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™

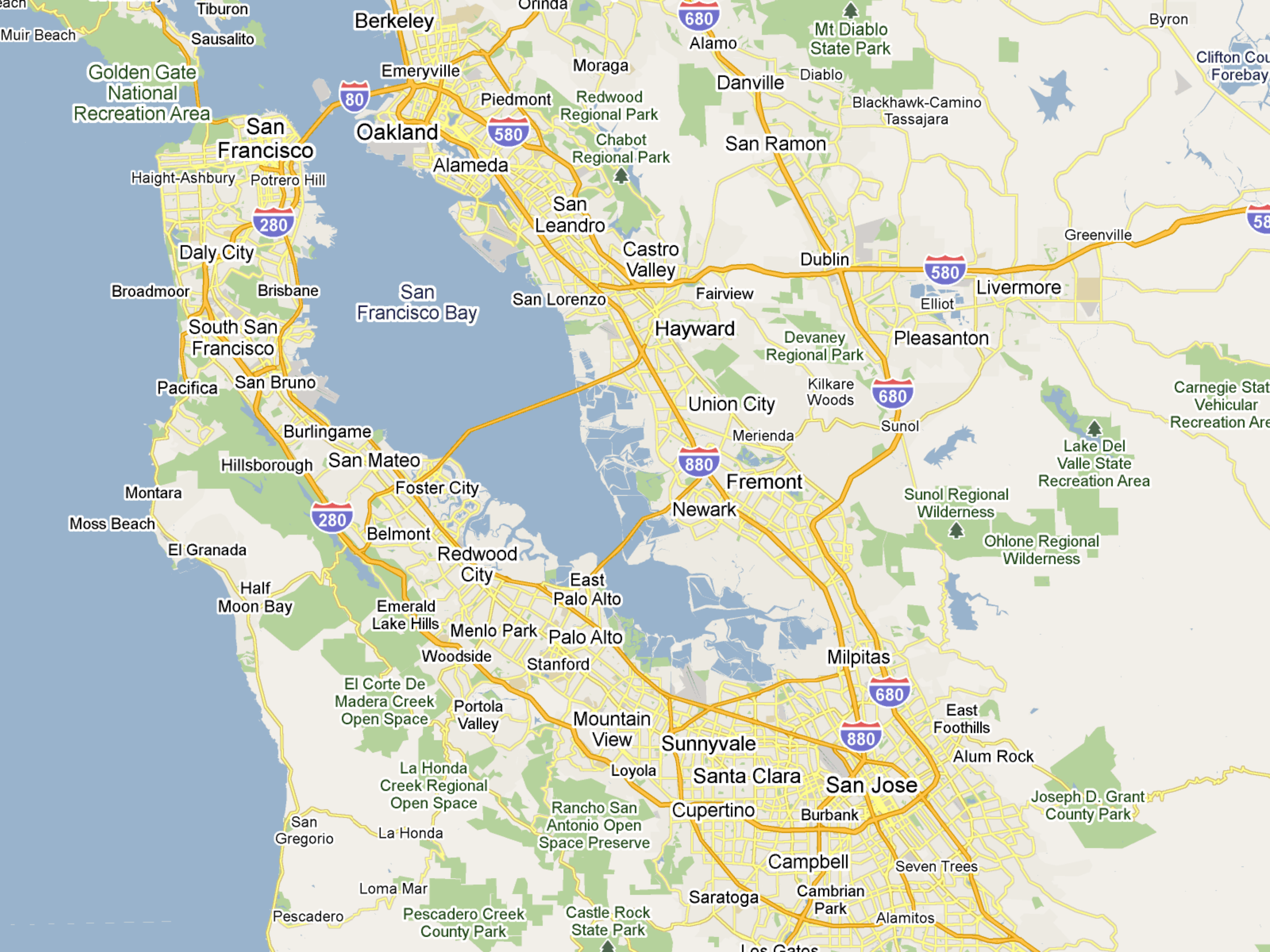
Goal: World maps in Japanese, Mandarin, Russian, etc.

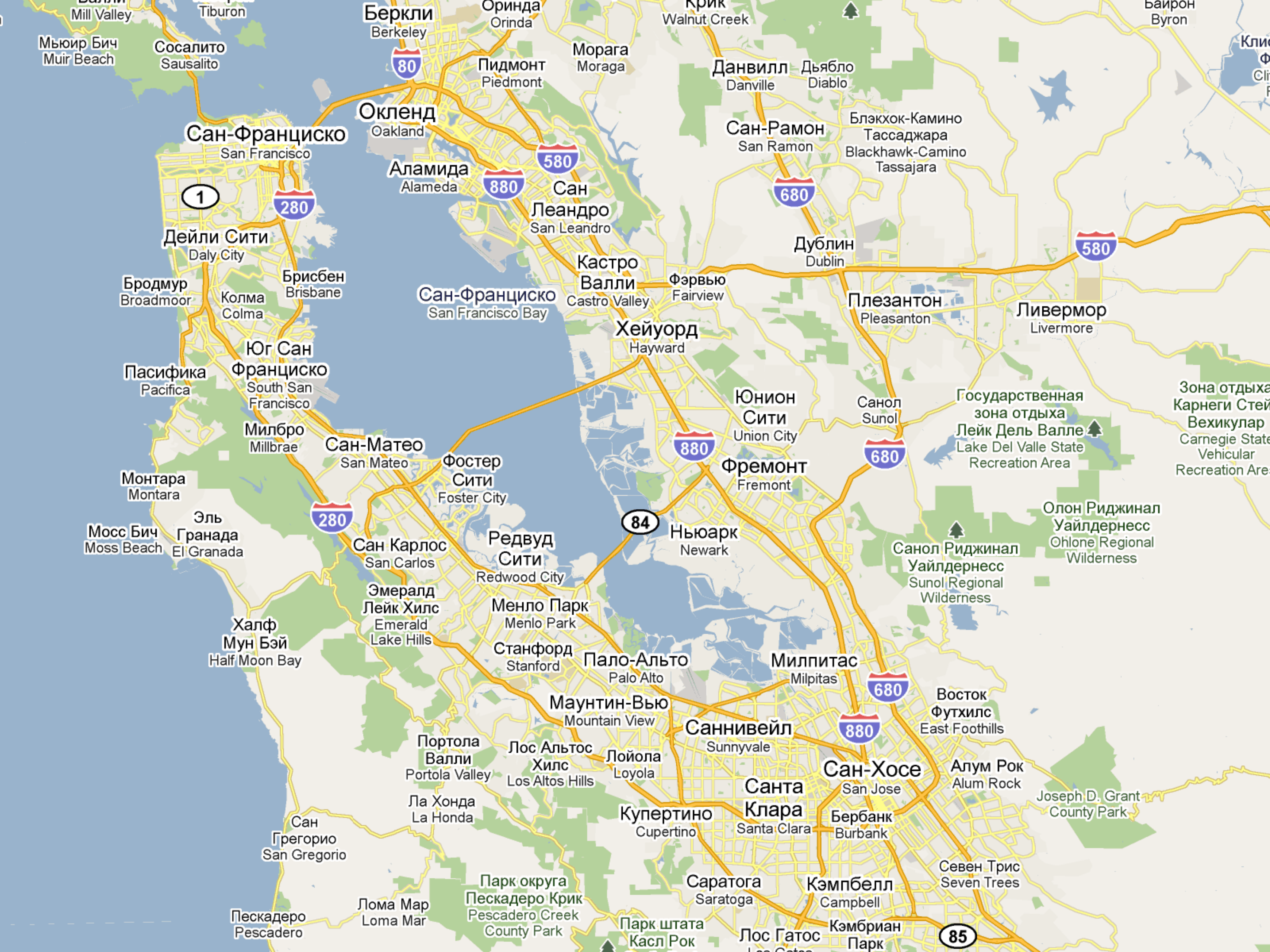
Internationalization of geographic names for display.

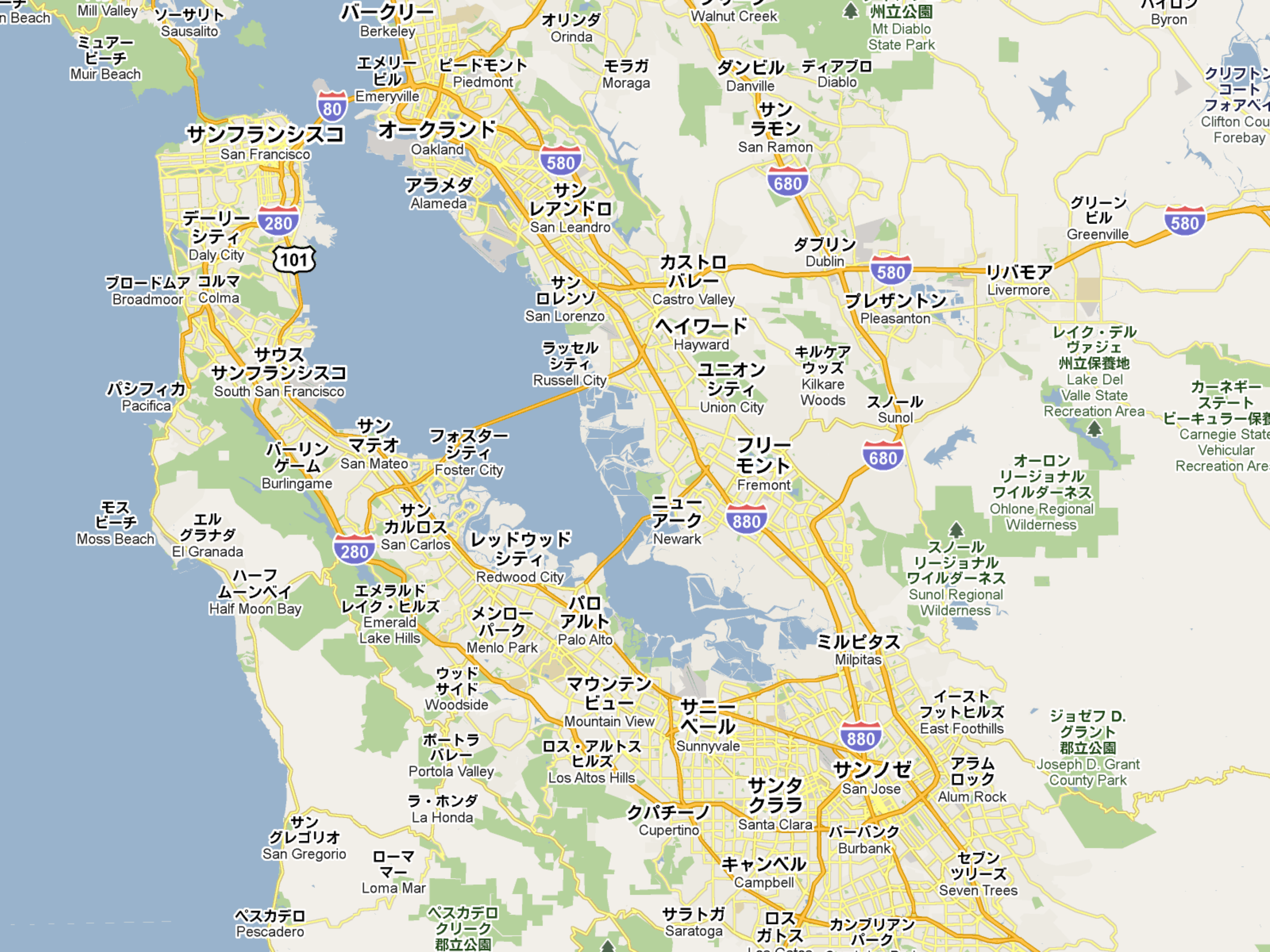
For example, Reading becomes レディング (redingu) in Japanese, or Рединг (Reding) in Russian.

Millions of labels are affected.

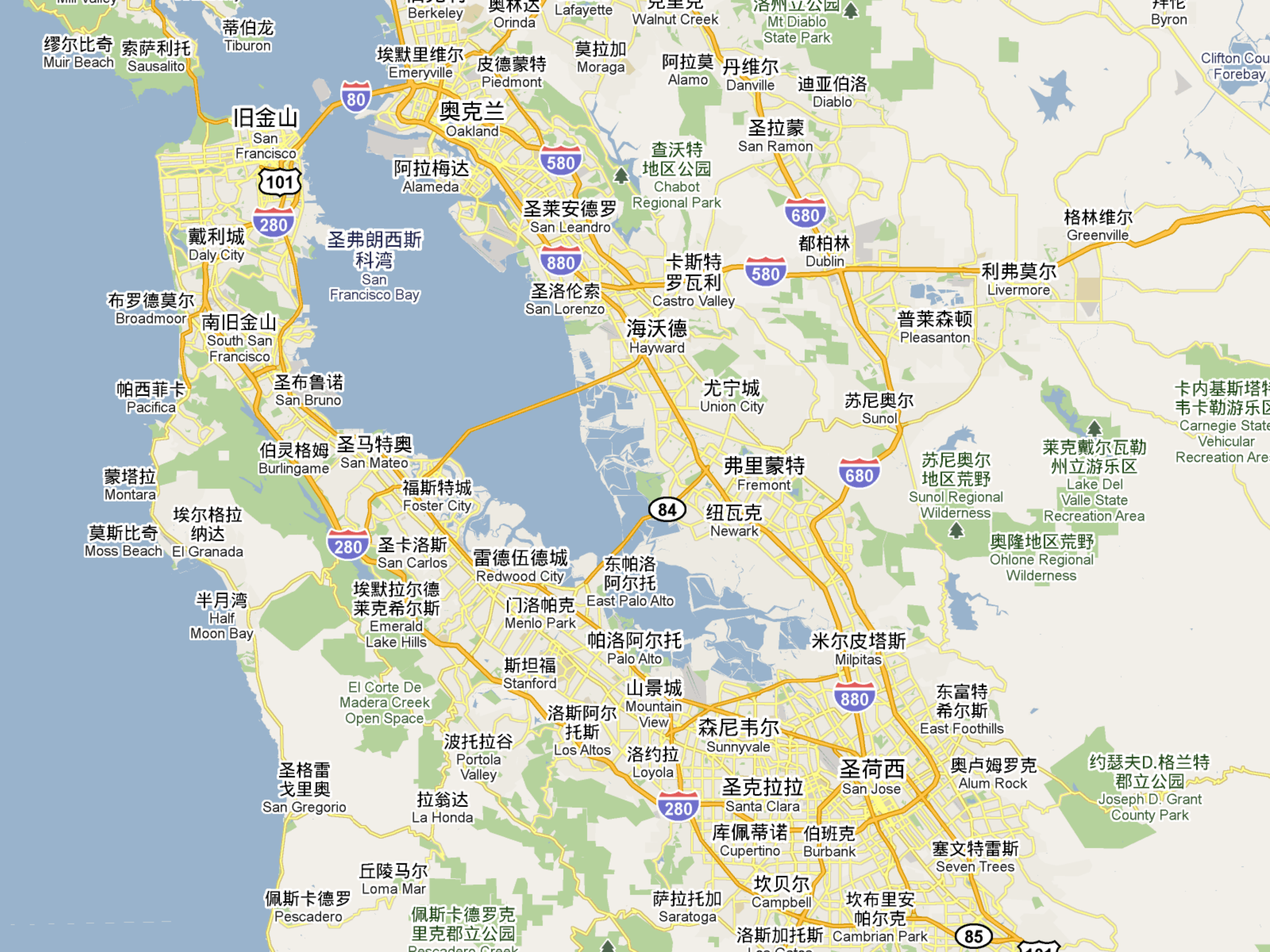
A large part of the transliteration effort can be automated.


Tiburon
Berkeley
Orinda
680
Mt Diablo State Park
Byron
Muir Beach
Sausalito
Alamo
Clifton Cou Forebay
Golden Gate National Recreation Area
Emeryville
Moraga
Danville
Diablo
80
Piedmont
Redwood Regional Park
Blackhawk-Camino Tassajara
San Francisco
Oakland
580
Chabot Regional Park
San Ramon
Alameda
Haight-Ashbury
Potrero Hill
San Leandro
280
Greenville
58
Daly City
Castro Valley
Dublin
580
Livermore
Broadmoor
Brisbane
San Francisco Bay
San Lorenzo
Fairview
Elliot
Hayward
South San Francisco
Devaney Regional Park
Pleasanton
Pacifica
San Bruno
Kilkare Woods
680
Carnegie Stat Vehicular Recreation Are
Union City
Sunol
Burlingame
Merienda
Lake Del Valle State Recreation Area
San Mateo
880
Hillsborough
Fremont
Montara
Foster City
Sunol Regional Wilderness
Moss Beach
280
Newark
Belmont
Ohlone Regional Wilderness
El Granada
Redwood City
East Palo Alto
Half Moon Bay
Emerald Lake Hills
Menlo Park
Palo Alto
Woodside
Stanford
Milpitas
El Corte De Madera Creek Open Space
680
Portola Valley
East Foothills
Mountain View
Sunnyvale
880
Alum Rock
La Honda Creek Regional Open Space
Loyola
Santa Clara
San Jose
Rancho San Antonio Open Space Preserve
Cupertino
Joseph D. Grant County Park
Burbank
San Gregorio
La Honda
Campbell
Seven Trees
Loma Mar
Cambrian Park
Saratoga
Pescadero
Pescadero Creek County Park
Castle Rock State Park
Alamitos
Los Gatos


Mill Valley
Tiburon
Беркли
Berkeley
Оринда
Orinda
Крик
Walnut Creek
Байрон
Byron
Клис
Cli
Мьюир Бич
Muir Beach
Сосалито
Sausalito
80
Пидмонт
Piedmont
Морага
Moraga
Данвилл
Danville
Дьябло
Diablo
Окленд
Oakland
Сан-Франциско
San Francisco
Блэкхок-Камино
Тассаджара
Blackhawk-Camino
Tassajara
Сан-Рамон
San Ramon
580
Аламида
Alameda
880
Сан
Леандро
San Leandro
1
280
680
Дейли Сити
Daly City
Дублин
Dublin
580
Кастро
Валли
Castro Valley
Фэрвью
Fairview
Бродмур
Broadmoor
Брисбен
Brisbane
Колма
Colma
Сан-Франциско
San Francisco Bay
Плезантон
Pleasanton
Ливермор
Livermore
Хейуорд
Hayward
Юг Сан
Франциско
South San
Francisco
Пасифика
Pacifica
Юнион
Сити
Union City
Санол
Sunol
Государственная
зона отдыха
Лейк Дель Валле
Lake Del Valle State
Recreation Area
Зона отдыха
Карнеги Стей
Вехикулар
Carnegie State
Vehicular
Recreation Area
Милбро
Millbrae
Сан-Матео
San Mateo
Фостер
Сити
Foster City
880
680
Фремонт
Fremont
Монтара
Montara
Олон Риджинал
Уайлдернесс
Ohlone Regional
Wilderness
Эль
Гранада
El Granada
Мосс Бич
Moss Beach
280
84
Ньюарк
Newark
Сан Карлос
San Carlos
Редвуд
Сити
Redwood City
Санол Риджинал
Уайлдернесс
Sunol Regional
Wilderness
Эмералд
Лейк Хилс
Emerald
Lake Hills
Менло Парк
Menlo Park
Халф
Мун Бэй
Half Moon Bay
Станфорд
Stanford
Пало-Альто
Palo Alto
Милпитас
Milpitas
680
Восток
Футхилс
East Foothills
Маунтин-Вью
Mountain View
Саннивейл
Sunnyvale
880
Портола
Валли
Portola Valley
Лос Альтос
Хилс
Los Altos Hills
Лойола
Loyola
Сан-Хосе
San Jose
Алум Рок
Alum Rock
Санта
Клара
Santa Clara
Joseph D. Grant
County Park
Ла Хонда
La Honda
Купертино
Cupertino
Бербанк
Burbank
Сан
Грегорио
San Gregorio
Парк округа
Пескадеро Крик
Pescadero Creek
County Park
Саратога
Saratoga
Кэмпбелл
Campbell
Севен Трис
Seven Trees
Лома Мар
Loma Mar
Пескадеро
Pescadero
Парк штата
Касл Рок
Лос Гатос
Кэмбриан
Парк
85



n Beach
Mill Valley
ソーサリト
Sausalito
バークリー
Berkeley
オリンダ
Orinda
Walnut Creek
州立公園
Mt Diablo
State Park
Byron
ミュアー
ビーチ
Muir Beach
エメリー
ビル
Emeryville
ピードモント
Piedmont
モラガ
Moraga
ダンビル
Danville
ディアブロ
Diablo
クリフトン
コート
フォアベイ
Clifton Cou
Forebay
80
サンフランシスコ
San Francisco
オークランド
Oakland
580
サン
ラモン
San Ramon
680
アラメダ
Alameda
サン
レアンドロ
San Leandro
デーリー
シティ
Daly City
280
101
グリーン
ビル
Greenville
580
ダブリン
Dublin
カストロ
バレー
Castro Valley
580
リバモア
Livermore
ブロードムア
Broadmoor
コルマ
Colma
サン
ロレンゾ
San Lorenzo
プレザントン
Pleasanton
ヘイワード
Hayward
レイク・デル
ヴァジェ
州立保養地
Lake Del
Valle State
Recreation Area
ラッセル
シティ
Russell City
キルケア
ウッズ
Kilkare
Woods
サウス
サンフランシスコ
South San Francisco
ユニオン
シティ
Union City
パシフィカ
Pacifica
スノール
Sunol
カーネギー
ステート
ビーキュラー保養
Carnegie State
Vehicular
Recreation Area
サン
マテオ
San Mateo
フォスター
シティ
Foster City
フリー
モント
Fremont
680
オーロン
リージョナル
ワイルダーネス
Ohlone Regional
Wilderness
バーリン
ゲーム
Burlingame
モス
ビーチ
Moss Beach
エル
グラナダ
El Granada
サン
カルロス
San Carlos
ニュー
アーク
Newark
880
280
レッドウッド
シティ
Redwood City
スノール
リージョナル
ワイルダーネス
Sunol Regional
Wilderness
ハーフ
ムーンベイ
Half Moon Bay
エメラルド
レイク・ヒルズ
Emerald
Lake Hills
メンロー
パーク
Menlo Park
パロ
アルト
Palo Alto
ミルピタス
Milpitas
ウッド
サイド
Woodside
マウンテン
ビュー
Mountain View
サニー
ベール
Sunnyvale
880
イースト
フットヒルズ
East Foothills
ジョゼフ D.
グラント
郡立公園
Joseph D. Grant
County Park
ポートラ
バレー
Portola Valley
ロス・アルトス
ヒルズ
Los Altos Hills
サンノゼ
San Jose
アラム
ロック
Alum Rock
サンタ
クララ
Santa Clara
ラ・ホンダ
La Honda
クパチーノ
Cupertino
バーバンク
Burbank
サン
グレゴリオ
San Gregorio
ローマ
マー
Loma Mar
キャンベル
Campbell
セブン
ツリーズ
Seven Trees
ペスカデロ
Pescadero
ペスカデロ
クリーク
郡立公園
サラトガ
Saratoga
ロス
ガトス
カンブリアン
パーク


Berkeley
奥林达
Orinda
Lafayette
Walnut Creek
洛州立公园
Mt Diablo
State Park
拜伦
Byron
缪尔比奇
Muir Beach
索萨利托
Sausalito
蒂伯龙
Tiburon
埃默里维尔
Emeryville
皮德蒙特
Piedmont
莫拉加
Moraga
阿拉莫
Alamo
丹维尔
Danville
迪亚伯洛
Diablo
Clifton Cou
Forebay
80
旧金山
San
Francisco
奥克兰
Oakland
圣拉蒙
San Ramon
查沃特
地区公园
Chabot
Regional Park
101
阿拉梅达
Alameda
580
680
圣莱安德罗
San Leandro
280
戴利城
Daly City
圣弗朗西斯
科湾
San
Francisco Bay
880
都柏林
Dublin
格林维尔
Greenville
卡斯特
罗瓦利
Castro Valley
580
利弗莫尔
Livermore
圣洛伦索
San Lorenzo
布罗德莫尔
Broadmoor
南旧金山
South San
Francisco
海沃德
Hayward
普莱森顿
Pleasanton
圣布鲁诺
San Bruno
帕西菲卡
Pacifica
尤宁城
Union City
苏尼奥尔
Sunol
卡内基斯塔特
韦卡勒游乐区
Carnegie State
Vehicular
Recreation Area
伯灵格姆
Burlingame
圣马特奥
San Mateo
莱克戴尔瓦勒
州立游乐区
Lake Del
Valle State
Recreation Area
苏尼奥尔
地区荒野
Sunol Regional
Wilderness
弗里蒙特
Fremont
680
蒙塔拉
Montara
福斯特城
Foster City
84
纽瓦克
Newark
埃尔格拉
纳达
El Granada
莫斯比奇
Moss Beach
280
圣卡洛斯
San Carlos
奥隆地区荒野
Ohlone Regional
Wilderness
雷德伍德城
Redwood City
东帕洛
阿尔托
East Palo Alto
埃默拉尔德
莱克希尔斯
Emerald
Lake Hills
半月湾
Half
Moon Bay
门洛帕克
Menlo Park
帕洛阿尔托
Palo Alto
米尔皮塔斯
Milpitas
斯坦福
Stanford
El Corte De
Madera Creek
Open Space
山景城
Mountain
View
880
东富特
希尔斯
East Foothills
洛斯阿尔
托斯
Los Altos
森尼韦尔
Sunnyvale
波托拉谷
Portola
Valley
洛约拉
Loyola
约瑟夫D.格兰特
郡立公园
Joseph D. Grant
County Park
圣格雷
戈里奥
San Gregorio
圣荷西
San Jose
奥卢姆罗克
Alum Rock
圣克拉拉
Santa Clara
拉翁达
La Honda
280
库佩蒂诺
Cupertino
伯班克
Burbank
塞文特雷斯
Seven Trees
丘陵马尔
Loma Mar
坎贝尔
Campbell
佩斯卡德罗
Pescadero
佩斯卡德罗克
里克郡立公园
Pescadero Creek
萨拉托加
Saratoga
坎布里安
帕尔克
Cambrian Park
洛斯加托斯
Los Gatos
85
101

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

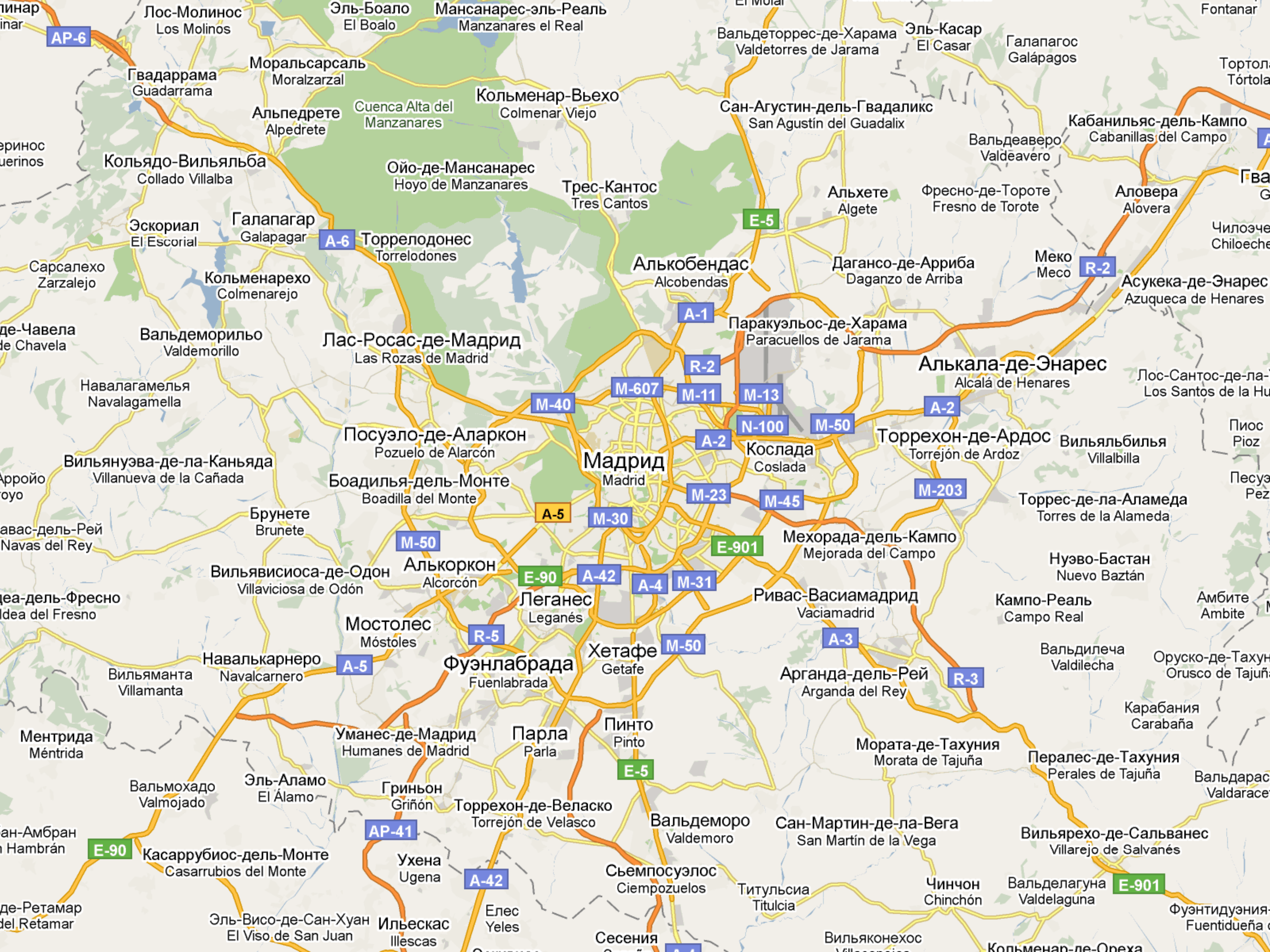
Lesson #1: Figuring out what to transliterate and how is typically much harder than implementing the transliteration scheme.

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


Лос-Молинос
Los Molinos
Эль-Боало
El Boalo
Мансанарес-эль-Реаль
Manzanares el Real
El Molar
Fontanar
AP-6
Вальдеторрес-де-Харама
Valdetorres de Jarama
Эль-Касар
El Casar
Галапагос
Galápagos
Моральсарсаль
Moralzarzal
Гвадаррама
Guadarrama
Cuenca Alta del Manzanares
Кольменар-Вьехо
Colmenar Viejo
Сан-Агустин-дель-Гвадаликс
San Agustín del Guadalix
Альпедрете
Alpedrete
Кабанильяс-дель-Кампо
Cabanillas del Campo
Вальдеаверо
Valdeavero
Кольядо-Вильяльба
Collado Villalba
Ойо-де-Мансанарес
Hoyo de Manzanares
Трес-Кантос
Tres Cantos
Альхете
Algete
Фресно-де-Торote
Fresno de Torote
Аловера
Alovera
E-5
Эскориал
El Escorial
Галапагар
Galapagar
A-6
Торрелодонес
Torrelodones
Меко
Meco
R-2
Сарсалехо
Zarzalejo
Кольменарехо
Colmenarejo
Алькобендас
Alcobendas
Дагансо-де-Арриба
Daganzo de Arriba
Асукека-де-Энарес
Azuqueca de Henares
A-1
Парaкуэльос-де-Харама
Paracuellos de Jarama
Вальдеморильо
Valdemorillo
Лас-Росас-де-Мадрид
Las Rozas de Madrid
R-2
Алькала-де-Энарес
Alcalá de Henares
Навалагамелья
Navalagamella
M-607
M-11
M-13
M-40
A-2
Лос-Сантос-де-ла
Los Santos de la Hu
N-100
M-50
Посуэло-де-Аларкон
Pozuelo de Alarcón
A-2
Торрехон-де-Ардос
Torrejón de Ardoz
Вильяльбилья
Villalbilla
Пиос
Pioz
Вильянуэва-де-ла-Каньяда
Villanueva de la Cañada
Мадрид
Madrid
Кослада
Coslada
Боадилья-дель-Монте
Boadilla del Monte
M-23
M-45
M-203
Торрес-де-ла-Аламеда
Torres de la Alameda
Брунете
Brunete
A-5
M-30
M-50
E-901
Мехорада-дель-Кампо
Mejorada del Campo
Нуэво-Бастан
Nuevo Baztán
Вильявисиоса-де-Одон
Villaviciosa de Odón
Алькоркон
Alcorcón
E-90
A-42
A-4
M-31
Леганес
Leganés
Ривас-Васиамадрид
Vaciamadrid
Кампо-Реаль
Campo Real
Амбите
Ambite
Мостолес
Móstoles
R-5
A-3
Хетафе
Getafe
M-50
Навалькарнеро
Navalcarnero
A-5
Фуэнлабрада
Fuenlabrada
Вильяманта
Villamanta
Вальдилеча
Valdilecha
Оруско-де-Тахуня
Orusco de Tajuña
Арганда-дель-Рей
Arganda del Rey
R-3
Карабания
Carabaña
Пинто
Pinto
Ментрида
Méntrida
Уманес-де-Мадрид
Humanes de Madrid
Парла
Parla
Мората-де-Тахуния
Morata de Tajuña
Пeралес-де-Тахуния
Perales de Tajuña
E-5
Вальмохадо
Valmojado
Эль-Аламо
El Álamo
Гриньон
Griñón
Торрехон-де-Веласко
Torrejón de Velasco
Вальдеморо
Valdemoro
Сан-Мартин-де-ла-Вега
San Martín de la Vega
Вильярехо-де-Сальванес
Villarejo de Salvanés
AP-41
E-90
Касаррубиос-дель-Монте
Casarrubios del Monte
Ухена
Ugena
A-42
Сьемпосуэлос
Ciempozuelos
Титульсиа
Titulcia
Чинчон
Chinchón
Вальделагуна
Valdelaguna
E-901
Фуэнтидуэния
Fuentidueña
Елес
Yeles
Эль-Висо-де-Сан-Хуан
El Viso de San Juan
Ильескас
Illescas
Сесения
Вильяконехос
Кольменар-де-Орexa

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:

{G, J, X}iménez	[ximeneθ]	Хименес
Gijón	[xixon]	Хихон
Jerez (Xeres)	[xeɾeθ]	Херес
Cuenca	[kweŋka]	Куэнка
Quesada	[kesaða]	Кесада
Quintana	[kintana]	Кинтана
Ceuta	[θewta]	Сеута
Zaragoza	[θaɾagoθa]	Сарагоса
La Mancha	[la mantʃa]	Ла-Манча
Huelva	[welβa]	Уэльва

Spanish-to-Russian (continued)

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

Griñón	[gɾiɲon]	Гриньон
Logroño	[loɣɾoɲo]	Логроньо
Llobregat	[ʎoβɾeɣat]	Льобрегат
Sagra	[saɣɾa]	Сагра
Zagra	[θaɣɾa]	Сагра
Ascensión	[asθensjon]	Асенсьон
Laredo	[laɾeðo]	Ларедо
Casalarreina	[kasalarːejna]	Касаларрейна
Casa la Reina	[kasalarːejna]	Каса-ла-Рейна

Spanish-to-Russian (continued)

Even though Spanish b and v are pronounced exactly the same, the orthographic distinction is retained in Russian:

Bargas [baɾɣas]	Баргас
Vargas [baɾɣas]	Варгас
Ribera [rːiβeɾa]	Рибера
Rivera [rːiβeɾa]	Ривера
Exception:	
Córdoba [koɾðoβa]	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.

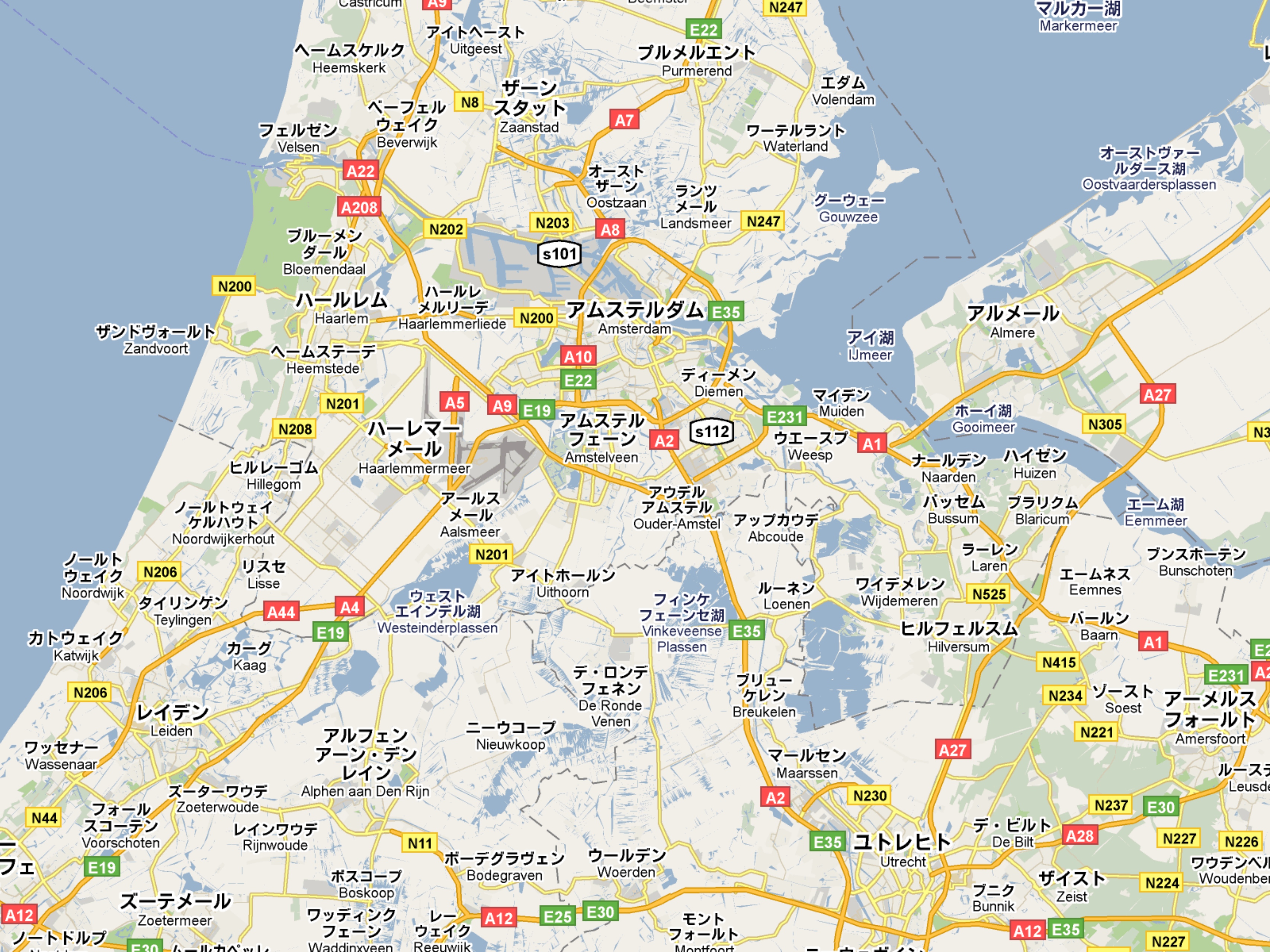
Spanish-to-Russian ICU rules

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

```
         b         → б ;
         ch        → ч ;
         c } [ei]  → с ;  # before ’e’ or ’i’
         c         → к ;
         d         → д ;
[-\ $] { e         → э ;  # at beginning of word
         e         → е ;
         f         → ф ;
         gu } [ei] → г ;
         g } [ei]  → х ;
         g         → г ;
         h         →   ;  # empty string
```

Proper Name Transliteration with ICU Transforms

Target: Japanese


Castricum
A9
Beemster
N247
マルカー湖
Markermeer
アイトヘースト
Uitgeest
E22
ヘームスケルク
Heemskerk
プルメルエント
Purmerend
エダム
Volendam
ザーン
スタット
Zaanstad
ベーフェル
ウェイク
Beverwijk
N8
A7
フェルゼン
Velsen
ワーテルラント
Waterland
A22
オーストヴァー
ルダース湖
Oostvaardersplassen
オースト
ザーン
Oostzaan
ランツ
メール
Landsmeer
A208
グーウェー
Gouwzee
N202
N203
A8
N247
ブルーメン
ダール
Bloemendaal
s101
N200
ハールレ
メルリーデ
Haarlemmerliede
ハールレム
Haarlem
N200
アムステルダム
Amsterdam
E35
アルメール
Almere
ザンドヴォールト
Zandvoort
アイ湖
IJmeer
ヘームステーデ
Heemstede
A10
E22
ディーメン
Diemen
A27
マイデン
Muiden
N201
A5
A9
E19
ホーイ湖
Gooimeer
E231
N305
N208
ハーレマー
メール
Haarlemmermeer
アムステル
フェーン
Amstelveen
A2
s112
ウエースプ
Weesp
A1
N3
ナールデン
Naarden
ハイゼン
Huizen
ヒルレーゴム
Hillegom
アウデル
アムステル
Ouder-Amstel
アールス
メール
Aalsmeer
バッセム
Bussum
ブラリクム
Blaricum
エーム湖
Eemmeer
ノールトウェイ
ケルハウト
Noordwijkerhout
アップカウデ
Abcoude
ラーレン
Laren
N201
ブンスホーテン
Bunschoten
ノールト
ウェイク
Noordwijk
N206
リスセ
Lisse
アイトホールン
Uithoorn
エームネス
Eemnes
ルーネン
Loenen
ワイデメレン
Wijdemeren
N525
タイリンゲン
Teylingen
A44
A4
ウェスト
エインデル湖
Westeinderplassen
フィンケ
フェーンセ湖
Vinkeveense
Plassen
E35
ヒルフェルスム
Hilversum
バールン
Baarn
カトウェイク
Katwijk
E19
A1
E2
カーグ
Kaag
N415
E231
A2
N206
デ・ロンデ
フェネン
De Ronde
Venen
ブリュー
ケレン
Breukelen
N234
ゾースト
Soest
アーメルス
フォールト
Amersfoort
レイデン
Leiden
ワッセナー
Wassenaar
アルフェン
アーン・デン
レイン
Alphen aan Den Rijn
ニーウコープ
Nieuwkoop
N221
A27
マールセン
Maarssen
ルースデ
Leusde
ズーターワウデ
Zoeterwoude
A2
N230
N237
E30
N44
フォール
スコーテン
Voorschoten
レインワウデ
Rijnwoude
N11
E35
ユトレヒト
Utrecht
デ・ビルト
De Bilt
A28
N227
N226
フェ
E19
ボーデグラヴェン
Bodegraven
ウールデン
Woerden
ワウデンベル
Woudenber
ボスコープ
Boskoop
ザイスト
Zeist
ブニク
Bunnik
N224
ズーテメール
Zoetermeer
A12
ワッディンク
フェーン
Waddinxveen
レー
ウェイク
Reeuwijk
A12
E25
E30
モント
フォールト
Montfoort
A12
E35
N227
ノートドルプ
E30
ムールカペッレ
ニーウェガイン

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:

{G, J, X}iménez	[ximeneθ]	ヒメネス	himenesu
Gijón	[xixon]	ヒホン	hihon
Cuenca	[kweŋka]	クエンカ	kuenka
Quesada	[kesaða]	ケサダ	kesada
Quintana	[kintana]	キンタナ	kintana
Ceuta	[θewta]	セウタ	seuta
Zaragoza	[θaɾagoθa]	サラゴサ	saragosa
La Mancha	[la mantʃa]	ラ・マンチャ	ra mancha
Huelva	[welβa]	ウエルバ	ueruba

Spanish-to-Japanese (continued)

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

Griñón [gɾiɲon]	グリニョン	gurinyon
Logroño [loɣɾoɲo]	ログロニョ	roguronyo
Llobregat [ʎoβɾeɣat]	リョブレガト	ryoburegato
Sagra [saɣɾa]	サグラ	sagura
Zagra [θaɣɾa]	サグラ	sagura

Spanish-to-Japanese (conclusion)

Some aspects of transliteration are governed by Japanese extended phonology:

Santiago	サンティアゴ	santiago
not	サンチアゴ	sanchiago
Alcorcón	アルコルコン	arukorukon
Chimbote	チンボテ	chinbote
Motril	モトリル	motoriru
Rodríguez	ロドリゲス	rodorigesu
Ciudad [θjuðað]	シウダー	shiudā
closer than	シウダド	shiudado

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-SpanishPhonemic 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 ;
rː → ɾ ;
l  → ɾ ;
⋮
:: Null;  # Second pass: Phonemes to Katakana.
a   → ア ;
ba  → バ ;
bi  → ビ ;
bu  → ブ ;
be  → ベ ;
bo  → ボ ;
b   → ブ ;
⋮
```

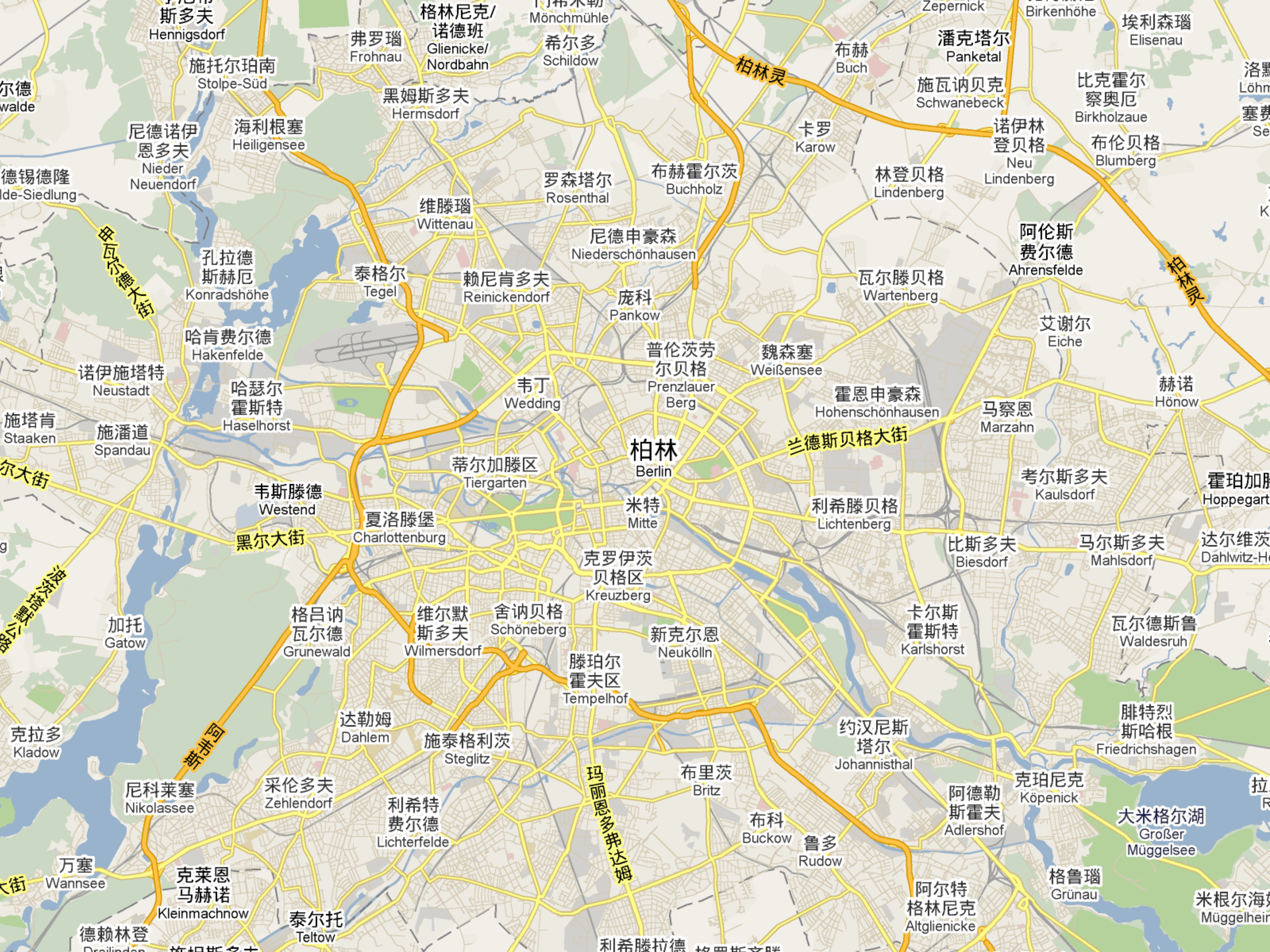
Summary: Japanese

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


斯多夫
Hennigsdorf
弗罗瑙
Frohnau
格林尼克/
诺德班
Glienicke/
Nordbahn
Mönchmühle
希尔多
Schildow
Zepernick
Birkenhöhe
埃利森瑙
Elisenau
布赫
Buch
潘克塔尔
Panketal
施托尔珀南
Stolpe-Süd
柏林灵
施瓦纳贝克
Schwanebeck
比克霍尔
察奥厄
Birkholzaue
洛默
Löhm
塞费
Se
尔德
walde
黑姆斯多夫
Hermsdorf
海利根塞
Heiligensee
尼德诺伊
恩多夫
Nieder
Neuendorf
卡罗
Karow
诺伊林
登贝格
Neu
Lindenberg
布伦贝格
Blumberg
德锡德隆
lde-Siedlung
布赫霍尔茨
Buchholz
林登贝格
Lindenberg
罗森塔尔
Rosenthal
维滕瑙
Wittenau
中瓦尔德大街
尼德申豪森
Niederschönhausen
阿伦斯
费尔德
Ahrensfelde
柏林灵
孔拉德
斯赫厄
Konradshöhe
泰格尔
Tegel
赖尼肯多夫
Reinickendorf
庞科
Pankow
瓦尔滕贝格
Wartenberg
艾谢尔
Eiche
哈肯费尔德
Hakenfelde
普伦茨劳
尔贝格
Prenzlauer
Berg
魏森塞
Weißensee
诺伊施塔特
Neustadt
哈瑟尔
霍斯特
Haselhorst
韦丁
Wedding
霍恩申豪森
Hohenschönhausen
赫诺
Hönow
马察恩
Marzahn
施塔肯
Staaken
施潘道
Spandau
柏林
Berlin
兰德斯贝格大街
尔大街
蒂尔加滕区
Tiergarten
考尔斯多夫
Kaulsdorf
霍珀加
Hoppegart
韦斯滕德
Westend
米特
Mitte
利希滕贝格
Lichtenberg
夏洛滕堡
Charlottenburg
黑尔大街
比斯多夫
Biesdorf
马尔斯多夫
Mahlsdorf
达尔维茨
Dahlwitz-H
克罗伊茨
贝格区
Kreuzberg
阿弗斯高速公路
加托
Gatow
格吕讷
瓦尔德
Grunewald
维尔默
斯多夫
Wilmersdorf
舍讷贝格
Schöneberg
新克尔恩
Neukölln
卡尔斯
霍斯特
Karlshorst
瓦尔德斯鲁
Waldesruh
滕珀尔
霍夫区
Tempelhof
腓特烈
斯哈根
Friedrichshagen
克拉多
Kladow
阿韦斯
达勒姆
Dahlem
施泰格利茨
Steglitz
约汉尼斯
塔尔
Johannisthal
玛丽恩多弗达姆
布里茨
Britz
克珀尼克
Köpenick
拉
R
尼科莱塞
Nikolassee
采伦多夫
Zehlendorf
利希特
费尔德
Lichterfelde
阿德勒
斯霍夫
Adlershof
大米格尔湖
Großer
Müggelsee
布科
Buckow
鲁多
Rudow
万塞
Wannsee
大街
克莱恩
马赫诺
Kleinmachnow
格鲁瑙
Grünau
阿尔特
格林尼克
Altglienicke
米根尔海
Müggelhei
泰尔托
Teltow
德赖林登
Dreilinden
利希滕拉德

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

Transliteration into Mandarin (contd.)

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):

	∅-	b-	p-	d-/ð-	t-	g-/ɣ-	k-
-∅		布 bù	普 pǔ	德 dé	特 tè	格 gé	
-a	阿 ā	巴 bā	帕 pà	达 dá	塔 tǎ	加 jiā	卡 kǎ
-e	埃 āi	贝 bèi	佩 pèi	德 dé	特 tè	格 gé	克 kè
-ej	埃 āi	贝 bèi	佩 pèi	代 dài	泰 tài	盖 gài	凯 kǎi
-i	伊 yī	比 bǐ	皮 pí	迪 dí	蒂 dì	吉 jí	基 jī
-o	奥 ào	博 bó	波 bō	多 duō	托 tuō	戈 gē	科 kē
-ow	欧 ōu	博 bó	波 bō	多 duō	托 tuō	戈 gē	科 kē
-u	乌 wū	布 bù	普 pǔ	杜 dù	图 tú	古 gǔ	库 kù

Example: Pico del Teide → [tej.ðe] → 泰德峰

Spanish-to-Mandarin peculiarities (I)

The transliteration scheme always collapses the [l] / [ɾ] / [rː] distinction phonetically, but maintains it sometimes in the orthography:

Logroño	[loɣɾoɲo]	洛格罗尼奥	luò gé luó ní ào
Malón	[malon]	马隆	mǎ lóng
marrón	[marːon]	马龙	mǎ lóng

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 transcibe the [ʒ] sound, e.g. in words of French origin.)

Spanish-to-Mandarin peculiarities (II)

The transliteration scheme distinguishes [b] from [β], but conflates [d] and [ð] (as well as [g] and [ɣ]):

Valencia	[balenθja]	巴伦西亚	bā lún xī yà
Córdoba	[koɾðoβa]	科尔多瓦	kē ěr duō wǎ
Mondoñedo	[mondoɲeðo]	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:

```
ʎ → ʝ ;
θ → 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.

Proper Name Transliteration with ICU Transforms

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.

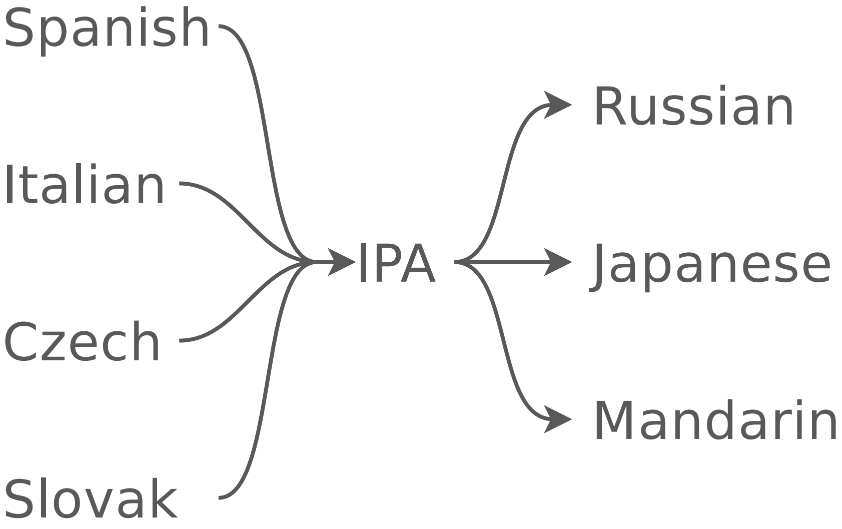
Pronunciation rules/models are highly reusable, since they represent an abstract truth about the source language, with no reference to transliteration.

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):

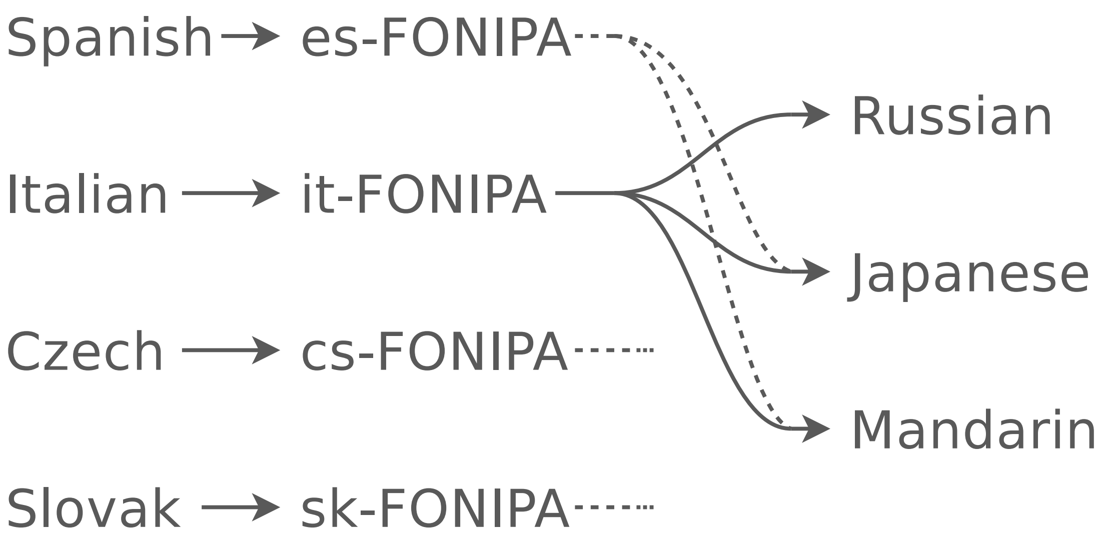


No interlingua for transliteration

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.

		Russian	**Japanese**		**Mandarin**
German	Schönau	Шёнау	シェーナウ	Busch	布施
	[ʃønaʊ]	Šënau	shēnau	[bʊʃ]	bùshī
French	Dreux	Дрё	ドルー	Franche	弗朗什
	[dʁø]	Drë	dorū	[fʁãʃ]	fúlǎngshí

The cross-product reality



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.

Not all is lost

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