

# English-Chinese Personal Name Transliteration by 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

This paper reports on our participation in the NEWS 2011 shared task on transliteration generation with a syllable-based Backward Maximum Matching system. The system 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 official results suggest that our system could potentially be improved with a re-ranking module for English-to-Chinese transliteration, while its performance on Chinese-to-English back transliteration reached the state of the art.

## 1 Introduction

This paper describes our system participating in two tracks of the NEWS 2011 shared task on transliteration generation, including English-to-Chinese transliteration (EnCh) and Chinese-to-English back transliteration (ChEn).

Our system is essentially a syllable-based Backward Maximum Matching (BMM) system, which works bi-directionally for EnCh and ChEn. 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 BMM with syllables as the basic units. Target candidates were generated by looking up the bilingual lexicon and ranked by unigram probabilities.

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 and its performance reported in Section 4, followed by future work and conclusion in Section 5.

## 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. Wutiw-watchai 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, 2009a).

## 3 Datasets

The transliteration data provided by the shared task organiser are mostly based on name pairs from Xinhua News Agency (1992). For EnCh, there are 37,753 English-Chinese name pairs in the training set, 2,802 pairs in the development set, and another 2,000 English names in the test set. For ChEn, there are 28,678 Chinese-English name pairs in the training set, 2,719 pairs in the development set, and another 2,266 Chinese 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.

In the current study, we focused entirely on personal name transliteration. The small proportion of place names in the data was not handled. Most of them contain multiple English words or otherwise are not entirely phonemically rendered in Chinese (e.g. Africa 非洲, transcribed as *fei1 zhou1* in Hanyu Pinyin). They are better dealt with by a specific lookup table of place names, but since we only participated in the standard runs and did not use any external resources, those cases were practically ignored.

All English names are in 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*). 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.

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

Table 1. Examples of Transliteration by Chunks

### 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.

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.

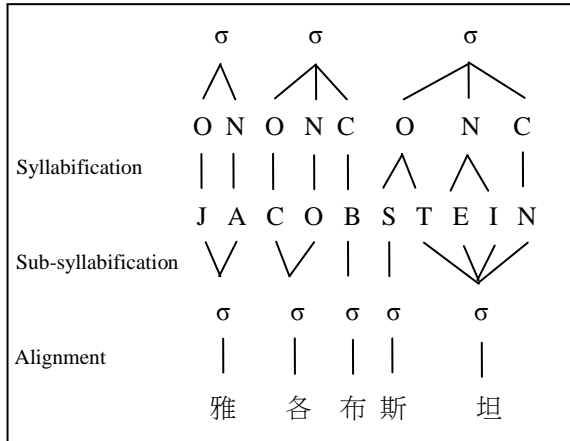


Figure 1. Syllabification and Alignment

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”.

(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.

## 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 some heuristics would be used to attempt some complex alignments. 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*). 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*), and 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 Backward Maximum Matching

During transliteration, an English source name was first syllabified using the syllabification and

sub-syllabification procedures described above, except the contraction part. The name was then segmented using Backward Maximum Matching with the lexicon. 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.

The same procedures were applied to EnCh and ChEn, as 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. In the current study, the candidates were simply ranked by unigram probabilities. Figure 2 shows an example of Maximum Matching and candidate generation.

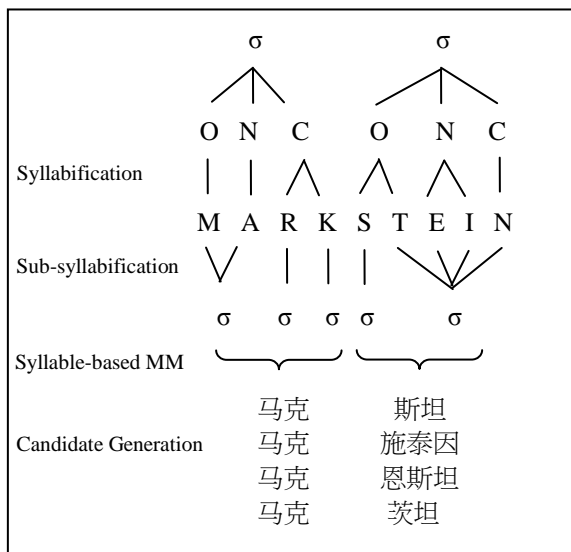


Figure 2. Max Matching and Candidate Generation

#### 4.6 Official Results

Table 2 and Table 3 show the official results for the two standard runs we submitted and the best system in EnCh and ChEn respectively. The first run used segment pairs with frequency two or above, and the second run used those with frequency five or above. The evaluation metrics follow the definitions in the whitepaper of the shared task (Zhang *et al.*, 2011).

The performance of our system on EnCh is in the mid-range, and re-ranking with n-gram features is apparently important. For instance, VE/夫 (*fu1*) is more frequent than VE/维 (*wei2*), but

the former is often restricted to the end of a name. This would not be realised for now, unless a longer segment can be matched, e.g. "VELO" could only be matched on single syllables, so "夫洛" (*fu1 luo4*) came before "维洛" (*wei2 luo4*), but "VELASCO" found a longer match with "维拉斯科" (*wei2 la1 si1 ke1*) as the first candidate. This suggests that Maximum Matching is useful, but re-ranking is needed for better performance.

ChEn is apparently more difficult, and scores are lower in general. Nevertheless, our system came in the top three, giving even better Mean F-score and MRR than the system with the best ACC. The more severe graphemic ambiguity for ChEn may make it a more difficult task. According to Kwong (2009b), on average one English segment (syllable) has 1.7 Chinese renditions but one Chinese character can be mapped to 10 different English segments. Another major problem for ChEn is unseen characters and the spelling conventions of English or other European languages. For example, "云" (*yun2*) was not found in the training and development data and therefore "云格" (*yun2 ge2*) could not be properly back transliterated. Also, some candidates end up with triple consonants which are obviously not acceptable in English and should be avoided.

Metric	Run 1	Run 2	Best
ACC	0.305	0.285	0.348
Mean F-score	0.672	0.660	0.700
MRR	0.378	0.349	0.462
MAP <sub>ref</sub>	0.297	0.276	0.342

Table 2. Official EnCh Results on Test Data

Metric	Run 1	Run 2	Best
ACC	0.155	0.154	0.167
Mean F-score	0.766	0.757	0.765
MRR	0.215	0.206	0.202
MAP <sub>ref</sub>	0.155	0.154	0.167

Table 3. Official ChEn Results on Test Data

#### 5 Future Work and Conclusion

Thus the performance of our approach on EnCh has room for improvement, possibly with a re-ranking module, and that on ChEn is close to the state of the art. Forward and Backward Maximum Matching could potentially be used together to better handle overlapping ambiguity so as not to miss other possible candidates.

#### Acknowledgements

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

## References

- Dobrovolsky, M. and Katamba, F. (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 Sumita E. (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., Zhang, M. and Tan, C.L. (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 Graehl, J. (1998) Machine Transliteration. *Computational Linguistics*, 24(4):599-612.
- Kwong, O.Y. (2009a) Homophones and Tonal Patterns in English-Chinese Transliteration. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, Singapore, pp.21-24.
- Kwong, O.Y. (2009b) Graphemic Approximation of Phonological Context for English-Chinese Transliteration. In *Proceedings of NEWS 2009*, Singapore, pp.186-193.
- Ladefoged, P. (2006) *A Course in Phonetics*. Thomson Wadsworth.
- Li, H., Zhang, M. and Su, J. (2004) A Joint Source-Channel Model for Machine Transliteration. In *Proceedings of ACL 2004*, 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 ACL 2007*, Prague, Czech Republic, pp.120-127.
- Li, H., Kumaran, A., Pervouchine, V. and Zhang, M. (2009) Report of NEWS 2009 Machine Transliteration Shared task. In *Proceedings of NEWS 2009*, Singapore.
- Li, H., Kumaran, A., Zhang, M. and Pervouchine, V. (2010) Report of NEWS 2010 Transliteration Generation Shared Task. In *Proceedings of NEWS 2010*, Uppsala, Sweden.
- 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.
- Shishtla, P., Ganesh, V.S., Sethuramalingam, S. and Varma, V. (2009) A language-independent transliteration schema using character aligned models. In *Proceedings of NEWS 2009*, Singapore.
- Song, Y., Kit, C. and Zhao, H. (2010) Reranking with multiple features for better transliteration. In *Proceedings of NEWS 2010*, Uppsala, Sweden.
- 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.
- Wutiw WATCHAI, C. and Thangthai, A. (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., Kumaran, A. and Li, H. (2011) Whitepaper of NEWS 2011 Shared Task on Machine Transliteration. In *Proceedings of NEWS 2011*, Chiang Mai, Thailand.