SOME LINGUISTIC ISSUES IN THE MACHINE TRANSLITERATION OF CHINESE, JAPANESE, AND ARABIC NAMES

Keynote address

6th NEWS Named Entities Workshop, Berlin, 2016

Jack Halpern

CEO, The CJK Dictionary Institute, Niiza, Japan

Summary

Non-Latin scripts:

- Arabic
- Chinese
- Japanese

Clear up differences between:

- transliteration
- transcription
- romanization

Focus on:

- special characteristics of the CJA scripts that impact MT
- role played by lexical resources such as personal name dictionaries

Transcription and Transliteration

"Never the Twain Shall Meet"

Transliteration: representing the source script letters (graphemes not phonemes) with the characters of another script

Transcription: representing the source script of a language in the target script in a manner that reflects pronunciation

محمد > \mHmd\

1. Phonetic transcription represents the actual speech sounds.

محمد > [muħɛmm ¨ɗ̪] (IPA)

2. Phonemic transcription represents the phonemes of the source.

محمد > /muHammad/

3. Popular transcription roughly represents pronunciation.

محمد > Mohammed, Muhammad, Moohammad, Moohamad…+200

Arabic Transcription and Romanization

Orthographic Ambiguity

1. Short vowel omission (كاتب\kAtb\)
2. Short vowel Representation (جامعة /jaami`a/)
3. Multiple long /aa/
   'alif Tawiila (سوريا ا)
   'alif maduuda (آسيا آ)
   'alif maqSuura (الوسطى آسيا ى.)
4. Long vowel omission (هدا /haadha/)
5. Long /aa/ ambiguity (شكرا, انا) (/an/, /a/)
6. Otiose alif is silent كتبوا > /katabuu/
7. Omission of shadda (مُحَمَّد > محمد) /Muhammad/
8. Omission of tanwiin شكرا \$ukrAF\ (شُكْرًا)
9. Complex hamza rules (فوكوأوكا vs فوكوؤوكا)
10. Hamza omission (سائيتاما > سايتاما)
11. Phonological alternation (الرجل الطويل)
    /'alrajulu alTawiilu/> /'arrajulu-Ttawiilu >
12. Shortening long vowels (القاهرة في)
    /fii-lqaahira/ > /fi-lqaahira/

Variants and Errors for "Alexandria"

V=variant E=error S=Standard N=normalized

Rank	Type	Arabic	Buckwalter	Google Hits	Remarks
1	N	الاسكندرية	AlAskndryp	2930000	Normalized, no hamza
2	S	الإسكندرية	Al<skndryp	690000	Standard form, with hamza
3	E	الاسكندريه	AlAskndryh	89200	No hamza, taa marbuta replaced by haa
4	V	الإسكندريّة	Al<skndry~p	954	Explicit shadda
5	E	الإسكندريه	Al<skndryh	897	taa marbuta replaced by haa
6	V	الاسكندريّة	AlAskndry~p	245	no hamza, shadda explicit
7	E	الاسكندريا	AlAskndryA	80	hamza omitted, taa marbuuta replaced by alif
8	V	الإسْكَنْدَرِيَّة	Al<sokanodary~ap	24	fully vocalized
9	E	الاسكندريّه	AlAskndry~h	12	no hamza, shadda explicit, taa marbuta replaced by haa
10	E	الإسكندريا	Al<skndryA	7	taa marbuta replaced by alif tawiila
11	E	الإسكندريّه	Al<skndry~h	5	taa marbuuta replaced by haa, shadda explicit

Major Arabic Romanization Systems

Example: شولوخ

Description	Example	System
Romanization standard of the American Library Association-Library of Congress.	shwlwkh	ALC-LC
This refers to DIN 31635, the DIN standard for Arabic transliteration.	šūlūḫ	DIN
International Phonetic Alphabet, a scientific system of uniquely and accurately representing speech sounds.	ʃuːluːx	IPA
One of many (at least 10) possible popular transcriptions.	Shoulokh	English
A strict transliteration system widely used in information processing.	$wlwx	Buckwalter

Variation in Arabic Names

Remarks	Error	Variant	English	Buckwalter	Standard
V: omit hamza E: alif maqsura replaces yaa'	أبو ظبى	ابو ظبي	Abu Dhabi	>bw Zby	أبو ظبي
V: omit hamza E: haa' replaces taa' marbuuTa	الإسكندريه	الاسكندرية	Alexandria	Al<skndryp	الإسكندرية
V: explicit shadda E: haa' replaces taa' marbuuTa	جده	جدّة	Jeddah	jdp	جدة
V: omit hamza		الاردن	Jordan	Al>rdn	الأردن
V1: omit hamza V2: madda replaces hamza		بالو التو بالو آلتو	Palo Alto	bAlw >ltw	بالو ألتو
V: explicit shadda		الرّياض	Riyadh	AlryAD	الرياض
E: taa' replaces Taa'			Tokyo	Twkyw	طوكيو

Database of Arab Names (DAN)

Variants of عبدالرحيم

Sub ID	Variants	Frequency
U000261	Abderrahim	0000382000
U000763	Abderrahim	0000382000
U000425	Abdurrahim	0000172000
U000928	Abdurrahim	0000172000
U000385	Abdulrahim	0000082100
U000887	Abdulrahim	0000082100
U000236	Abdelrahim	0000054200
U000739	Abdelrahim	0000054200
U000359	Abdul Rahim	0000040000

DAN includes 1100-plus variants of the popular name عبدالرحيم *'Abd Al Raheem*

Diphthong Ambiguity for 福井 /fu-ku-i/

Buckwalter	Google hits	Arabic	No.
fwkw}y	468	فوكوئي	1
fwkw}	9	فوكوئ	2
Fwkwy	1950	فوكوي	3
Fwkwyy	335	فوكويي	4

Long and Short Vowels

Arab3	Arab2	Arab1	Phonemic	Kana	Kanji	No.
		أوتا	oota	おおた	太田	1
		فوما	fuuma	ふうま	風馬	2
	كيكو	كييكو	keiko	けいこ	敬子	3
		كونو	kuuno	くうの	空野	4
		كونو	kuno	くの	久野	5
هيئيدا	هيئدا	هييدا	hieda	ひえだ	日枝	6
يوشيئي	يوشيئه	يوشيي	yoshie	よしえ	芳江	7

Ambiguity of كاتب kaatib

Placeholder

file:///E:/desk/conferen/caasl2/presentation/extended_sample.htm

Japanese Orthographic Variation

Database of Japanese Name Variants

Romanized Variants for 純一郎 (Junichiro)

Type	Romanization	Rank
Variant	Junichiro	A
Eng	Jun'ichiro	A
Variant	Jun-ichiro	A
Variant	Junichirô	A
Hybrid	Juniciro	A
Hepburn	Jun'ichirō	A
Variant	Jun-ichirō	A
Variant	Junichirou	B
Variant	Jun'ichirou	B
Variant	Jun-ichirou	B
Variant	Junichirō	B

Top 11 variants of "Junichiro", out of 169 total.

「幕張国際展示場」の方式別翻訳例

Type	English	Simplified Chinese	Traditional Chinese	Comments
字訳	–	幕张国际展示场	幕張國際展示場	日本語表記を文字対応レベルで変換
音訳	Makuhari kokusai tenjijo	–	–	日本語読みの音通りに表記
意訳	Makuhari International Exhibition Area	幕张国际展览馆	幕張國際展覽館	地名等の固有名以外、全て意訳
意音訳	Makuhari Kokusai Exhibition Area	–	–	主要な構成要素部分を意訳、それ以外は音訳
翻訳	International Exhibition Halls, Makuhari Messe	幕张国际展览中心	幕張國際展覽中心	施設自体が公式に定めた名称

店舗・建物・ビル名等の翻訳レベルについて

例「東京中央ゴルフ場」の訳語を定める場合			
	翻訳レベル	**翻訳例**	**解説**
1	音訳１	Tokyo Chuo Gorufujo	完全な音訳。読みさえあれば、ほぼ自動処理が可能。
2	音訳２	Tokyo Chuo Golf-jo	音訳＋外来語由来のカタカナのみ正しいスペルを調査。語源不明の造語等は音訳で処理する場合もある。
3	意音訳	Tokyo Chuo Golf Course	ゴルフ場 ="Golf Course" 、病院 ="Hospital" 、デパート ="Department Store" のように大まかな訳語を定めておき、それ以外の部分を音訳する。
4	意訳	Tokyo Central Golf Course	固有名以外の部分を全て意訳する。
5	翻訳	The Central Golf Club, Tokyo	ウェブサイトや電話調査で、個々の正式訳を調査する。現地での表示と一致するので最も望ましい形だが、作成の時間・経費が嵩む。

Okurigana Variants

Headword	Reading	Normalized
書き著す	かきあらわす	書き著す
書き著わす	かきあらわす	書き著す
書著す	かきあらわす	書き著す
書著わす	かきあらわす	書き著す

Variants of *toriatsukai*

toriatsukai	Type of Variant
取り扱い	"standard" form
取扱い	okurigana variant
取扱	All kanji
とり扱い	replace kanji with hiragana
取りあつかい	replace kanji with hiragana
とりあつかい	All hiragana

Cross-Script Variants

Type	Example
Kanji vs. Hiragana	大勢　おおぜい
Kanji vs. Katakana	硫黄　イオウ
Kanji vs. hiragana vs. katakana	猫　ねこ　ネコ
Katakana vs. hybrid	ワイシャツ　Ｙシャツ
Kanji vs. katakana vs. hybrid	皮膚　ヒフ　皮フ
Kanji vs. hybrid	彗星　すい星
Hiragana vs. katakana	ぴかぴか　ピカピカ

Chinese-to-Chinese Cross-Script Conversion

The Three Conversion Levels

Level 1	Code	Character-to-character, code-based substitution
Level 2	Orthographic	Word-to-word, character-based conversion
Level 3	Lexemic	Word-to-word, lexicon-based conversion

Code Conversion

SC	TC1	TC2	TC3	TC4	Remarks
门	們				one-to-one
汤	湯				one-to-one
发	發	髮			one-to-many
暗	暗	闇			one-to-many
干	幹	乾	干	榦	one-to-many

Orthographic Conversion

English	SC	TC1	TC2	Incorrect	Comments
telephone	电话	電話			unambiguous
we	我们	我們			unambiguous
start-off	出发	出發		出髮 齣髮 齣發	one-to-many
dry	干燥	乾燥		干燥 幹燥 榦燥	one-to-many
	阴干	陰乾	陰干		depends on context

Lexemic Conversion

English	SC	Taiwan TC	Hong Kong TC	Incorrect TC
Software	软件	軟體	軟件	軟件
File	文件	檔案	檔案	文件
Program	程序	程式	程式	程序
Taxi	出租汽车	計程車	的士	出租汽車
Osama Bin Laden	奥萨马 本拉登	奧薩瑪 賓拉登	奧薩瑪 賓拉丹	奧薩馬 本拉登
Kennedy	肯尼迪	甘迺迪	堅尼地	肯尼迪

Chinese

Category	Variants	Surname 艾	Surname 单	Given Name 业经	Given Name 爱博
	Simplified Chinese	艾	单	业经	爱博
Simplified Chinese	Toned Pinyin	ài	shàn	yèjīng	àibó
	Numbered Pinyin	ai4	shan4	ye4-jing1	ai4-bo2
	Wade-Giles	ai	shan	yehching	aipo
	Yale System	ai	shan	yejing	aibwo
	Tongyong	ai	shan	yejing	aibo
Traditional Chinese	Traditional Chinese	艾	單	業經	愛博
	Zhuyin	ㄞˋ	ㄕㄢˋ	ㄧㄝˋ ㄐㄧㄥ	ㄞˋ ㄅㄛˊ

Chinese Dialects

Category	Variants	Surname 艾	Surname 单	Given Name 业经	Given Name 爱博
Simplified Chinese		艾	单	业经	爱博
Cantonese	LAU1	ngaai6	daan1	yip6-ging1	ngoi3-bok3
	LAU2	ngaai	daan	yipging	ngoibok
	YALE1	ngaai6	daan1	yip6-ging1	ngoi3-bok3
	YALE2	ngaai	daan	yipging	ngoibok
	JYUTPING1	ngaai6	daan1	jip6-ging1	ngoi3-bok3
	JYUTPING2	ngaai	daan	jipging	ngoibok
Hokkien		gai	sean	yapkeng	ai po
Hakka		ngioi	shan	ngiapgin	oibok

Japanese, Korean, and Others

Category	Variants	Surname 艾	Surname 单	Given Name 业经	Given Name 爱博
Japanese	Japanese	艾	単	業経	愛博
	Hiragana	あい	だん	ぎょうけい	あいはく
	Katakana	アイ	ダン	ギョウケイ	アイハク
Korean	Korean Hanja	艾	單	業經	愛博
	Hanja reading	아이	산	예징	아이보
	Korean Hangul	애	단	업경	애박
	MOE	ai	san	yejing	aibo
	NRS	ai	san	yejing	aibo
	KLS	ai	san	yejing	aibo
	ISO DPRK	ai	san	yecing	aipo
	ISO ROK	ai	san	yejing	aibo
Other languages	Vietnamese	ngải / nghệ	đơn / đan	Nghiệp Kinh	ái bác
	English	ai	shan	yejing	aibo

Thank You

شكرا

谢谢

ありがとうございました