Proper Name Transliteration with ICU Transforms

Sascha Brawer
Martin Jansche
Hiroshi Takenaka 竹中 浩
Yui Terashima 寺島 有為
Proper Name Transliteration with ICU Transforms

Motivation
Case study: Google Maps™

Goal: World maps in Japanese, Mandarin, Russian, etc.

Internationalization of geographic names for display.

For example, Reading becomes レディング (redingu) in Japanese, or Рединг (Reding) in Russian.

Millions of labels are affected.

A large part of the transliteration effort can be automated.
What’s in a name?

Labels in Google Maps are not limited to toponyms. Many labels have a complex internal structure and contain names of persons (e.g. in street names) and organizations (e.g. in labels for points of interest).

Labels in Google Maps also contain common nouns like city or river, which are not always transliterated.

We focus on those (parts of) labels that need to be transliterated.
Transliteration

We use **transliteration** in a loose sense to mean the transformation of a word from a source language into a sequence of similar **sounds** in the target language.

Contrast this with **translation**, which preserves the **meaning** of a word, phrase, or text. We won’t have anything to say here about translation.
More precisely: Strict transliteration

We use transliteration in a broad sense to cover a range of relations ranging from strict transliteration to transcription.

Strict transliteration: A function, often bijective, from strings in one script to strings in a different script. Can be independent of language and writing system: e.g. it is possible to define a transliteration of Cyrillic into Latin script that applies to Russian, Ukrainian, Serbian, etc.

Strict transliteration systems have limited practical uses. They used to be necessary in scholarly work when it was not feasible to represent text in mixed scripts, e.g. for Western scholars working on Russian or vice versa.
More precisely: Transcription

Transcription for our purposes: Representing the approximate sounds of loanwords from a source language in the writing system of a target language.

Phonetic borrowing of loanwords itself does not assume writing or literacy and is independent of writing systems. E.g. French loanwords in Turkish used to be written in Arabic script and are now written in Latin script.

In present-day Turkish, kampüs (from French campus) or prodüktör (from French producteur) follow the conventions of Turkish orthography to represent the sounds of words borrowed from French.
Phonetic/phonemic transcription

General **phonetic transcription** systems like the International Phonetic Alphabet have dedicated symbols and conventions for writing a wide variety of naturally occurring human speech sounds.

**Phonemic transcription** represents the sound system of a particular language in a phonemic writing system. Examples: Hanyu Pinyin or Bopomofo for Mandarin; Hiragana, Katakana, or Romaji for Japanese; Korean respelling for Korean; Thai respelling for Thai; Dania lydskrift for Danish; etc.
Practically useful transliteration

Strict transliteration preserves a maximum of orthographic detail of the source language (including phonetically meaningless distinctions), at the expense of readability in the target language.

Phonetic transcription preserves a maximum of phonetic detail of the source language (including sounds which do not exist in the target language or which are subtly different from ones that do), at the expense of readability.

Practical transliteration systems are hybrids that strike a balance between these two extremes, sacrificing both orthographic and phonetic details.
Overview

Lesson #1: Figuring out what to transliterate and how is typically much harder than implementing the transliteration scheme.

We illustrate several practical issues with transliteration into three target languages of increasing complexity: Russian, Japanese, and Mandarin.

We will use Spanish as our default source language. This is purely for expository convenience: Spanish is widely spoken and studied, and has a simple orthography. We could have used Italian, Czech, Korean, etc. instead while leaving the take-home points unchanged.
Proper Name Transliteration with ICU Transforms

Target: Russian
Transliteration into Russian

Not very different from Romanization.

Caveat: There usually is no common cyrillization that would target Russian, Ukrainian, Bulgarian, Serbian, and the many other languages written in Cyrillic script. This is because the inventory of letters and the pronunciation conventions differ between these languages.

Illustrates our point that the target of transcription is a language and its writing system, rather than merely a script.
### Spanish-to-Russian transliteration

Based mostly on phonology:

<table>
<thead>
<tr>
<th>Spanish Place</th>
<th>Russian Transliteration</th>
</tr>
</thead>
<tbody>
<tr>
<td>{G, J, X}iménez</td>
<td>Хименес</td>
</tr>
<tr>
<td>Gijón</td>
<td>Хихон</td>
</tr>
<tr>
<td>Jerez (Xeres)</td>
<td>Херес</td>
</tr>
<tr>
<td>Cuenca</td>
<td>Куэнка</td>
</tr>
<tr>
<td>Quesada</td>
<td>Кесада</td>
</tr>
<tr>
<td>Quintana</td>
<td>Кинтана</td>
</tr>
<tr>
<td>Ceuta</td>
<td>Сеута</td>
</tr>
<tr>
<td>Zaragoza</td>
<td>Сарагоса</td>
</tr>
<tr>
<td>La Mancha</td>
<td>Ла-Манча</td>
</tr>
<tr>
<td>Huelva</td>
<td>Уэльва</td>
</tr>
</tbody>
</table>
Spanish-to-Russian (continued)

Many Spanish sounds have close equivalents in Russian, but some distinctions are lost or only retained in spelling:

<table>
<thead>
<tr>
<th>Spanish</th>
<th>Russian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Griñón [ɡɾiɲon]</td>
<td>Гриньон</td>
</tr>
<tr>
<td>Logroño [loɣɾoɲo]</td>
<td>Логроньо</td>
</tr>
<tr>
<td>Llobregat [ʎoβɾeɣat]</td>
<td>Льобрегат</td>
</tr>
<tr>
<td>Sagra [saɣɾa]</td>
<td>Сагра</td>
</tr>
<tr>
<td>Zagra [θaɣɾa]</td>
<td>Сагра</td>
</tr>
<tr>
<td>Ascensión [asθensjon]</td>
<td>Асенсьон</td>
</tr>
<tr>
<td>Laredo [laɾeðo]</td>
<td>Ларедо</td>
</tr>
<tr>
<td>Casalarreina [kasalar:r:ejna]</td>
<td>Касаларрейина</td>
</tr>
<tr>
<td>Casa la Reina [kasalar:r:ejna]</td>
<td>Каса-ла-Рейна</td>
</tr>
</tbody>
</table>
Even though Spanish \( b \) and \( v \) are pronounced exactly the same, the orthographic distinction is retained in Russian:

- **Bargas** [barɣas] \( \rightarrow \) Баргас
- **Vargas** [barɣas] \( \rightarrow \) Варгас
- **Ribera** [rɨβeɾa] \( \rightarrow \) Рибера
- **Rivera** [rɨβeɾa] \( \rightarrow \) Ривера

**Exception:**
- **Córdoba** [koɾðoβa] \( \rightarrow \) Кордова

**Lesson #2:** Names of well-known places, persons, etc. often have exceptional transliterations.
Spanish-to-Russian (conclusion)

Some aspects of transliteration follow Russian orthographic conventions:

- **Yanguas** → Янгвас
- **Yecla** → Екла
- **Estepona** → Эстепона
- **Yuncos** → Юнкос
- **Soria** → Сория, **not** Сориа
- **El Escorial** → Эль-Эскориаль

**Lesson #3:** Established conventions or constraints of the target language and/or writing system must be taken into account.
Spanish-to-Russian ICU rules

A simplified version of Spanish-to-Russian transliteration expressed in the ICU transform rule language:

\[
\begin{align*}
&b &\rightarrow & б ; \\
&ch &\rightarrow & ч ; \\
&c &\rightarrow & с ; & # \\ & & \text{before ’e’ or ’i’} \\
&d &\rightarrow & к ; \\
&[-\ \$] \{ &e &\rightarrow & э ; & # \\ & & \text{at beginning of word} \\
&e &\rightarrow & е ; \\
&f &\rightarrow & ф ; \\
&gu &\rightarrow & г ; \\
&g &\rightarrow & х ; \\
&h &\rightarrow & ; & # \\ & & \text{empty string}
\end{align*}
\]
Proper Name Transliteration
with ICU Transforms

Target: Japanese
Transliteration into Japanese

Katakana script is used for all types of transcription. Somewhat similar to use of italics in western typesetting to denote foreign words.

Transliteration is usually purely phonemic, subject to the constraints of Japanese extended phonology, which permits a few sound combinations not found in native Japanese words.

We also have to deal with the moraic nature of the Katakana writing system.
## Spanish-to-Japanese transliteration

Based mostly on phonology:

<table>
<thead>
<tr>
<th>Spanish</th>
<th><a href="KEI">音</a></th>
<th>Japanese</th>
<th><a href="KEI">音</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>{G, J, X}iménez</td>
<td>[ximeneθ]</td>
<td>ヒメネス</td>
<td>himenesu</td>
</tr>
<tr>
<td>Gijón</td>
<td>[xiθon]</td>
<td>ヒホン</td>
<td>hihon</td>
</tr>
<tr>
<td>Cuenca</td>
<td>[kweŋka]</td>
<td>クエンカ</td>
<td>kuenka</td>
</tr>
<tr>
<td>Quesada</td>
<td>[kesadøa]</td>
<td>ケサダ</td>
<td>kesada</td>
</tr>
<tr>
<td>Quintana</td>
<td>[kintana]</td>
<td>キンタナ</td>
<td>kintana</td>
</tr>
<tr>
<td>Ceuta</td>
<td>[θewta]</td>
<td>セウタ</td>
<td>seuta</td>
</tr>
<tr>
<td>Zaragoza</td>
<td>[θaragoθa]</td>
<td>サラゴサ</td>
<td>saragosa</td>
</tr>
<tr>
<td>La Mancha</td>
<td>[la mantʃa]</td>
<td>ラ・マンチャ</td>
<td>ra mancha</td>
</tr>
<tr>
<td>Huelva</td>
<td>[welβa]</td>
<td>ウエルバ</td>
<td>ueruba</td>
</tr>
</tbody>
</table>
Spanish-to-Japanese (continued)

Many Spanish sounds have close equivalents in Japanese, but some distinctions are inevitably lost:

<table>
<thead>
<tr>
<th>Spanish</th>
<th>Japanese</th>
<th>Transcription</th>
</tr>
</thead>
<tbody>
<tr>
<td>Griñón</td>
<td>グリニョン</td>
<td>gurinyon</td>
</tr>
<tr>
<td>Logroño</td>
<td>ログロニョ</td>
<td>roguɾonyo</td>
</tr>
<tr>
<td>Llobregat</td>
<td>リョブレガト</td>
<td>ryoburegato</td>
</tr>
<tr>
<td>Sagra</td>
<td>サグラ</td>
<td>sagura</td>
</tr>
<tr>
<td>Zagra</td>
<td>サグラ</td>
<td>sagura</td>
</tr>
</tbody>
</table>
Spanish-to-Japanese (conclusion)

Some aspects of transliteration are governed by Japanese extended phonology:

<table>
<thead>
<tr>
<th>Spanish</th>
<th>Japanese</th>
<th>Japanese (consonant modification)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Santiago</td>
<td>サンティアゴ</td>
<td>santiago</td>
</tr>
<tr>
<td>not</td>
<td>サンチアゴ</td>
<td>sanchiago</td>
</tr>
<tr>
<td>Alcorcón</td>
<td>アルコルコン</td>
<td>arukorukon</td>
</tr>
<tr>
<td>Chimbote</td>
<td>チンボテ</td>
<td>chinbote</td>
</tr>
<tr>
<td>Motril</td>
<td>モトリル</td>
<td>motoriru</td>
</tr>
<tr>
<td>Rodríguez</td>
<td>ロドリゲス</td>
<td>rodorigesu</td>
</tr>
<tr>
<td>Ciudad</td>
<td>シウダー [θuðað]</td>
<td>shiudā</td>
</tr>
<tr>
<td>closer</td>
<td>シウダド</td>
<td>shiudado</td>
</tr>
<tr>
<td>than</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Phonotactic constraints are met by vowel epenthesis.
Spanish-to-Japanese ICU rules

Since transliteration into Japanese is almost purely phonemic, it makes sense to go via a phonemic representation of the source language:

:: Spanish-SpanishPhonemic;
:: SpanishPhonemic-Japanese;
Spanish pronunciation rules

The Spanish-Spanish Phonemic transform produces Spanish pronunciations, expressed here in IPA:

:: Lower;
b → β ;
ch → tʃ ;
c } [ei] → θ ;
c → k ;
d → ð ;
:: Null; # Second pass: positional allophones.
[mnɲŋ $] { β → b ;
[mnɲŋ $] { ð → d ;
::
# From Spanish phonemes to Katakana

# First pass: Collapse irrelevant phonemic distinctions.
θ → s;
ɾː → ɾ;
l → ɾ;
:: Null; # Second pass: Phonemes to Katakana.
a → ア;
ba → バ;
bi → ビ;
bu → ブ;
be → ベ;
bo → ボ;
b → ブ;
::
Summary: Japanese

Japanese is an easy language to work with, since transliteration is usually straightforwardly phonemic. If pronunciations for the source language are readily available, rule-based transliteration is easy.

Google Maps in Japanese uses transliteration from English, French, Italian, German, Spanish, Dutch, Russian, Polish, Czech, Catalan, Welsh, etc., though not all cases use ICU transforms.
Proper Name Transliteration with ICU Transforms

Target: Mandarin
Transliteration into Mandarin

Note that the target language is Mandarin, not “Chinese”. Transliteration into e.g. Cantonese is different.

All transliteration schemes target a small subset of common Chinese characters. E.g. transliteration from Spanish uses fewer than 250 characters, most of them shared with other transliteration schemes. When used in transliteration, these characters only denote sounds; their meanings are ignored.

There are official standards for a few important source languages, and de facto standards based on reference works for many other source languages.
Transliteration vs. translation

Partial translations and calques make Mandarin tricky:

**Washington, D.C.**
华盛顿 哥伦比亚 特区
huáshèngdùn gēlúnbǐyà tèqū
(phonetic) (phonetic) special district

**Little Rock** (Arkansas)
小 石 城
xiǎo shí chéng
little rock city

**Treinta y Tres** (Uruguay)
三十三 人 城
sānshísān rén chéng
33 people city
Assume that we have sorted out translation vs. transliteration and ordinary vs. exceptional transliteration. Ordinary transliteration is then quite straightforward.

Ordinary transliteration rules are usually expressed in the form of tables, owing to a long history of rhyme tables in Chinese phonology.

Tables show onsets and codas of syllables, conforming to the constraints of Mandarin phonology. Transliteration amounts to a greedy leftmost-longest rewriting of the input string according to the table.
## Spanish-to-Mandarin table

Transliteration table from Chinese national standard GB/T 17693.5–1999 (excerpt, slightly simplified):

<table>
<thead>
<tr>
<th></th>
<th>b-</th>
<th>p-</th>
<th>d-/ð-</th>
<th>t-</th>
<th>g-/ɣ-</th>
<th>k-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ø-</td>
<td>布</td>
<td>bù</td>
<td>普</td>
<td>pǔ</td>
<td>德</td>
<td>dé</td>
</tr>
<tr>
<td>-Ø</td>
<td>布</td>
<td>bù</td>
<td>普</td>
<td>pǔ</td>
<td>德</td>
<td>dé</td>
</tr>
<tr>
<td>-a</td>
<td>阿</td>
<td>ā</td>
<td>巴</td>
<td>bā</td>
<td>帕</td>
<td>pà</td>
</tr>
<tr>
<td>-e</td>
<td>埃</td>
<td>āi</td>
<td>贝</td>
<td>bèi</td>
<td>佩</td>
<td>pèi</td>
</tr>
<tr>
<td>-ej</td>
<td>埃</td>
<td>āi</td>
<td>贝</td>
<td>bèi</td>
<td>佩</td>
<td>pèi</td>
</tr>
<tr>
<td>-i</td>
<td>伊</td>
<td>yī</td>
<td>比</td>
<td>bǐ</td>
<td>皮</td>
<td>pí</td>
</tr>
<tr>
<td>-o</td>
<td>奥</td>
<td>ào</td>
<td>博</td>
<td>bó</td>
<td>波</td>
<td>bō</td>
</tr>
<tr>
<td>-ow</td>
<td>欧</td>
<td>ōu</td>
<td>博</td>
<td>bó</td>
<td>波</td>
<td>bō</td>
</tr>
<tr>
<td>-u</td>
<td>乌</td>
<td>wū</td>
<td>布</td>
<td>bù</td>
<td>普</td>
<td>pǔ</td>
</tr>
</tbody>
</table>

**Example:** Pico del Teide → [tej.ðe] → 泰德峰
The transliteration scheme always collapses the [l] / [ɾ] / [rː] distinction phonetically, but maintains it sometimes in the orthography:

<table>
<thead>
<tr>
<th>Spanish</th>
<th>Mandarin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logroño</td>
<td>洛格罗尼奥</td>
</tr>
<tr>
<td>Malón</td>
<td>马隆</td>
</tr>
<tr>
<td>Marrón</td>
<td>马龙</td>
</tr>
</tbody>
</table>

The Spanish [r] phonemes are systematically lost, even though Mandarin has somewhat similar sounds: e.g. 若 (ruò) could have been used instead of 罗 (luó), or 容 (róng) instead of 龙 (lóng). (They are used instead to transcribe the [ʒ] sound, e.g. in words of French origin.)
The transliteration scheme distinguishes $[b]$ from $[\beta]$, but conflates $[d]$ and $[\ddot{a}]$ (as well as $[g]$ and $[\gamma]$):

- Valencia $[\text{ba}l\text{e}n\thetaja]$ 巴伦西亚 bā lún xī yà
- Córdoba $[\text{ko}\ddot{o}\betaa]$ 科尔多瓦 kē ěr duō wǎ
- Mondoñedo $[\text{mon}do\etae\ddot{o}]$ 蒙多涅多 méng duō niè duō

Contrast this with transliteration into Japanese, where these distinctions are never preserved; and with transliteration into Russian, where the orthographic ‘b’ / ‘v’ distinction is preserved instead.
Which Spanish?

The GB/T 17693.5 standard mandates that one of the key differences in pronunciation between European Spanish and Latin American Spanish – the so-called *yeísmo* – be reflected in the transliteration.

Castilla, Spain [kastiʎa] 卡斯蒂利亚 kǎ sī dì lì yà
Castilla Province, Peru [kastiʝa] 卡斯蒂亚 kǎ sī dì yà

In order to transliterate a Spanish-language place name correctly according to the Chinese national standard, we must know which continent the place is found on. Since our main application is Google Maps, this turned out not to be a problem.
Spanish-to-Mandarin ICU rules

At the highest level, Spanish-to-Mandarin transliteration consists of three components:

:: Spanish-SpanishPhonemic;
:: EuropeanSpanish-LatinAmericanSpanish; # as needed
:: SpanishPhonemic-Mandarin;

We can simply re-use the pronunciation rules we developed earlier for Spanish-to-Japanese transliteration.

The rules for transforming European into Latin American Spanish pronunciations are trivial:

\[ \breve{\text{a}} \rightarrow \text{j} ; \]
\[ \theta \rightarrow s ; \]
From Spanish phonemes to Mandarin

# First pass: Collapse irrelevant phonemic distinctions.
θ → s ;
ð → d ;
ɣ → g ;
:
:: Null;  # Second pass: Phonemes to Hanzi.
aj → 艾;
an } [^aeiou] → 安;
aw → 奥;
a → 阿;
baj → 拜;
ban } [^aeiou] → 班;
baw → 包;
ba → 巴;
:
ICU rules from transliteration tables

Chinese reference works like 《世界人名翻译大辞典》（“Names of the World’s Peoples”）provide transliteration tables for more than 50 source languages.

From a machine-readable version of a transliteration table an equivalent ICU transform can be derived automatically. That is the easy part.

The hard part is making sense of the tables. They are all expressed in terms of the source-language orthography and its pronunciation given in IPA. Often, these don’t cover all the corner cases, and the given orthography and pronunciation conflict sometimes.
Lessons Learned
How to develop ICU transforms

Test-driven development, empirical validation.

Official standards are often fuzzy around the edges. They cannot be treated as rigorous specifications that can be implemented as written. One typically has to look at actual usage in order to get the corner cases right. If there are no standards, then observed usage is often the only guidance.

Gather a corpus of transliteration pairs. Some corpora can be purchased (e.g. English-Chinese from Xinhua via LDC), but for the bulk of language pairs there are no off-the-shelf corpora. Crawl the Web, Wikipedia, etc.
How to modularize

Key component: Pronunciation rules for the source languages.

Often, the hardest part of transliteration is knowing how a name is pronounced in its source language.

Pronunciation rules/models are highly reusable, since they represent an abstract truth about the source language, with no reference to transliteration.

We saw one example of this reuse: we used the same Spanish pronunciation rules as the first step in transliterating into Japanese and Chinese.
The cross-product problem

Is there an interlingua that would make transliterating from M source languages into N target languages simple? It is tempting to wish that phonetic notation might fill this role.

By pivoting through IPA, we would have to implement only M + N individual transforms (compare with compilers):
No interlingua for transliteration

Practical transliteration schemes are messy, with no single authority in charge globally. Even for the same target language, where national academies could impose uniform standards, transliteration schemes vary across source languages.

<table>
<thead>
<tr>
<th>Language</th>
<th>Russian</th>
<th>Japanese</th>
<th>Mandarin</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td>Schöna [ʃønaʊ]</td>
<td>Šënau shënau</td>
<td>Busch [bʊʃ]</td>
</tr>
<tr>
<td>French</td>
<td>Dreux [dʁœ]</td>
<td>Drē dorū</td>
<td>Franche [fʁɑʃ]</td>
</tr>
</tbody>
</table>
Pronunciation rules for the source languages are very general, highly reusable components. They greatly simplify the construction of transliteration schemes between many languages, even if the full cross-product cannot be avoided in the worst case.
By reusing common pronunciation rules, the transliteration problem becomes much simpler. The mapping between source-language phonemes and target-language phonemes may need to be written separately for each language pair, but this is usually very straightforward.

For transliteration into Mandarin, a different transliteration table is used for each source language. Here too the bulk of the work lies elsewhere, since ICU rules can be generated automatically from a given transliteration table.
Conclusions

We discussed transliteration into three very different target languages. We saw that ICU transforms can express the transliteration schemes in a modular way that allows for reuse of core components. Our approach has been used as part of a larger effort to internationalize Google Maps. Millions of geographic names have been automatically transliterated using a variety of ICU transform rules, as well as other techniques.

ICU transforms are ideally suited for implementing transliteration schemes. In fact, the implementation itself is often the simplest part. Figuring out the intricacies of certain transliteration schemes can be much harder.
Proper Name Transliteration with ICU Transforms

Thank You