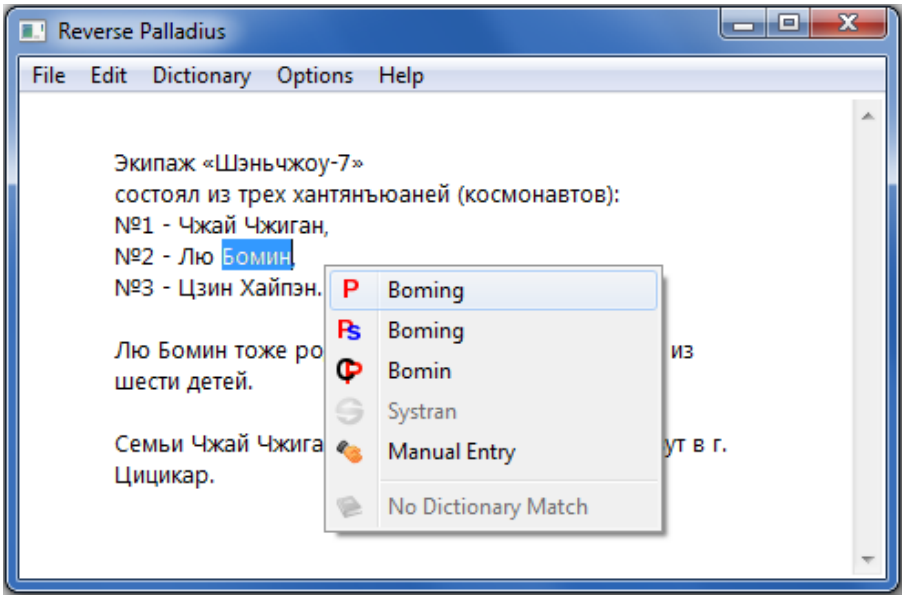

Figure 1. The RevP Transliteration Menu

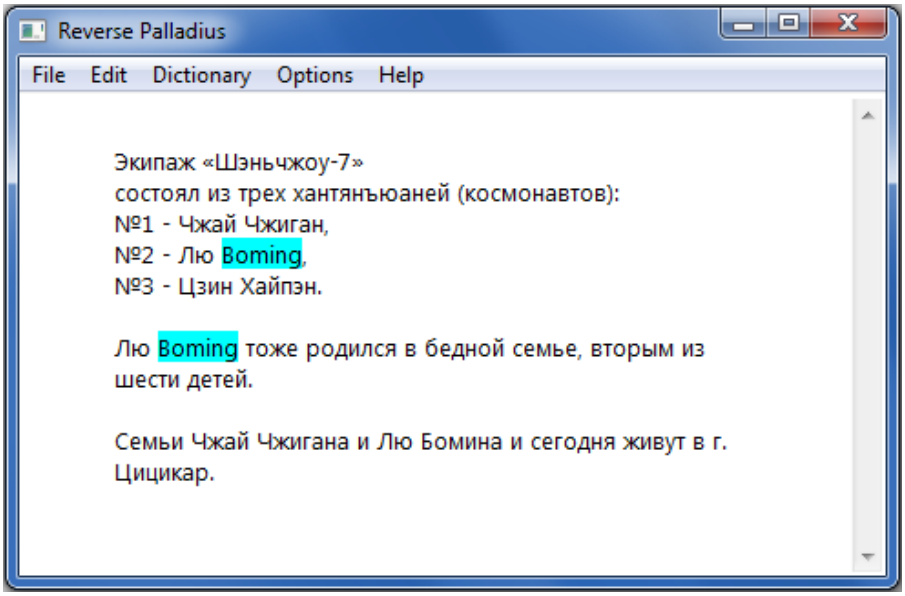

Figure 2. Example of Changing a Word throughout the Document

In this example, describing the crew of the Shenzhou-7 spaceship, the name Бомин is selected, and the first option, applying the reverse Palladius mapping, gives the correct result, Boming. There is no potential inflectional ending, so the stemmed option defaults to the same form, Boming. Traditional Cyrillic transliteration gives the incorrect variant, Bomin; no Systran translation is provided in this example.

After the user makes a selection, the word is changed within the document, and a color highlight is applied. The program offers the option to change all instances of the word throughout the document; these changes are also highlighted. See Figure 2, in which two instances of the name Бомин have been changed to Boming.

If Systran is available in the user's system, RevP provides a preview of the Systran translation for the selected word. Sometimes the preview shows that Systran already knows how to translate a name, and the RevP transliteration is not necessary. The connection to Systran is effected using Systran 7's SOAP API interface. Using Apache AXIS to create a Java binding from the provided WSDL file, we can provide a reasonable translation for the selected word or phrase. Additional translation system interfaces could also be implemented without much effort.

The traditional Cyrillic transliteration is provided in case the selected word is not a Chinese word, but a word borrowed from another language, such as English. The name John Glenn, for example, follows traditional sound mappings to become Джон Гленн in Russian texts. In such cases, the traditional mapping is more useful than the reverse Palladius mapping. Also, we have seen some more recent documents in which the Russian writer has not used the Palladius system for Chinese names, but instead applied traditional Cyrillic sound mappings, writing the name Guo Guangchang, for example, as Гуо Гуангчанг.[5] See Table 5.

Cyrillic	Джон Гленн
Palladius	Drong Gliengn
Cyrillic Sounds	Dzhon Glenn
Correct Spelling	John Glenn
Cyrillic	Гуо Гуангчанг
Palladius	Guo Guanggchangg
Cyrillic Sounds	Guo Guangchang
Correct Spelling	Guo Guangchang

Table 5. Names Written with Traditional Cyrillic Sound Mappings

Manual entry is provided for words like Джон 'John' above, for which none of the automatic transliterations are correct. This is also useful if the original Russian text has been misspelled, leading to errors in the transliteration. Manual entry also allows the user to correct names from Taiwan, which may need to be written according to Tongyong pinyin instead of the Hanyu pinyin generated by RevP.

RevP records the user's choices in a dictionary file, as shown in Figure 3. In this example, the user has made corrections to the names of the three cosmonauts from the text in Figure 1.

```
#ENCODING=UTF-8
#SUMMARY=REVP-Systran/cosmonauts.rpd
#MULTI
#RU          EN
Хайпэн       Haipeng
Лю           Liu
Чжай         Zhai
Цзин         Jing
Бомин        Boming
Чжиган       Zhigang
```

Figure 3. A RevP Dictionary File

RevP saves the dictionary file in a Systran-compatible format.

2.2 Applying the Dictionary to a New Document

After a dictionary of name mappings has been developed, the linguist can use RevP to apply that dictionary to new documents. This can be done with or without Systran.

When not using Systran, the user can open a new document in RevP, load the previously created dictionary, and automatically apply the dictionary to pre-translate any matching names in the new document. The modified document can then be submitted to a machine translation system.

If Systran is available, the user can instead compile the RevP dictionary as a user dictionary within Systran; any new documents submitted to Systran will then follow the translations specified in the user dictionary.

For example, we took the user dictionary shown in Figure 3 and compiled it as a Systran user dictionary. We then used Systran to translate a new document containing some of the same names. Table 6 shows the difference in the Systran translation with and without the RevP dictionary. This example was conducted using Systran Version 5.0, with the option of marking unknown, transliterated words with square brackets.

[5] http://www.peoples.ru/undertake/founder/guo_guangchang/

Russian
Тем временем Чжай Чжиган и Лю Бомин перешли в орбитальный модуль и начали подготовку к выходу.

Systran, no user dictionary
Meanwhile [Chzhay] Of [chzhigan] and [Lyu] Of [bomin] passed into the orbital module and they began preparation for the output.

Systran, with user dictionary
Meanwhile Zhai Zhigang and Liu Boming they passed into the orbital module and they began preparation for the output.

Table 6. Using the RevP Dictionary in Systran to Improve Translation of a New Document

2.3 Future Work

The RevP program relies on human decision-making, both in selecting the Chinese names for transliteration and in choosing the correct transliteration option. For future versions of RevP, we hope to integrate named entity tagging to facilitate the identification of the Chinese names in the Russian text.

We would also like to be able to tag our pre-translated names as do-not-translate elements for Systran, to protect them from re-ordering or other processing during the Russian-to-English machine translation. In particular, we want to protect these words from the Systran program's tendency to insert the word "of" before English words. We currently provide an option within RevP to create a supplemental English-to-English dictionary for all the pre-translated words; this can be compiled as a second user dictionary, forcing Systran to translate these words without changes. Do-not-translate tags would be a more useful solution to this problem.

The RevP program could also be improved in the future by adding a transliteration mapping for the Tongyong pinyin system used in Taiwan.

Finally, we consider that the RevP framework for the automatic generation of transliteration options could be used for other languages that have multiple systems for transliterating foreign words, by specifying the syllabification and transliteration rules needed for the new language. The RevP framework could also be adapted to serve as part of a search engine interface in which a foreign-word search term might need to be instantiated in various transliterated forms.

3 Conclusion

The RevP program has proven to be a useful tool for dealing with the specialized transliteration of Chinese names within Russian text. By automating the reversal of the Palladius mapping of Chinese sounds in Cyrillic, we provide the user the ability to correctly pre-translate these names before submitting a document to Russian-to-English machine translation. As the user continues to apply RevP to various documents, a larger name dictionary is created, allowing improved translation of new documents. The dictionary can be applied to specific documents via RevP, creating pre-translated documents suitable for further hand or machine translation, or it can be compiled in Systran as a user dictionary and applied to all future documents.