# Homophones and Tonal Patterns in English-Chinese Transliteration

**Oi Yee Kwong**
Department of Chinese, Translation and Linguistics
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong
Olivia.Kwong@cityu.edu.hk

## Abstract

The abundance of homophones in Chinese significantly increases the number of similarly acceptable candidates in English-to-Chinese transliteration (*E2C*). The dialectal factor also leads to different transliteration practice. We compare *E2C* between Mandarin Chinese and Cantonese, and report work in progress for dealing with homophones and tonal patterns despite potential skewed distributions of individual Chinese characters in the training data.

## 1 Introduction

This paper addresses the problem of automatic English-Chinese forward transliteration (referred to as *E2C* hereafter).

There are only a few hundred Chinese characters commonly used in names, but their combination is relatively free. Such flexibility, however, is not entirely ungoverned. For instance, while the Brazilian striker Ronaldo is rendered as 朗拿度 *long5-naa4-dou6* in Cantonese, other phonetically similar candidates like 朗娜度 *long5-naa4-dou6* or 郎拿刀 *long4-naa4-dou1*[1] are least likely. Beyond linguistic and phonetic properties, many other social and cognitive factors such as dialect, gender, domain, meaning, and perception, are simultaneously influencing the naming process and superimposing on the surface graphemic correspondence.

The abundance of homophones in Chinese further complicates the problem. Past studies on phoneme-based *E2C* have reported their adverse effects (e.g. Virga and Khudanpur, 2003). Direct orthographic mapping (e.g. Li *et al.*, 2004), making use of individual Chinese graphemes, tends to overcome the problem and model the character choice directly. Meanwhile, Chinese is a typical tonal language and the tone information can help distinguish certain homophones. Phoneme mapping studies seldom make use of tone information. Transliteration is also an open problem, as new names come up everyday and there is no absolute or one-to-one transliterated version for any name. Although direct orthographic mapping has implicitly or partially modelled the tone information via individual characters, the model nevertheless heavily depends on the availability of training data and could be skewed by the distribution of a certain homophone and thus precludes an acceptable transliteration alternative. We therefore propose to model the sound and tone together in *E2C*. In this way we attempt to deal with homophones more reasonably especially when the training data is limited. In this paper we report some work in progress and compare *E2C* in Cantonese and Mandarin Chinese.

Related work will be briefly reviewed in Section 2. Some characteristics of *E2C* will be discussed in Section 3. Work in progress will be reported in Section 4, followed by a conclusion with future work in Section 5.

## 2 Related Work

There are basically two categories of work on machine transliteration. First, various alignment models are used for acquiring transliteration lexicons from parallel corpora and other resources (e.g. Kuo and Li, 2008). Second, statistical models are built for transliteration. These models could be phoneme-based (e.g. Knight and Graehl, 1998), grapheme-based (e.g. Li *et al.*, 2004), hybrid (Oh and Choi, 2005), or based on phonetic (e.g. Tao *et al.*, 2006) and semantic (e.g. Li *et al.*, 2007) features.

Li *et al.* (2004) used a Joint Source-Channel Model under the direct orthographic mapping

[1] Mandarin names are transcribed in Hanyu Pinyin and Cantonese names are transcribed in Jyutping published by the Linguistic Society of Hong Kong.

*Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 21–24,
Suntec, Singapore, 4 August 2009. ©2009 ACL and AFNLP

(DOM) framework, skipping the middle phonemic representation in conventional phoneme-based methods, and modelling the segmentation and alignment preferences by means of contextual n-grams of the transliteration units. Although DOM has implicitly modelled the tone choice, since a specific character has a specific tone, it nevertheless heavily relies on the availability of training data. If there happens to be a skewed distribution of a certain Chinese character, the model might preclude other acceptable transliteration alternatives. In view of the abundance of homophones in Chinese, and that sound-tone combination is important in names (i.e., names which sound "nice" are preferred to those which sound "monotonous"), we propose to model sound-tone combinations in transliteration more explicitly, using pinyin transcriptions to bridge the graphemic representation between English and Chinese. In addition, we also study the dialectal differences between transliteration in Mandarin Chinese and Cantonese, which is seldom addressed in past studies.

## 3 Some *E2C* Properties

### 3.1 Dialectal 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 *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 *baa1-gaak3-daat6* in Cantonese.

Such phonological difference between Mandarin Chinese and Cantonese might also account for the observation that Cantonese transliterations often do not introduce extra syllables for certain consonant segments in the middle of an English name, as in Dickson, transliterated as 迪克遜 *di2-ke4-xun4* in Mandarin Chinese and 迪臣 *dik6-san4* in Cantonese.

### 3.2 Ambiguities from Homophones

The homophone problem is notorious in Chinese. As far as personal names are concerned, the "correctness" of transliteration is not clear-cut at all. For example, to transliterate the name Hilary into Chinese, based on Cantonese pronunciations, the following are possibilities amongst many others: (a) 希拉利 *hei1-laai1-lei6*, (b) 希拉莉 *hei1-laai1-lei6*, and (c) 希拉里 *hei1-laai1-lei5*.

The homophonous third character gives rise to multiple alternative transliterations in this example, where orthographically 利 *lei6*, 莉 *lei6* and 里 *lei5* are observed for "ry" in transliteration data. One cannot really say any of the combinations is "right" or "wrong", but perhaps only "better" or "worse". Such judgement is more cognitive than linguistic in nature, and apparently the tonal patterns play an important role in this regard. Hence naming is more of an art than a science, and automatic transliteration should avoid over-reliance on the training data and thus missing unlikely but good candidates.

## 4 Work in Progress

### 4.1 Datasets

A common set of 1,423 source English names and their transliterations[2] in Mandarin Chinese (as used by media in Mainland China) and Cantonese (as used by media in Hong Kong) were collected over the Internet. The names are mostly from soccer, entertainment, and politics. The data size is admittedly small compared to other existing transliteration datasets, but as a preliminary study, we aim at comparing the transliteration practice between Mandarin speakers and Cantonese speakers in a more objective way based on a common set of English names. The transliteration pairs were manually aligned, and the pronunciations for the Chinese characters were automatically looked up.

### 4.2 Preliminary Quantitative Analysis

| | Cantonese | Mandarin |
|---|---|---|
| Unique name pairs | 1,531 | 1,543 |
| Total English segments | 4,186 | 4,667 |
| Unique English segments | 969 | 727 |
| Unique grapheme pairs | 1,618 | 1,193 |
| Unique seg-sound pairs | 1,574 | 1,141 |

Table 1. Quantitative Aspects of the Data

As shown in Table 1, the average segment-name ratios (2.73 for Cantonese and 3.02 for Mandarin) suggest that Mandarin transliterations often use more syllables for a name. The much smaller number of unique English segments for Mandarin and the difference in token-type ratio of grapheme pairs (3.91 for Mandarin and 2.59 for Cantonese) further suggest that names are more consistently segmented and transliterated in Mandarin.

[2] Some names have more than one transliteration.

### 4.2.1 Graphemic Correspondence

Assume grapheme pair mappings are in the form <$e_k$, {$c_{k1}$,$c_{k2}$,…,$c_{kn}$}>, where $e_k$ stands for the *k*th unique English segment from the data, and {$c_{k1}$,$c_{k2}$,…,$c_{kn}$} for the set of *n* unique Chinese segments observed for it. It was found that *n* varies from 1 to 10 for Mandarin, with 34.9% of the distinct English segments having multiple grapheme mappings, as shown in Table 2. For Cantonese, *n* varies from 1 to 13, with 31.5% of the distinct English segments having multiple grapheme mappings. The proportion of multiple mappings is similar for Mandarin and Cantonese, but the latter has a higher percentage of English segments with 5 or more Chinese renditions. Thus Mandarin transliterations are relatively more “standardised”, whereas Cantonese transliterations are graphemically more ambiguous.

| *n* | Cantonese | Mandarin |
|---|---|---|
| >=5 | 5.3% | 3.3% |
| 4 | 4.0% | 4.4% |
| 3 | 6.2% | 7.2% |
| 2 | 16.0% | 20.0% |
| 1 | 68.5% | 65.1% |
| Example | <le, {列, 利, 勒, 尼, 李, 歷, 烈, 爾, 理, 萊, 路, 里, 雷}> | <le, {列, 利, 勒, 歷, 爾, 理, 萊, 裏, 路, 雷}> |

Table 2. Graphemic Ambiguity of the Data

### 4.2.2 Homophone Ambiguity (Sound Only)

Table 3 shows the situation with homophones (ignoring tones). For example, all five characters 利莉李里理 correspond to the Jyutping *lei*. Despite the tone difference, they are considered homophones in this section.

| n | Cantonese | Mandarin |
|---|---|---|
| >=5 | 3.3% | 1.9% |
| 4 | 4.0% | 2.5% |
| 3 | 5.8% | 5.7% |
| 2 | 16.3% | 20.7% |
| 1 | 70.5% | 69.2% |
| Example | <le, {ji, laak, lei, leoi, lik, lit, loi, lou, nei}> | <le, {er, lai, le, lei, li, lie, lu}> |

Table 3. Homophone Ambiguity (Ignoring Tone)

Assume grapheme-sound pair mappings are in the form <$e_k$, {$s_{k1}$,$s_{k2}$,…,$s_{kn}$}>, where $e_k$ stands for the *k*th unique English segment, and {$s_{k1}$,$s_{k2}$,…,$s_{kn}$} for the set of *n* unique pronunciations (regardless of tone). For Mandarin, *n* varies from 1 to 7, with 30.8% of the distinct English segments having multiple sound mappings. For Cantonese, *n* varies from 1 to 9, with 29.5% of the distinct English segments having multiple sound mappings. Comparing with Table 2 above, the downward shift of the percentages suggests that much of the graphemic ambiguity is a result of the use of homophones, instead of a set of characters with very different pronunciations.

### 4.2.3 Homophone Ambiguity (Sound-Tone)

Table 4 shows the situation of homophones with both sound and tone taken into account. For example, the characters 利莉 all correspond to *lei6* in Cantonese, while 李里理 all correspond to *lei5*, and they are thus treated as two groups.

Assume grapheme-sound/tone pair mappings are in the form <$e_k$, {$st_{k1}$,$st_{k2}$,…,$st_{kn}$}>, where $e_k$ stands for the *k*th unique English segment, and {$st_{k1}$,$st_{k2}$,…,$st_{kn}$} for the set of *n* unique pronunciations (sound-tone combination). For Mandarin, *n* varies from 1 to 8, with 33.5% of the distinct English segments corresponding to multiple Chinese homophones. For Cantonese, *n* varies from 1 to 10, with 30.8% of the distinct English segments having multiple Chinese homophones.

| *n* | Cantonese | Mandarin |
|---|---|---|
| >=5 | 4.1% | 2.8% |
| 4 | 4.8% | 3.3% |
| 3 | 6.1% | 6.8% |
| 2 | 15.8% | 20.7% |
| 1 | 69.2% | 66.5% |
| Example | <le, {ji5, laak6, lei5, lei6, leoi4, lik6, lit6, loi4, lou6, nei4}> | <le, {er3, lai2, le4, lei2, li3, li4, lie4, lu4} |

Table 4. Homophone Ambiguity (Sound-Tone)

The figures in Table 4 are somewhere between those in Table 2 and Table 3, suggesting that a considerable part of homophones used in the transliterations could be distinguished by tones. This supports our proposal of modelling tonal combination explicitly in *E2C.*

### 4.3 Method and Experiment

The Joint Source-Channel Model in Li *et al.* (2004) was adopted in this study. However, instead of direct orthographic mapping, we model the mapping between an English segment and the pronunciation in Chinese. Such a model is expected to have a more compact parameter space as individual Chinese characters for a certain English segment are condensed into homophones defined by a finite set of sounds and tones. The model could save on computational effort, and is less affected by any bias or sparseness of the data. We refer to this approach as SoTo hereafter.

Hence our approach with a bigram model is as follows:

$$
\begin{aligned}
P(E,ST) &= P(e_1,e_2,...,e_k,st_1,st_2,...,st_k) \\
&= P(<e_1,st_1>,<e_2,st_2>,...,<e_k,st_k>) \\
&= \prod_{k=1}^{K} P(<e_k,st_k>|<e_{k-1},st_{k-1}>)
\end{aligned}
$$

where $E$ refers to the English source name and $ST$ refers to the sound/tone sequence of the transliteration, while $e_k$ and $st_k$ refer to $k$th segment and its Chinese sound respectively. Homophones in Chinese are thus captured as a class in the phonetic transcription. For example, the expected Cantonese transliteration for Osborne is 奧斯邦尼 *ou3-si1-bong1-nei4*. Not only is it ranked first using this method, its homophonous variant 奧施邦尼 is within the top 5, thus benefitting from the grouping of the homophones, despite the relatively low frequency of <s,施>. This would be particularly useful for transliteration extraction and information retrieval.

Unlike pure phonemic modelling, the tonal factor is modelled in the pronunciation transcription. We do not go for phonemic representation from the source name as the transliteration of foreign names into Chinese is often based on the surface orthographic forms, e.g. the silent h in Beckham is pronounced to give 漢姆 *han4-mu3* in Mandarin and 咸 *haam4* in Cantonese.

Five sets of 50 test names were randomly extracted from the 1.4K names mentioned above for 5-fold cross validation. Training was done on the remaining data. Results were also compared with DOM. The Mean Reciprocal Rank (MRR) was used for evaluation (Kantor and Voorhees, 2000).

### 4.4 Preliminary Results

| Method | Cantonese | Mandarin |
|---|---|---|
| DOM | 0.2292 | 0.3518 |
| SoTo | 0.2442 | 0.3557 |

Table 5. Average System Performance

Table 5 shows the average results of the two methods. The figures are relatively low compared to state-of-the-art performance, largely due to the small datasets. Errors might have started to propagate as early as the name segmentation step. As a preliminary study, however, the potential of the SoTo method is apparent, particularly for Cantonese. A smaller model thus performs better, and treating homophones as a class could avoid over-reliance on the prior distribution of individual characters. The better performance for Mandarin data is not surprising given the less "standardised" Cantonese transliterations as discussed above. From the research point of view, it suggests more should be considered in addition to grapheme mapping for handling Cantonese data.

## 5 Future Work and Conclusion

Thus we have compared *E2C* between Mandarin Chinese and Cantonese, and discussed work in progress for our proposed SoTo method which more reasonably treats homophones and better models tonal patterns in transliteration. Future work includes testing on larger datasets, more in-depth error analysis, and developing better methods to deal with Cantonese transliterations.

## Acknowledgements

The work described in this paper was substantially supported by a grant from City University of Hong Kong (Project No. 7002203).

## References

Kantor, P.B. and Voorhees, E.M. (2000) The TREC-5 Confusion Track: Comparing Retrieval Methods for Scanned Text. *Information Retrieval, 2(2-3)*: 165-176.

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

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

Li, H., Zhang, M. and Su, J. (2004) A Joint Source-Channel Model for Machine Transliteration. In *Proceedings of the 42nd Annual Meeting of ACL*, Barcelona, Spain, pp.159-166.

Li, H., Sim, K.C., Kuo, J-S. and Dong, M. (2007) Semantic Transliteration of Personal Names. In *Proceedings of the 45th Annual Meeting of ACL*, Prague, Czech Republic, pp.120-127.

Oh, J-H. and Choi, K-S. (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.

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

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