Regularity and Flexibility in English-Chinese Name Transliteration

Oi Yee Kwong
Department of Translation
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
oykwong@arts.cuhk.edu.hk

Abstract

This paper reflects on the nature of English-Chinese personal name transliteration and the limitations of state-of-the-art language-independent automatic transliteration generation systems. English-Chinese name pairs from various sources were analysed and the complex interaction of factors in transliteration is discussed. Proposals are made for fuller error analysis in shared tasks and for expanding transliteration systems for computer-aided translation with an integrated model.

1 Introduction

Name transliteration is defined as the rendition of a name originating from a source language in a target language, such that its representation in the target language (i) is phonemically equivalent to the source name, (ii) conforms to the phonology of the target language, and (iii) matches the user intuition of the equivalent of the source language name in the target language, considering the culture and orthographic character usage in the target language. Such a definition has been adopted in the NEWS shared task on transliteration generation since 2009 (Li et al., 2009).

Automatic transliteration, or transliteration generation, has to do with the production of a transliterated name for a given source name by a trained system. Criteria (i) and (ii) above are relatively straightforward and are often the primary, if not only, concerns between most language pairs. For instance, the English name Clinton is rendered in Japanese by katakana as クリントン (ku-ri-n-to-n). The Japanese form is entirely based on phonemic resemblance as individual characters bear no particular meanings. For this reason, the correspondence is very likely to be unambiguous as long as the pronunciation is correctly figured out. Criterion (iii) above originally intends to ensure the usefulness of transliteration for downstream applications, in case the normal or expected form of the target name slightly violates the other two criteria. Nevertheless, this third criterion also applies quite specifically to target languages like Chinese. With its ideographic nature, each character does not only bear a phonetic but more importantly also a semantic component. This implies multiple possibilities for representing a particular phoneme, and consequently leads to the problem of character selection in transliteration. With the example of Clinton, the Chinese forms 克林頓 (Hanyu Pinyin: ke4-lin2-dun4) and 柯林頓 (ke1-lin2-dun4), bearing almost the same pronunciation in Mandarin Chinese, are thus both acceptable, while other homophonic forms like 刻林頓 (ke4-lin2-dun4) and 課林頓 (ke4-lin2-dun4) are not normally used.

Hence, for English-Chinese transliteration, there is obviously much greater flexibility which also encompasses a certain degree of regularity. The relatively free combination of characters in Chinese proper names is not a random phenomenon. In this paper, we show that beyond phonemic consideration, English-Chinese transliteration is actually governed by a complex but systematic interaction of various linguistic, social, cognitive, and cultural factors. Current evaluation metrics thus have limitations. With their underlying assumptions, they are good for evaluating the usefulness of transliteration systems for language processing applications, but they may not be adequate to accommodate the

Proceedings of the Fifth Named Entity Workshop, joint with 53rd ACL and the 7th IJCNLP, pages 34–42,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

whole range of possibilities which may be more appreciated by actual translation tasks. We therefore propose deeper error analysis in transliteration evaluation, and an integrated model for transliteration.

Section 2 reviews related work. Section 3 describes the data sources for our analysis. Section 4 presents general observations for English-Chinese personal name transliteration, substantiated with quantitative comparisons in Section 5 with respect to various factors. In Section 6, deeper error analysis and an integrated model of name transliteration for computer-aided translation are proposed, followed by a conclusion with future work in Section 7.

2 Related Work

There are basically two categories of work on machine transliteration. On the one hand, various alignment models are used for acquiring transliteration lexicons from parallel corpora and other resources (e.g. Lee et al., 2006; Jin et al., 2008; Kuo and Li, 2008). On the other hand, statistical transliteration models are built for transliterating personal names and other proper names, and these models can be based on phonemes (e.g. Knight and Graehl, 1998; Virga and Khudanpur, 2003), graphemes (e.g. Li et al., 2004), or their combination (e.g. Oh and Choi, 2005). They may operate on characters (e.g. Shishtla et al., 2009), syllables (e.g. Wutiwiwatchai and Thangthai, 2010), as well as hybrid units (e.g. Oh and Choi, 2005). In addition to phonetic features, others like temporal, semantic, and tonal features have also been found useful in transliteration (e.g. Tao et al., 2006; Li et al., 2007; Yoon et al., 2007; Kwong, 2009).

The baseline in current English-Chinese transliteration generation research often refers to Li et al. (2004). They used a Joint Source-Channel Model under the direct orthographic mapping (DOM) framework, which skips the middle phonemic representation in conventional phoneme-based methods, and models the segmentation and alignment preferences by means of contextual n-grams of the transliteration units. Their method was shown to outperform phoneme-based methods and those based on the noisy channel model. In fact, transliteration of foreign names into Chinese is often based on the surface orthographic forms, as exemplified in the transliteration of Beckham, where the supposedly silent h in “ham” is taken as pronounced, resulting in 漢姆 (Hanyu Pinyin: han4-mu3) in Mandarin Chinese and 咸 (Jyutping: haam4) in Cantonese.

The reports of the shared task in NEWS 2009 (Li et al., 2009) and NEWS 2010 (Li et al., 2010) highlighted two particularly popular approaches for transliteration generation among the participating systems. One is phrase-based statistical machine transliteration (e.g. Song et al., 2010; Finch and Sumita, 2010) and the other is Conditional Random Fields which treats the task as one of sequence labelling (e.g. Shishtla et al., 2009). More recent shared tasks have shown a wider array of promising techniques (Zhang et al., 2011; Zhang et al., 2012), although the absolute results as measured by Word Accuracy in Top-1 (ACC), Fuzziness in Top-1 (Mean F-score), and Mean Reciprocal Rank (MRR) have not really demonstrated any remarkable boost.

3 Resources

English and Chinese personal names obtained from various resources were analysed to illustrate the properties of Chinese naming practice and English-Chinese name transliteration. The datasets used in this study are described below.

3.1 Monolingual Chinese Names (N1)

About 40,000 distinct names written in Chinese, including authentic Chinese names (e.g. 胡錦濤 Hu Jintao, 曾蔭權 Donald Tsang) and those transliterated from foreign origins (e.g. 克林頓 Bill Clinton, 奧尼爾 Shaquille O’Neal), were obtained from the Hong Kong, Beijing and Taipei sub-corpora of the LIVAC synchronous corpus[1] (Tsou and Lai, 2003). Names from Japanese and Korean (e.g. 酒井法子 Noriko Sakai, 金大中 Kim Dae-jung) and code-mixed names (e.g. C 朗拿度 Cristiano Ronaldo, A 卡達 Anthony Carter) were excluded. Since the names are personalities appearing on news media in the various places, there are overlaps but we assume that local names also occupy a substantial proportion in each place respectively.

3.2 English-Chinese Name Pairs (N2)

About 20,000 bilingual (English-Chinese) name pairs have been manually collected from various

[1] The corpus has been word segmented and annotated with part of speech (POS). The names were extracted with the relevant POS tag.

sources including the Internet, name dictionaries, and books on naming practice. These names cover commonly used given names, big names in history, and contemporary personalities in politics, sports, entertainment, and other fields. The data were pre-processed and categorised according to:

- Region of transliteration: Hong Kong, Mainland China, or Taiwan region
- Domain: politics, sports, entertainment, or others
- Gender (if known): male or female
- Name type (if known): last name or given name

This collection was organised into sub-datasets, two of which were used in the current study. Dataset N2a is a parallel collection containing transliterations from the three Chinese speech communities for a common set of English names, mostly for celebrities. Dataset N2b consists of the transliterations for a set of common English given names, for both male and female, used predominantly in Mainland China and Taiwan region respectively. The transliterated names were also automatically mapped to Hanyu Pinyin and Jyutping for their pronunciations in Mandarin and Cantonese respectively. The mappings were manually verified.

4 General Observations

Transliteration of foreign names can lead to different possibilities across various Chinese speech communities. Phonemic equivalence is often considered with Cantonese pronunciation in Hong Kong, and Mandarin pronunciation in Mainland China and Taiwan region. This difference in pronunciation has led to very observable differences in the choice of characters, not only between Cantonese and Mandarin speaking communities, but also even between Mandarin speaking communities as in Mainland China and Taiwan region. For example, the English segment “son” as in Richardson is often rendered as 臣 (Jyutping: san4; Hanyu Pinyin: chen2) in Hong Kong, but always as 森 (Hanyu Pinyin: sen1; Jyutping: sam1) in Mainland China and 遜 (Hanyu Pinyin: xun4; Jyutping: seon3) in Taiwan region[2]. The difference in phonological properties between Mandarin and Cantonese also leads to noticeable differences in syllabification. For example, extra syllables are often introduced for certain consonant segments in the middle of an English name, as in Hamilton, transliterated as 漢密爾頓 (Hanyu Pinyin: han4-mi4-er3-dun4) in Mainland China but 咸美頓 (Jyutping: haam4-mei5-deon6) in Hong Kong. The abundance of homophones and significance of tones in Chinese also introduces much more variability, thus Rivaldo could be acceptably transliterated as 里華度 (Jyutping: lei5-waa4-dou6) or 李華度 (Jyutping: lei5-waa4-dou6). Both forms have exactly the same pronunciation except that the first may more readily suggest that it is a foreign name while the second starts with a character which is also a common Chinese surname. The phonological context embedding a particular English segment also influences the pronunciation of the segment and thus the choice of Chinese characters. Such graphemic ambiguity is an important element in transliteration.

The domain in which a personality is active often plays a role in name transliteration. For instance, names of foreign stars in the showbiz are usually fully transliterated, with given names followed by last names, e.g. Julia Roberts is known as 茱莉亞蘿拔絲 (Jyutping: zyu1-lei6-aa3-lo4-bat6-si1) in Hong Kong. On the contrary, sports stars and people in politics are often only known by their transliterated last names, such as Wayne Rooney and Bill Clinton, which usually only appear as 朗尼 (Jyutping: long5-nei4) and 克林頓 (Jyutping: haak1-lam4-deon6) in Hong Kong. In addition, the gender of the person can somehow be reflected from the transliteration via character choice among homophones. In the case of Julia Roberts, the characters 茱, 莉, 蘿 and 絲 very strongly suggest the female gender, as the first three characters all relate to flowers and plants, and the fourth character relates to silk. This practice serves to meet the social and cultural preference

[2] Both Cantonese and Mandarin pronunciations, in Jyutping and Hanyu Pinyin respectively, are given for these examples so that the readers can have some idea of their difference. For the examples in the rest of this paper, only the relevant pronunciation will be shown, according to the region in which the transliterations are used.

and the cognitive expectation of the perceivers, and it seems to be more seriously observed in Hong Kong and Taiwan region. Transliterations in Mainland China often stick quite strictly to the pronunciation, and tend to be more gender-neutral especially when only last names are transliterated. For example, the Danish tennis player Caroline Wozniacki is known by most Hong Kong media as 禾絲妮雅琪 (Jyutping: wo4-si1-nei4-ngaa5-kei4) but as 沃伊尼亞茨基 (Hanyu Pinyin: wo4-yi1-ni2-ya4-ci2-ji1) by Mainland media. The former is apparently more feminine, as the characters 絲, 妮, 雅 and 琪 are predominantly used for female names.

These general observations thus suggest that in addition to phonemic resemblance, English-Chinese name transliteration is a result of the interaction among different factors which could be linguistic, social, cognitive, and cultural in nature. In the following we will look into these factors more thoroughly with our collected data.

The interplay of these factors means that English-Chinese transliteration enjoys much more flexibility, while this freedom is accompanied by a certain degree of regularity. It also points to the need for cautious interpretation of transliteration results measured by common evaluation metrics like ACC, Mean F-score and MRR. They are based on two assumptions. One is treating the transliteration task as a closed-set problem, and the other is pre-supposing a standard reference set of "correct" transliterations. These assumptions would be reasonable and realistic for language pairs where phonemic resemblance is the entire consideration. For English-Chinese name transliteration, however, these assumptions do not take into account the possibility and acceptability (and creativity) beyond those phonemically neutral and conventional transliterations. These limitations have to be fully realised so as to perceive the performance of individual systems in a fair way.

5 Beyond Phonemic Resemblance

English-Chinese transliteration is not different from transliteration between other language pairs as phonemic resemblance is still the foremost consideration, and in this regard objective system evaluation is feasible. However, the abundance of homophones makes the naming process so much more flexible that the space for "correct" transliteration is considerably, though not unlimitedly, expanded.

5.1 Character Choice and Culture

To start with, we look at the characters often used in personal names. With the Hong Kong data in Dataset N1, we took all three-character names with Chinese origin and all foreign transliterated names, and compared the most frequent characters used in them. For the Chinese names, we only considered the second and third characters, ignoring the last names for the current comparison. Table 1 shows the top 30 characters used in the two kinds of names. It is very obvious that Chinese names and transliterated names appearing in Hong Kong media are composed of very different characters. This is possibly a result of the different phonology between English and Chinese, and thus very different pronunciations or sounds are found, leading to the use of characters in transliterated names which are not commonly found in traditional Chinese names. Among the top 100 characters in both kinds of names, only 10 characters were found in common: 德 (dak1), 亞 (aa3), 維 (wai4), 基 (gei1), 世 (sai3), 林 (lam4), 安 (ngon1), 金 (gam1), 文 (man4), and 海 (hoi2).

A similar comparison was done on the Mainland China and Taiwan region data in Dataset N1, and a similar difference between the characters used for Chinese names and transliterated names is observed. For instance, within the top 100 characters, 12% and 11% overlap were observed for Mainland China data and Taiwan region data respectively. The common characters between the two types of names in Mainland China are 德 (de2), 克 (ke4), 維 (wei2), 亞 (ya4), 基 (ji1), 林 (lin2), 安 (an1), 梅 (mei2), 金 (jin1), 文 (wen2), 小 (xiao3), and 海 (hai3); while those for the Taiwan region are 德 (de2), 瑞 (rui4), 維 (wei2), 達 (da2), 安 (an1), 吉 (ji2), 傑 (jie2), 林 (lin2), 雅 (ya3), 金 (jin1), and 華 (hua2).

Comparing the characters used for transliterated names among the three regions, it is apparent that Hong Kong and Mainland China tend to use more similar characters (although the precise syllabification and correspondence between English and Chinese segments might be different, as discussed in Section 5.2 below), while Taiwan region has a somewhat different character choice. Table 2 shows the commonality and difference among the three communities with respect to the top 100 characters in individual regions.

No.	HK Chinese	HK Foreign	No.	HK Chinese	HK Foreign
1	國	斯	16	英	布
2	文	爾	17	東	阿
3	華	德	18	雄	納
4	明	拉	19	生	巴
5	志	克	20	清	科
6	建	特	21	家	迪
7	德	夫	22	仁	亞
8	永	里	23	小	森
9	偉	羅	24	輝	伊
10	光	卡	25	中	維
11	平	利	26	林	姆
12	榮	馬	27	麗	雷
13	強	尼	28	金	普
14	玉	哈	29	慶	米
15	成	格	30	昌	基

Table 1: Top 30 characters used in Chinese and transliterated names from Dataset N1 (HK)

Comparison \ Region	Hong Kong	Mainland China	Taiwan region
Common	斯 克 爾 拉 特 德 尼 卡 羅 夫 里 布 艾 諾 利 馬 巴 格 維 洛 亞 阿 納 茲 哈 西 迪 麥 森 曼 達 普 塔 安 雷 魯 瓦 貝 伊 吉 恩 米 希 蘭 波 姆 威 奇 莫 萊 伯 勒 沙 薩 凱 基 比 托 倫 索 多 蒂 塞 林 法 奧 蘇 梅 杜 頓 科 金 帕 菲 赫 耶 費 加 穆		
Unique	世 高 二 盧	什 蒙 小 朗 茨	瑞 絲 莉 歐 柯 佛 瑪 葛 提 娜 柏 傑 妮 可 雅 莎 華 賈 丹 娃

Table 2: Comparison of character choice in individual communities from Dataset N1

5.2 Linguistic Factors

According to Dobrovolsky and Katamba (1996), native speakers of any language intuitively know that certain words that come from other languages sound unusual and they often adjust the segment sequences of these words to conform to the pronunciation requirements of their own language. These intuitions are based on a tacit knowledge of the permissible syllable structures of the speaker's own language. The difference between transliterations based on Mandarin and Cantonese is particularly obvious between Mainland China and Hong Kong, where the resulting number of syllables in the transliterated names is on average higher for the former. With Dataset N2a, we can compare the regional differences with respect to a more or less common set of transliterated names.

Among the common set of names, it was found that the average number of syllables (which correspond to the Chinese characters) is 2.60, 2.88, and 2.74 for Hong Kong, Mainland China, and Taiwan region respectively. This is mostly due to phonological differences. English and Chinese have very different phonological properties. A well cited example is a syllable initial /d/ may surface as in Baghdad 巴格達 (Hanyu Pinyin: ba1-ge2-da2), but the syllable final /d/ is not represented. This is true for Mandarin Chinese, but since ending stops like -p, -t, and -k are allowed in Cantonese syllables, the syllable final /d/ in Baghdad is already captured in the last syllable of 巴格達 (Jyutping: baa1-gaak3-daat6) in Cantonese. This difference in allowable codas sometimes surfaces in the form of an additional syllable in transliterations based on Mandarin. For example, Dickson is transliterated as 迪克遜 (Hanyu Pinyin: di2-ke4-xun4) in Mandarin Chinese and 迪臣 (Jyutping: dik6-san4) in Cantonese, where no extra syllable is introduced in the latter. This possibly accounts for the greater number of syllables for transliterations found in Mainland China and Taiwan region, as both these communities transliterate by Mandarin pronunciations. This is also reflected in the top English-Chinese segment pairs found from the three places, as shown in Table 3. From the table, we can see that English segments like "D", "T", "C", and "K" occupy the top positions for Mainland China and Taiwan region, where they consistently demand an additional syllable in the transliteration based on Mandarin. Although the corresponding segments are sometimes found in Hong Kong transliterations, they are nevertheless not as apparent and frequent.

As far as intra-regional variability is concerned, it is interesting to note that there are 1,974 distinct English-Chinese segment pairs in the Hong Kong data, but only 1,411 and 1,734 distinct pairs in the Mainland China and Taiwan region data respectively. This suggests that transliterations in Mainland China are most consistent, if not perfectly standardised. For instance, in the *Chinese Transliteration of Foreign Personal Names* published by the Xinhua News Agency (1992), a table showing the prescriptive Chinese rendition of individual English syllables is included. Transliterations in Hong Kong, however, are much more variable, and there are many ways to render a particular syllable.

No.	**Hong Kong**		**Mainland China**		**Taiwan Region**	
1	S	斯	S	斯	S	斯
2	SON	遜	L	爾	D	德
3	S	史	D	德	T	特
4	L	爾	T	特	K	克
5	TON	頓	C	克	B	布
6	G	格	SON	森	SON	森
7	O	奧	RI	里	C	克
8	A	亞	B	布	S	史
9	A	艾	G	格	RO	羅
10	BA	巴	K	克	TON	頓

Table 3: Top 10 English-Chinese segment pairs from Dataset N2a

5.3 Cognitive Factors

There are only a few hundred Chinese characters commonly used in transliterated names. Although their choice and combination are relatively free, the flexibility is not entirely ungoverned. For instance, the former Brazilian striker Ronaldo is typically rendered as 朗拿度 (Jyutping: long5-naa4-dou6) in Cantonese, but never as phonetically equivalent candidates like 朗娜度 (Jyutping: long5-naa4-dou6) or 郎拿刀 (Jyutping: long4-naa4-dou1). In this example, the second candidate is not preferred, as 娜 is conventionally restricted to female names (further discussed in Section 5.4 below). The third candidate is also not suitable. Even though 郎 is masculine, 刀 is probably not a character with enough positive meanings and is only occasionally found in Chinese names. This consideration in character choice is apparently cognitively based, with regard to the positive and negative connotations of individual characters, and thus their suitability for names. Apart from that, cognitive factors may involve the intonation of a name, which may also make a difference in the preference of a name. In particular, Chinese is a typical tonal language. Cantonese, in particular, has more tones than Mandarin, and the sound-tone combination is more important in names pronounced in Cantonese. Names which sound “nice” (or more “musical”) are often preferred to those which sound “monotonous”. It is thus important to consider the tone combination in transliteration. To this end, Kwong (2009) has shown that the improvement from including tones in a Joint Source-Channel model for automatic transliteration was more apparent for Cantonese data.

5.4 Social Factors

Gender difference is often reflected in the character choice for the transliterated names. Table 4 shows the most frequent characters for transliterating male and female given names in Mainland China and Taiwan region as analysed from Dataset N2b.

No.	Mainland China		Taiwan Region	
	Male	Female	Male	Female
1	斯	娜	斯	莉
2	爾	麗	爾	娜
3	里	莉	克	拉
4	特	拉	瑞	絲
5	德	爾	德	妮
6	克	特	特	瑪
7	利	絲	艾	西
8	尼	妮	尼	琳
9	羅	德	羅	安
10	雷	婭	利	凱

Table 4: Top 10 characters for male and female names in Dataset N2b

In terms of gender difference, the character sets are quite different for male and female names, in both regions alike. For instance, 219 and 256 distinct Chinese characters were found for female names and male names respectively from the Mainland China data, with 174 characters in common. For Taiwan region data, 230 and 275 distinct Chinese characters were found for female names and male names respectively, with only 136 characters in common. In other words, it suggests that the gender difference is much more apparent and significant for transliterations in Taiwan region, whereas transliterations in Mainland China tend to use more gender-neutral characters (as already shown in the example of Wozniacki earlier).

The actual characters used in transliteration in both regions are also considerably different. For instance, among the 219 and 230 characters for female names in Mainland China and Taiwan region respectively, 139 are in common (that is, around 60%); whereas among the 256 and 275 characters for males names in the two places respectively, 199 characters are in common (that is, over 70%). Hence the difference between the two regions is greater in the transliteration of female names than that of male names.

Table 5 shows an example for the English segment “LI”. It is obvious that the general graphemic and homophone ambiguity can somehow be reduced when gender is taken into account. For instance, 莉 (li4) and 麗 (li4) are mostly restricted to female names, whereas 力 (li4) and 立 (li4) are predominantly used for male names. Others like 利 (li4) and 里 (li3) are more or less gender-neutral.

The ability to distinguish the gender from the transliterated name is particularly useful as it could help resolve ambiguity in translation especially when there are more than one possible candidate bearing the same last name, such as John Williams the musician and Venus Williams the woman tennis player. The gender factor in transliteration thus bears important implications not only in (back) transliteration but also in translation in general.

	Male	**Female**
Mainland China	利 (Cliff 克利夫) 里 (Ali 阿里) 萊 (Clive 克萊夫)	利 (Melissa 梅利莎) 里 (Ali 阿里) 莉 (Alisha 阿莉莎) 萊 (Carolina 卡羅萊娜) 麗 (Alice 艾麗斯)
Taiwan Region	力 (Philip 菲力普) 立 (Oliver 奧立佛) 利 (Julian 朱利安) 里 (Cliff 克里夫) 萊 (Linus 萊納斯) 賴 (Elijah 伊賴嘉)	里 (Celia 賽里雅) 莉 (Alisha 艾莉夏) 琳 (Carolina 卡蘿琳娜) 麗 (Lisa 麗莎)

Table 5: Examples of gender-specific rendition of the English segment “LI” from Dataset N2b

Thus in this section, we have discussed the impact of various factors on English-Chinese personal name transliteration with empirical evidence. In particular, we have investigated the complex interaction among syllabification, phonological difference, homophones, tones, gender, and domain, in transliteration across three Chinese speech communities, namely Hong Kong, Mainland China, and Taiwan region.

6 Proposals

6.1 Deeper Error Analysis

With the current paradigm adopted in the shared task on transliteration generation, systems are evaluated by how often the first-ranked transliteration generated by a system matches the “answer” given in the evaluation data, and on average when will the “answer” appear in the top 10 transliterations given by the system. In terms of providing a common platform for evaluation, this is a natural and reasonable approach. However, it should not be disregarded that even if the system-generated result is not exactly the same as that in the evaluation data, it does not necessarily mean it is “wrong” or useless. As discussed above, English-Chinese name transliteration involves the interaction of linguistic, cognitive, social and cultural factors, and multiple renditions could be considered acceptable. Thus, usually there is no right or wrong, but better or worse, for the system-generated transliteration candidates.

In fact, when we look at the evaluation results over the last few shared tasks, there is no remarkable breakthrough observed. Literally the figures seem to be deteriorating. Considering the system with top performance in the English-Chinese transliteration task (standard run), the ACC, F-score and MRR were 0.731, 0.895 and 0.812 respectively in 2009. In 2010, they are 0.477, 0.740 and 0.506 respectively. In 2011, they are 0.349, 0.700 and 0.462 respectively. In 2012, they are 0.330, 0.669 and 0.413 respectively.

It will certainly be unfair to compare the above figures directly since different datasets were used, but the situation also raises the issue of robustness. The shared task for this year is re-using the test data from one of the previous years in order to track system improvement.[3] In addition to this, we suggest that deeper error analysis would be useful to obtain a better idea of the limitation of state-of-the-art system performance. It would be important to find out whether the bottleneck is possibly caused by the difference in training data, and whether the “unmatched” transliteration candidates could

[3] According to the shared task results provided by one of the anonymous reviewers, results for standard runs in the EnCh task do not seem to demonstrate any remarkable improvement. Further information and discussion are expected with the release of the official analysis and comparison.

also be considered acceptable; or otherwise what might have led to their unacceptability. For instance, the inexact matches generated by systems could be further analysed and classified according to the nature of the "errors", such as phonemic non-equivalence, character mismatch, character misuse, unseen characters, tone problem or perceptual idiosyncrasy, and region compatibility, just to name a few possibilities. It will be worthwhile to pursue such a direction in future evaluation of machine transliteration, while a similar need to augment automatic metric with linguistic and perceptual considerations for machine translation evaluation has been realised and proposed by Farrús et al. (2012). One of the primary concerns would naturally be the balance between automatic and manual work to be involved in the whole evaluation process.

6.2 An Integrated Model

The ability for systems to produce linguistically and cognitively acceptable transliterations is particularly important. New names or unseen names appear every day in the media, and accurate and reasonable renditions of foreign names into Chinese will be very useful, not only for downstream language processing applications, but also as a significant component for computer-aided translation in practice. Transliteration is to render a source name in a phonemically similar way in a target language. The linguistic factors, considering the phonological properties of the two languages and thus the syllabification, should bear primary importance. Other interacting factors, including the intonation, gender difference, and domain, may be considered peripheral, but considering them would certainly help produce better perceived candidates. For the case of English-Chinese transliteration, the cultural differences must not be ignored. They must be taken into account to ensure that the resulting transliterations are intelligible and appropriate to the Chinese speakers in individual regions.

An integrated model for transliteration is therefore necessary, although this might be at the expense of a completely language-independent design. We propose that a transliteration system should contain three major components, for segmentation, candidate generation, and candidate ranking respectively. The segmentation module should consist of a linguistic model, to break up a source name into pronunciation segments. The linguistic model incorporates language-specific phonological properties (for both the source language and target language), for initial syllabification of the source name and reconstructing the segmentation structure into one which is compatible with the requirements of the target language. The candidate generation module should consist of a cultural model, which provides information on the naming practice adopted in various cultures and the range of orthographic renditions usually allowed for personal names. The candidate ranking module should consist of a social module to compare the candidate transliterations for their desirability according to social factors like gender difference and domain preference, as well as a cognitive module to consider factors like pleasantness of sound and intonation, and avoidance of unfavourable homophone strings.

7 Conclusion and Future Work

In this paper, we have reflected on the nature of English-Chinese name transliteration, which is distinct from transliteration between other language pairs in its much greater flexibility beyond pure phonemic equivalence. A complex yet systematic interplay of cultural, linguistic, cognitive and social factors was shown from empirical data. On the one hand, we suggest that deeper error analysis of transliteration systems be performed to realise the limitations of common evaluation metrics. On the other hand, we propose an integrated model for a robust English-Chinese transliteration system. Practical systems, especially those for computer-aided translation, should consider the art and science of the transliteration task. In order to consider a realistically wider range of transliteration candidates, a system should take into account various interacting factors while capitalising on statistical patterns. The implementation of such a system will constitute an important part of our future work.

References

Dobrovolsky, M. and F. Katamba. 1996. Phonology: the function and patterning of sounds. In W. O'Grady, M. Dobrovolsky and F. Katamba (Eds.), *Contemporary Linguistics: An Introduction*. Essex: Addison Wesley Longman Limited.

Farrús, M., M.R. Costa-jussà and M. Popović. 2012. Study and Correlation Analysis of Linguistic, Perceptual, and Automatic Machine Translation Evaluations. *Journal of the American Society for Information Science and Technology, 63(1)*:174-184.

Finch, A. and E. Sumita. 2010. Transliteration using a phrase-based statistical machine translation system to re-score the output of a joint multigram model. In *Proceedings of NEWS 2010*, Uppsala, Sweden.

Jin, C., S-H. Na, D-I. Kim and J-H. Lee. 2008. Automatic Extraction of English-Chinese Transliteration Pairs using Dynamic Window and Tokenizer. In *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing (SIGHAN-6)*, Hyderabad, India, pp.9-15.Katamba, F. 1989. *An Introduction to Phonology*. Essex: Longman Group UK Limited.

Knight, K. and J. Graehl. 1998. Machine Transliteration. *Computational Linguistics, 24(4)*:599-612.

Kuo, J-S. and H. Li. 2008. Mining Transliterations from Web Query Results: An Incremental Approach. In *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing (SIGHAN-6)*, Hyderabad, India, pp.16-23.

Kwong, O.Y. 2009. Homophones and Tonal Patterns in English-Chinese Transliteration. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, Singapore, pp.21-24.

Lee, C-J., J.S. Chang and J-S.R. Jang. 2006. Extraction of transliteration pairs from parallel corpora using a statistical transliteration model. *Information Sciences, 176*:67-90.

Li, H., A. Kumaran, V. Pervouchine and M. Zhang. 2009. Report of NEWS 2009 Machine Transliteration Shared task. In *Proceedings of NEWS 2009*, Singapore.

Li, H., A. Kumaran, M. Zhang and V. Pervouchine. 2010. Report of NEWS 2010 Transliteration Generation Shared Task. In *Proceedings of NEWS 2010*, Uppsala, Sweden.

Li, H., K.C. Sim, J-S. Kuo and M. Dong. 2007. Semantic Transliteration of Personal Names. In *Proceedings of ACL 2007*, Prague, Czech Republic, pp.120-127.

Li, H., M. Zhang and J. Su. 2004. A Joint Source-Channel Model for Machine Transliteration. In *Proceedings of ACL 2004*, Barcelona, Spain, pp.159-166.

Oh, J-H. and K-S. Choi. 2005. An Ensemble of Grapheme and Phoneme for Machine Transliteration. In R. Dale *et al.* (Eds.), *Natural Language Processing – IJCNLP 2005*. Springer, LNAI Vol. 3651, pp.451-461.

Shishtla, P., V.S. Ganesh, S. Sethuramalingam and V. Varma. 2009. A language-independent transliteration schema using character aligned models. In *Proceedings of NEWS 2009*, Singapore.

Song, Y., C. Kit and H. Zhao. 2010. Reranking with multiple features for better transliteration. In *Proceedings of NEWS 2010*, Uppsala, Sweden.

Tao, T., S-Y. Yoon, A. Fister, R. Sproat and C. Zhai. 2006. Unsupervised Named Entity Transliteration Using Temporal and Phonetic Correlation. In *Proceedings of EMNLP 2006*, Sydney, Australia, pp.250-257.

Tsou, B.K. 鄒嘉彥 and T.B.Y. Lai 黎邦洋. 2003. 漢語共時語料庫與資訊開發，《中文資訊處理若干重要問題》, pp.147-165. 北京:科學出版社.

Virga, P. and S. Khudanpur. 2003. Transliteration of Proper Names in Cross-lingual Information Retrieval. In *Proceedings of the ACL2003 Workshop on Multilingual and Mixed-language Named Entity Recognition*.

Wutiwiwatchai, C. and A. Thangthai. 2010. Syllable-based Thai-English Machine Transliteration. In *Proceedings of NEWS 2010*, Uppsala, Sweden, pp.66-70.

Xinhua News Agency. 1992. *Chinese Transliteration of Foreign Personal Names*. The Commercial Press.

Yoon, S-Y., K-Y. Kim and R. Sproat. 2007. Multilingual Transliteration Using Feature based Phonetic Method. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, Prague, Czech Republic, pp.112-119.

Zhang, M., H. Li, A. Kumaran and M. Liu. 2011. Report of NEWS 2011 Transliteration Generation Shared Task. In *Proceedings of NEWS 2011*, Chiang Mai, Thailand, pp.1-13.

Zhang, M., H. Li, A. Kumaran and M. Liu. 2012. Report of NEWS 2012 Machine Transliteration Shared task. In *Proceedings of NEWS 2012*, Jeju, Korea, pp.10-20.