English-Chinese Name Transliteration with Bi-Directional Syllable-Based Maximum Matching [*]

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. In this paper, we propose a simple and intuitive yet linguistically and practically motivated method for English-Chinese name transliteration generation. Our system is essentially a syllable-based Maximum Matching system. It uses the Onset First Principle to syllabify English names and align them with Chinese names. The bilingual lexicon containing aligned segments of various syllable lengths subsequently allows direct transliteration by chunks. The proposed method was tested on the data from the shared task of the Named Entities Workshop 2009. The results suggest that Forward Maximum Matching performed slightly better than Backward Maximum Matching, but when used together much better results comparable to those of state-of-the-art methods could be attained.

Keywords: Name transliteration, Syllable-based Maximum Matching.

1 Introduction

In this paper, we propose a syllable-based Maximum Matching method for English-Chinese name transliteration generation. Compared to most other popular methods for the purpose, our approach is relatively simple and intuitively appealing, but at the same time it has strong linguistic and practical motivation.

Our system is simply trained on a set of English-Chinese name pairs. Two major principles underlie its design. On the one hand, syllables form the primary units in transliteration. While the rendition of a name in the target language is expected to phonemically resemble its pronunciation in the source language, difference in phonotactics between the two languages would be compromised during transliteration, and incompatible phonological structures might result in extra syllables in the transliterated name. On the other hand, name transliteration tends to be more conventional than innovative, such that the same chunk is often transliterated in the same way even when embedded in different names. Thus, with our proposed method, the Onset First Principle in phonology was used to syllabify English names and align them with the Chinese renditions. A bilingual lexicon containing segment pairs of various syllable lengths was then produced from the aligned names. This lexicon was subsequently used in transliteration, during which a source name was first syllabified and then segmented using Maximum Matching with syllables as the basic units. Target candidates were generated by looking up the bilingual lexicon. Both directions of the Maximum Matching and various ways of ranking the candidates have been tested. Experimental results show that while Forward Maximum Matching performed slightly better than Backward Maximum Matching, their

[*] The work described in this paper was supported by a grant from City University of Hong Kong (Project No. 7008004).

combined use could lead to much better results comparable to those produced by state-of-the-art methods.

We will briefly review related work in Section 2, and introduce the datasets used in this study in Section 3. The system will be described in Section 4. Experiments and results will be reported in Section 5, followed by future work and conclusion in Section 6.

2 Related Work

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). Besides these popular methods, for instance, Huang *et al.* (2011) used a non-parametric Bayesian learning approach in a recent study.

Regarding the basic unit of transliteration, traditional systems are mostly phoneme-based (e.g. Knight and Graehl, 1998). Li *et al.* (2004) suggested a grapheme-based Joint Source-Channel Model within the Direct Orthographic Mapping framework. Models based 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), are also seen. 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; Kwong, 2009).

3 Datasets

In the current study, we used the transliteration data provided by the organiser of the transliteration generation shared task (English-to-Chinese track) in the Named Entities Workshop 2009 (NEWS 2009) for testing, which are mostly based on name pairs from Xinhua News Agency (1992). There are 31,961 English-Chinese name pairs in the training set, 2,896 pairs in the development set, and another 2,896 English names in the test set. The Chinese transliterations basically correspond to Mandarin Chinese pronunciations of the English names, as used by the media in Mainland China. The training set and development set were both used for training, and experiments were done on the test set.

All English names were first converted to upper case letters, and all occurrences of “X” were replaced by “KS” before processing to facilitate subsequent syllabification, as a single letter “X” in an English word often corresponds to the consonant cluster /ks/ when pronounced.

4 System Description

Our system is motivated linguistically and for practical reasons. On the one hand, transliteration is to render a source name in a phonemically similar way in a target language, and syllable is an important concept in pronunciation. According to Ladefoged (2006), for alphabetic writing systems, syllables are systematically split into their components. A syllable is composed of an optional onset containing consonants and a mandatory rhyme. The rhyme comprises a mandatory nucleus containing vowels and an optional coda containing consonants. English has complex onsets and codas, whereas Mandarin Chinese has simple onsets and only allows nasal consonants in the coda. 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. Hence, the complex onset in the English syllable “STEIN” (as in Figure 1) violates the onset constraints in

Chinese and is therefore resolved into two Chinese syllables as “斯坦” (*si1 tan3)*[1]. Hence syllable is apparently the proper basic unit for machine transliteration.

On the other hand, during transliteration, people tend not to re-invent the wheel for a similar chunk of syllables in the source name. The examples in Table 1 illustrate this observation. As seen, “JACOB” is consistently rendered as “雅各布” (*ya3 ge4 bu4*) and “STEIN” as “斯坦” (*si1 tan3*) when they appear as part of different names. So based on the concept of translation memory, if a larger chunk can be matched, transliteration becomes easier and less uncertain. In this way, the context embedding a syllable is incorporated, and it might also reduce error propagation in the pipeline during syllabification and phoneme mapping.

With the above linguistic and practical considerations, a syllable-based Maximum Matching approach is thus adopted, and the following subsections explain the steps involved.

Table 1: Examples of Transliteration by Chunks

English	**Chinese**	**Hanyu Pinyin**
JACOB	雅各布	*ya3 ge4 bu4*
JACOBS	雅各布斯	*ya3 ge4 bu4 si1*
JACOBSEN	雅各布森	*ya3 ge4 bu4 sen1*
JACOBSTEIN	雅各布斯坦	*ya3 ge4 bu4 si1 tan3*
ARENSTEIN	阿伦斯坦	*a4 lun2 si1 tan3*
BARTENSTEIN	巴滕斯坦	*ba1 teng2 si1 tan3*
DUBERSTEIN	杜伯斯坦	*du4 bo2 si1 tan3*

4.1 Syllabification

The English names in the training data and development data were first syllabified with the Onset First Principle. According to Katamba (1989), the principle suggests that syllable-initial consonants are first maximised to the extent consistent with the syllable structure conditions of the language in question, followed by the maximisation of syllable-final consonants.

In English, written symbols do not necessarily bear a one-to-one relationship with phonological segments. So in practice, with reference to common phonics patterns, we drew up a list of possible onsets containing graphemic units which may correspond to simple phonemes (e.g. “CH”, “TH”) or complex onsets (e.g. “PL”, “STR”) to be used in syllabification.

During syllabification, all vowels were first marked as nucleus (N). The longest acceptable consonant sequences on the left of the vowels were then marked as onset (O), and finally all remaining consonants were marked as coda (C). From left to right, syllables are marked for each longest matching chain of ONC, ON, NC, or N. The top half of Figure 1 illustrates these steps with the English name “JACOBSTEIN”.

Subsequently, the syllable chain was subject to sub-syllabification considering the difference in phonotactics between English and Chinese. In particular, Chinese syllables have no complex onsets and only allow nasal consonants for codas. So if the syllabification step produces fewer English syllables than Chinese syllables, the sub-syllabification process will try to expand the English syllables, with the number of syllables checked after each expansion. At any point if the English syllables outnumber the Chinese ones, the sub-syllabification process will try to contract the English syllables.

The expansion process will thus follow the order of precedence below:

(1) From left to right, split up complex onsets. For example, “STEIN” is split up into “S/TEIN”.

[1] Transcriptions of the Chinese characters in Hanyu Pinyin are given in italics throughout this paper.

(2) From right to left, split up complex codas or separate coda from nucleus if the coda is not available in the target language. For example, “COB” is sub-syllabified as “CO/B”.

(3) From right to left, separate liquids and glides (“L”, “R”, “W”) from the nucleus if the Chinese rendition has “尔” (*er3*) or “夫” (*fu1*) in it. For example, with the pair “MINKOWSKI” and “明科夫斯基” (*ming2 ke1 fu1 si1 ji1*), initial syllabification produces three syllables, “MIN/KOW/SKI”. During sub-syllabification, “SKI” will be split into “S/KI” with (1) above, but the English side is still one syllable short. So “KOW” will be split into “KO/W” in the next expansion.

(4) From left to right, expand diphthongs as necessary. For example, diphthongs like “IA” will be split up as in “A/ME/LI/A”.

The contraction process will follow the order of precedence below:

(1) Contract the name-initial “M/C”, if any, with the following syllable.

(2) From right to left, contract nasals, liquids and glides followed by “E” with the previous syllable. For example, “AALLIBONE” for “阿利本” (*a4 li4 ben3*) will be initially syllabified as “AA/LLI/BO/NE”, which will then be contracted to “AA/LLI/BONE”.

The middle part of Figure 1 illustrates the sub-syllabification process.

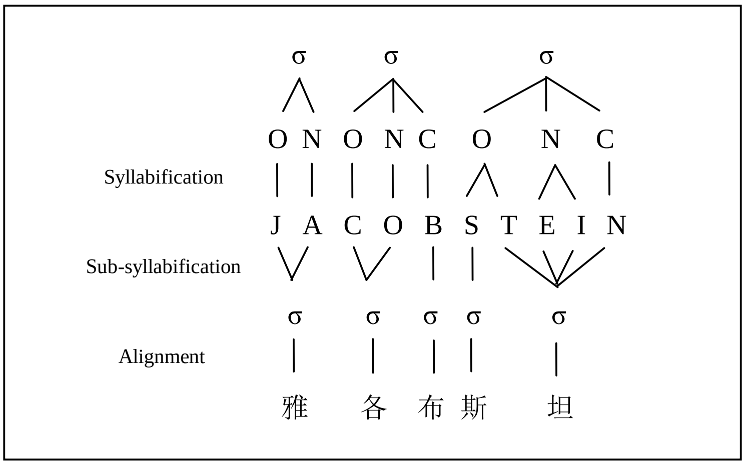

Figure 1: Syllabification and Alignment

4.2 Alignment

Upon syllabification and sub-syllabification, if the number of English syllables equals the number of Chinese syllables, alignment can be done directly in a one-to-one manner. Otherwise the following heuristics would be used to attempt some complex alignments.

(1) As long as Chinese syllables still outnumber English syllables, the next English syllable with four or more letters or starting with two different consonants will absorb two Chinese syllables, assuming such long segments are actually pronounced as two syllables. For example, “A/L/THOU/SE” does not have enough syllables to align with its Chinese rendition “奥尔特豪斯” (*ao4 er3 te4 hao2 si1*), so “THOU” will be forced to take up two Chinese syllables “特豪” (*te4 hao2*).

(2) At any point, if the remaining Chinese syllables fall short of English syllables, the rest will be aligned as a whole without further breaking into syllables. For example, “YON/GE” will simply be aligned with the Chinese name “扬” (*yang2*).

The bottom part of Figure 1 shows the alignment step.

4.3 Lexicon Production

Based on the aligned names, segment pairs of various syllable lengths were extracted to produce a bilingual lexicon as follows:

```
For i = 1 to n (# of aligned segment pairs)
    For j = i to n
        Extract segment-i to segment-j
    Next j
Next i
```

Hence for the aligned name in Figure 1, the following segment pairs will enter into the lexicon:

JA	雅 (*ya3*)
JACO	雅各 (*ya3 ge4*)
JACOB	雅各布 (*ya3 ge4 bu4*)
JACOBS	雅各布斯 (*ya3 ge4 bu4 si1*)
JACOBSTEIN	雅各布斯坦 (*ya3 ge4 bu4 si1 tan3*)
CO	各 (*ge4*)
COB	各布 (*ge4 bu4*)
COBS	各布斯 (*ge4 bu4 si1*)
COBSTEIN	各布斯坦 (*ge4 bu4 si1 tan3*)
B	布 (*bu4*)
BS	布斯 (*bu4 si1*)
BSTEIN	布斯坦 (*bu4 si1 tan3*)
S	斯 (*si1*)
STEIN	斯坦 (*si1 tan3*)
TEIN	坦 (*tan3*)

Note that we use “segment pairs” instead of “syllable pairs” here as the alignment may involve one or more syllables on either side.

4.4 Maximum Matching

During transliteration, an English source name was first syllabified using the syllabification and the first two sub-syllabification steps described above. The name was then segmented using Maximum Matching with the lexicon. It can be done in either direction. The matching was syllable-based, unless even the shortest syllable cannot be matched with the lexicon. In that case the syllable would be matched as a string of characters.

Although the current study focuses on English-to-Chinese transliteration, the same procedures can actually be applied to Chinese-to-English back transliteration as well, since the lexicon contains bilingual segment pairs, and can be looked up bi-directionally. Maximum Matching can be done with the English segments or Chinese segments accordingly. Chinese source names do not need particular syllabification as Chinese characters are syllabic.

4.5 Candidate Generation and Ranking

With the segmented source name, target candidates were generated by looking up the lexicon for each segment and its rendition(s) in the target language. The candidates can subsequently be ranked by different ways. In the current study, we tested ranking with unigram and contextual n-gram probabilities of the segment pairs. Figure 2 shows an example of Maximum Matching and candidate generation for the English name “MARKSTEIN”.

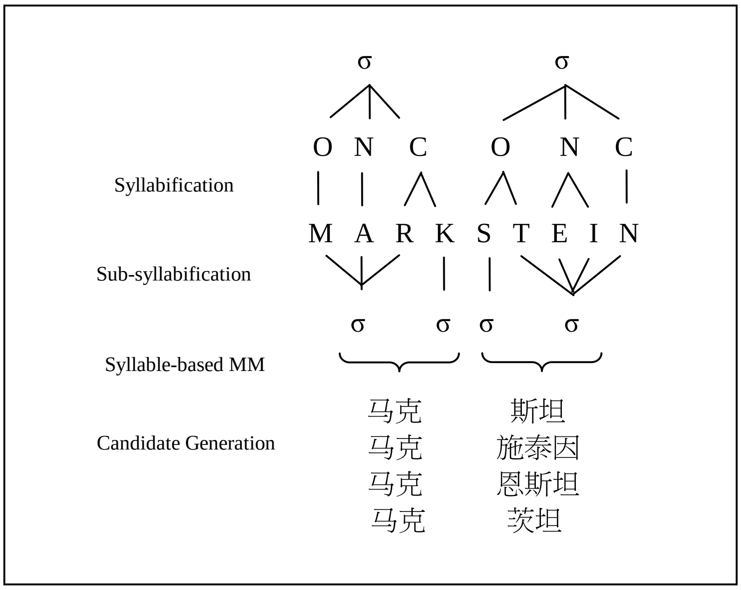

Figure 2: Maximum Matching and Candidate Generation

5 Experiments and Results

The proposed system was run on the test data, with variation on the lexicon containing bilingual segment pairs. One trial used all segment pairs in the lexicon, another used only segment pairs with frequency two or above, and the third trial used only segment pairs with frequency five or above. The lexicon thus contained about 94K entries, 16K entries, and 4K entries respectively. Table 2 and Table 3 show the system performance for Forward Maximum Matching and Backward Maximum Matching respectively, using the different lexicons. The ranking of candidates was by segment pair unigram probabilities. The different trials are indicated by "F" for Forward Maximum Matching and "B" for Backward Maximum Matching, "1/2/5" for the different lexicons, and "U" and "C" for unigram and contextual n-grams respectively. The evaluation metrics follow the definitions in the whitepaper of the most recent NEWS shared task on transliteration generation (Zhang *et al.*, 2011). ACC, Mean F-score, MRR, and MAP_{ref} refer to the Word Accuracy in Top-1, Fuzziness in Top-1, Mean Reciprocal Rank, and precision in the n-best candidates, respectively.

Table 2: Results with Forward Maximum Matching

Metric	F1U	F2U	F5U
ACC	0.6115	0.5974	0.5470
Mean F-score	0.8442	0.8357	0.8144
MRR	0.6887	0.6763	0.6221
MAP_{ref}	0.6115	0.5974	0.5470

Table 3: Results with Backward Maximum Matching

Metric	B1U	B2U	B5U
ACC	0.5967	0.5891	0.5425
Mean F-score	0.8378	0.8301	0.8063
MRR	0.6794	0.6674	0.6167
MAP_{ref}	0.5967	0.5891	0.5425

Table 2 and Table 3 suggest two points of interest. On the one hand, Forward Maximum Matching apparently performs slightly better than Backward Maximum Matching. On the other hand, many of the low frequency segment pairs of various syllable lengths are in fact very useful in transliteration, as is evident from the drop in performance when the lexicon was reduced in the second and third trials.

Studying the results more deeply, it is observed that Forward Maximum Matching and Backward Maximum Matching could be complementary. Many a time matching from both directions might give the same results, but at other times one might produce the favoured segmentation and thus correct transliteration at higher ranks, and vice versa. Take "ABERSFELLER" as an example, Forward Maximum Matching segments it into "ABER/SFE/LLER" and F1U gives "阿伯斯菲勒" (*a4 bo2 si1 fei1 le4*) as the first candidate, while Backward Maximum Matching segments it into "A/BERS/FELLER" and B1U gives "阿伯斯费勒" (*a4 bo2 si1 fei4 le4*) as the first candidate which is the answer in the reference set. So we tried to use both directions of matching in combination, evaluating the unigram rankings of all candidates together. In order to take the context into consideration, instead of segment pair unigrams, we also tested combined Maximum Matching with contextual n-gram probabilities for ranking. Since n-gram probabilities like bigrams for segment pairs directly are expected to suffer from severe data sparseness with larger chunks of syllables, we use the relative position of the segment pair as context, that is, whether the segment pair is at the beginning, middle or end of a name. Table 4 shows the results alongside those reported in other studies using the same datasets. FB1U and FB1C refer to our bi-directional Maximum Matching with unigram ranking and contextual n-gram ranking respectively. SAG refers to the non-parametric Bayesian learning method with Synchronous Adaptor Grammars (with syllable grammars) reported in Huang *et al.* (2011), and JSCM refers to the Joint Source-Channel Model, first used by Li *et al.* (2004) and implemented by Huang *et al.* (2011) on the same datasets as a baseline. Only ACC and Mean F-score were reported. NEWS refers to the best performing system in the NEWS 2009 English-to-Chinese task according to Li *et al.* (2009), and it is a phrase-based statistical machine transliteration system. Our results with bi-directional Maximum Matching are obviously better than Forward or Backward alone, and the results from ranking with simple positional information (that is, FB1C) are comparable to those of SAG and JSCM, though they are not as good as the best performing system in NEWS. Further analysis of the results suggests that while our method has the advantage of being simpler and intuitively appealing, it nevertheless has the following limitations.

Table 4: Results with Bi-directional Maximum Matching Compared with Other Studies

Metric	FB1U	FB1C	SAG	JSCM	NEWS
ACC	0.6119	0.6288	0.666	0.666	0.731
Mean F-score	0.8455	0.8549	0.866	0.857	0.895
MRR	0.7031	0.7176	N/A	N/A	0.812
MAP_{ref}	0.6119	0.6288	N/A	N/A	0.713

The first has to do with the nature of the method. By looking at the largest possible transliterated chunks, we inevitably miss the variations of the smaller units. The test item "ABARBANEL" is a good illustration of this problem. Both directions of Maximum Matching give "A/BARBA/NEL" as the segmentation, and "阿巴巴内尔" (*a4 ba1 ba1 nei4 er3*) is ranked first among other candidates. This is in fact a more than satisfactory transliteration of the name by itself, but the answer in the reference set is "阿巴伯内尔" (*a4 ba1 bo2 nei4 er3*). It happens that "BARBA" had no alignment with "巴伯" (*ba1 bo2*) in the training data, although the simpler syllable "BA" did with "伯" (*bo2*). Transliterating with the longest match has therefore precluded some other possible combinations from being evaluated, and this may account for the relatively lower ACC but similar mean F-score, compared with other systems. Interestingly, however, “巴伯” (*ba1 bo2*) had been aligned with “BARBAR”, “BARBE”, “BARBER”, and “BARBE” in the training data.

The second limitation is related to the origin of the English names. In this study we have based on the phonotactics of English for syllabification, which may not always work well with names from other origins. Take “DESLONGCHAMPS” as an example. As a French name, the “S” in “DES” and “CHAMPS” is usually not pronounced. Although we have the pair “DES/德” (*de2*) in the lexicon, its probability is much lower than the other pair “DES/德斯” (*de2 si1*). On the other hand, “CHAMPS” and “LONGCHAMPS” are not found in the training data, while “LONGCHAMP” is found, with a correspondence to “朗香” (*lang3 xiang1*) but not the expected “朗尚” (*lang3 shang4*). Unseen syllables or segments are thus another major source of errors.

6 Conclusion

We have thus proposed a syllable-based Maximum Matching method for English-Chinese name transliteration generation. Although it may not outperform systems based on phrase-based translation models, it is in fact effective and comparable to many other popular methods and has the advantage of being intuitively simple, linguistically motivated, and easy to implement. Future directions will be along possible improvements of our method, by means of more flexible manipulation of the syllable units, better detection of the origins of English names, as well as incorporation of more re-ranking criteria.

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.

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.

Huang, Y., M. Zhang and C.L. Tan. 2011. Nonparametric Bayesian Machine Transliteration with Synchronous Adaptor Grammars. In *Proceedings of ACL-HLT 2011: Short Papers*, Portland, Oregon, pp.534-539.

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.

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.

Ladefoged, P. 2006. *A Course in Phonetics*. Thomson Wadsworth.

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.

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.

Zhang, M., A. Kumaran and H. Li. 2011. Whitepaper of NEWS 2011 Shared Task on Machine Transliteration. In *Proceedings of NEWS 2011*, Chiang Mai, Thailand.