Corpus Effects on the Evaluation of Automated Transliteration Systems

Sarvnaz Karimi **Andrew Turpin** **Falk Scholer**
School of Computer Science and Information Technology
RMIT University, GPO Box 2476V, Melbourne 3001, Australia
{sarvnaz,aht,fscholer}@cs.rmit.edu.au

Abstract

Most current machine transliteration systems employ a corpus of known source-target word pairs to train their system, and typically evaluate their systems on a similar corpus. In this paper we explore the performance of transliteration systems on corpora that are varied in a controlled way. In particular, we control the number, and prior language knowledge of human transliterators used to construct the corpora, and the origin of the source words that make up the corpora. We find that the word accuracy of automated transliteration systems can vary by up to 30% (in absolute terms) depending on the corpus on which they are run. We conclude that at least four human transliterators should be used to construct corpora for evaluating automated transliteration systems; and that although absolute word accuracy metrics may not translate across corpora, the relative rankings of system performance remains stable across differing corpora.

1 Introduction

Machine transliteration is the process of transforming a word written in a source language into a word in a target language without the aid of a bilingual dictionary. Word pronunciation is preserved, as far as possible, but the script used to render the target word is different from that of the source language. Transliteration is applied to proper nouns and out-of-vocabulary terms as part of machine translation and cross-lingual information retrieval (CLIR) (AbdulJaleel and Larkey, 2003; Pirkola et al., 2006). Several transliteration methods are reported in the literature for a variety of languages, with their performance being evaluated on multilingual corpora. Source-target pairs are either extracted from bilingual documents or dictionaries (AbdulJaleel and Larkey, 2003; Bilac and Tanaka, 2005; Oh and Choi, 2006; Zelenko and Aone, 2006), or gathered explicitly from human transliterators (Al-Onaizan and Knight, 2002; Zelenko and Aone, 2006). Some evaluations of transliteration methods depend on a single unique transliteration for each source word, while others take multiple target words for a single source word into account. In their work on transliterating English to Persian, Karimi et al. (2006) observed that the content of the corpus used for evaluating systems could have dramatic affects on the reported accuracy of methods.

The effects of corpus composition on the evaluation of transliteration systems has not been specifically studied, with only implicit experiments or claims made in the literature such as introducing the effects of different transliteration models (AbdulJaleel and Larkey, 2003), language families (Lindén, 2005) or application based (CLIR) evaluation (Pirkola et al., 2006). In this paper, we report our experiments designed to explicitly examine the effect that varying the underlying corpus used in both training and testing systems has on transliteration accuracy. Specifically, we vary the number of human transliterators that are used to construct the corpus; and the origin of the English words used in the corpus.

Our experiments show that the word accuracy of automated transliteration systems can vary by up to 30% (in absolute terms), depending on the corpus used. Despite the wide range of absolute values

Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 640–647,
Prague, Czech Republic, June 2007. ©2007 Association for Computational Linguistics

in performance, the ranking of our two transliteration systems was preserved on all corpora. We also find that a human's confidence in the language from which they are transliterating can affect the corpus in such a way that word accuracy rates are altered.

2 Background

Machine transliteration methods are divided into grapheme-based (AbdulJaleel and Larkey, 2003; Lindén, 2005), phoneme-based (Jung et al., 2000; Virga and Khudanpur, 2003) and combined techniques (Bilac and Tanaka, 2005; Oh and Choi, 2006). Grapheme-based methods derive transformation rules for character combinations in the source text from a training data set, while phoneme-based methods use an intermediate phonetic transformation. In this paper, we use two grapheme-based methods for English to Persian transliteration. During a training phase, both methods derive rules for transforming character combinations (*segments*) in the source language into character combinations in the target language with some probability.

During transliteration, the source word s_i is segmented and rules are chosen and applied to each segment according to heuristics. The probability of a resulting word is the product of the probabilities of the applied rules. The result is a list of target words sorted by their associated probabilities, L^i.

The first system we use (SYS-1) is an n-gram approach that uses the last character of the previous source segment to condition the choice of the rule for the current source segment. This system has been shown to outperform other n-gram based methods for English to Persian transliteration (Karimi et al., 2006).

The second system we employ (SYS-2) makes use of some explicit knowledge of our chosen language pair, English and Persian, and is also on the collapsed-vowel scheme presented by Karimi et al. (2006). In particular, it exploits the tendency for runs of English vowels to be collapsed into a single Persian character, or perhaps omitted from the Persian altogether. As such, segments are chosen based on surrounding consonants and vowels. The full details of this system are not important for this paper; here we focus on the performance evaluation of systems, not the systems themselves.

2.1 System Evaluation

In order to evaluate the list L^i of target words produced by a transliteration system for source word s_i, a test corpus is constructed. The test corpus consists of a source word, s_i, and a list of possible target words $\{t_{ij}\}$, where $1 \leq j \leq d_i$, the number of distinct target words for source word s_i. Associated with each t_{ij} is a count n_{ij} which is the number of human transliterators who transliterated s_i into t_{ij}.

Often the test corpus is a proportion of a larger corpus, the remainder of which has been used for training the system's rule base. In this work we adopt the standard ten-fold cross validation technique for all of our results, where 90% of a corpus is used for training and 10% for testing. The process is repeated ten times, and the mean result taken. Forthwith, we use the term corpus to refer to the single corpus from which both training and test sets are drawn in this fashion.

Once the corpus is decided upon, a metric to measure the system's accuracy is required. The appropriate metric depends on the scenario in which the transliteration system is to be used. For example, in a machine translation application where only one target word can be inserted in the text to represent a source word, it is important that the word at the top of the system generated list of target words (by definition the most probable) is one of the words generated by a human in the corpus. More formally, the first word generated for source word s_i, L^i_1, must be one of $t_{ij}, 1 \leq j \leq d_i$. It may even be desirable that this is the target word most commonly used for this source word; that is, $L^i_1 = t_{ij}$ such that $n_{ij} \geq n_{ik}$, for all $1 \leq k \leq d_i$. Alternately, in a CLIR application, all variants of a source word might be required. For example, if a user searches for an English term "Tom" in Persian documents, the search engine should try and locate documents that contain both "تام" (3 letters:م-ا-ت) and " تم"(2 letters: م-ت), two possible transliterations of "Tom" that would be generated by human transliterators. In this case, a metric that counts the number of t_{ij} that appear in the top d_i elements of the system generated list, L^i, might be appropriate.

In this paper we focus on the "Top-1" case, where it is important for the most probable target word generated by the system, L^i_1 to be either the most popular t_{ij} (labeled the *Majority*, with ties broken arbitrarily), or just one of the t_{ij}'s (labeled *Uniform* because all possible transliterations are equally rewarded). A third scheme (labeled *Weighted*) is also possible where the reward for t_{ij} appearing as L_1^i is $n_{ij}/\sum_{j=1}^{d_i} n_{ij}$; here, each target word is given a weight proportional to how often a human transliterator chose that target word. Due to space considerations, we focus on the first two variants only.

In general, there are two commonly used metrics for transliteration evaluation: word accuracy (WA) and character accuracy (CA) (Hall and Dowling, 1980). In all of our experiments, CA based metrics closely mirrored WA based metrics, and so conclusions drawn from the data would be the same whether WA metrics or CA metrics were used. Hence we only discuss and report WA based metrics in this paper.

For each source word in the test corpus of K words, word accuracy calculates the percentage of correctly transliterated terms. Hence for the majority case, where every source word in the corpus only has one target word, the word accuracy is defined as

$$MWA = |\{s_i | L_1^i = t_{i1}, 1 \le i \le K\}|/K,$$

and for the *Uniform* case, where every target variant is included with equal weight in the corpus, the word accuracy is defined as

$$UWA = |\{s_i | L_1^i \in \{t_{ij}\}, 1 \le i \le K, 1 \le j \le d_i\}|/K.$$

2.2 Human Evaluation

To evaluate the level of agreement between transliterators, we use an agreement measure based on Mun and Eye (2004).

For any source word s_i, there are d_i different transliterations made by the n_i human transliterators ($n_i = \sum_{j=1}^{d_i} n_{ij}$, where n_{ij} is the number of times source word s_i was transliterated into target word t_{ij}). When any two transliterators agree on the same target word, there are two agreements being made: transliterator one agrees with transliterator two, and vice versa. In general, therefore, the total number of agreements made on source word s_i is $\sum_{j=1}^{d_i} n_{ij}(n_{ij}-1)$. Hence the total number of actual agreements made on the entire corpus of K words is

$$A_{act} = \sum_{i=1}^{K} \sum_{j=1}^{d_i} n_{ij}(n_{ij}-1).$$

The total number of possible agreements (that is, when all human transliterators agree on a single target word for each source word), is

$$A_{poss} = \sum_{i=1}^{K} n_i(n_i-1).$$

The proportion of overall agreement is therefore

$$P_A = \frac{A_{act}}{A_{poss}}.$$

2.3 Corpora

Seven transliterators (T1, T2, ..., T7: all native Persian speakers from Iran) were recruited to transliterate 1500 proper names that we provided. The names were taken from lists of names written in English on English Web sites. Five hundred of these names also appeared in lists of names on Arabic Web sites, and five hundred on Dutch name lists. The transliterators were not told of the origin of each word. The entire corpus, therefore, was easily separated into three sub-corpora of 500 words each based on the origin of each word. To distinguish these collections, we use E_7, A_7 and D_7 to denote the English, Arabic and Dutch sub-corpora, respectively. The whole 1500 word corpus is referred to as EDA_7.

Dutch and Arabic were chosen with an assumption that most Iranian Persian speakers have little knowledge of Dutch, while their familiarity with Arabic should be in the second rank after English. All of the participants held at least a Bachelors degree. Table 1 summarizes the information about the transliterators and their perception of the given task. Participants were asked to scale the difficulty of the transliteration of each sub-corpus, indicated as a scale from 1 (*hard*) to 3 (*easy*). Similarly, the participants' confidence in performing the task was rated from 1 (*no confidence*) to 3 (*quite confident*). The level of familiarity with second languages was also reported based on a scale of zero (*not familiar*) to 3 (*excellent knowledge*).

The information provided by participants confirms our assumption of transliterators knowledge of second languages: high familiarity with English, some knowledge of Arabic, and little or no prior knowledge of Dutch. Also, the majority of them found the transliteration of English terms of medium difficulty, Dutch was considered mostly hard, and Arabic as easy to medium.

	Second Language Knowledge				Difficulty,Confidence		
Transliterator	English	Dutch	Arabic	Other	English	Dutch	Arabic
1	2	0	1	-	1,1	1,2	2,3
2	2	0	2	-	2,2	2,3	3,3
3	2	0	1	-	2,2	1,2	2,2
4	2	0	1	-	2,2	2,1	3,3
5	2	0	2	Turkish	2,2	1,1	3,2
6	2	0	1	-	2,2	1,1	3,3
7	2	0	1	-	2,2	1,1	2,2

Table 1: Transliterator's language knowledge (0=not familiar to 3=excellent knowledge), perception of difficulty (1=hard to 3=easy) and confidence (1=no confidence to 3=quite confident) in creating the corpus.

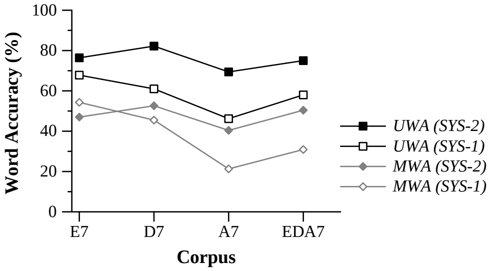

Figure 1: Comparison of the two evaluation metrics using the two systems on four corpora. (Lines were added for clarity, and do not represent data points.)

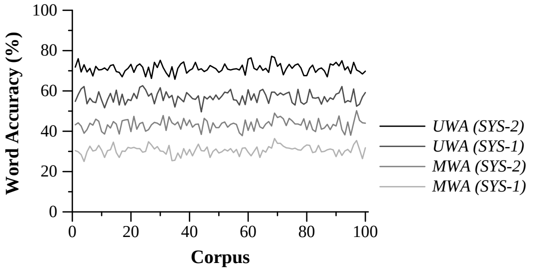

Figure 2: Comparison of the two evaluation metrics using the two systems on 100 randomly generated sub-corpora.

3 Results

Figure 1 shows the values of UWA and MWA for E_7, A_7, D_7 and EDA_7 using the two transliteration systems. Immediately obvious is that varying the corpora (x-axis) results in different values for word accuracy, whether by the *UWA* or *MWA* method. For example, if you chose to evaluate SYS-2 with the *UWA* metric on the D_7 corpus, you would obtain a result of 82%, but if you chose to evaluate it with the A_7 corpus you would receive a result of only 73%. This makes comparing systems that report results obtained on different corpora very difficult. Encouragingly, however, SYS-2 consistently outperforms the SYS-1 on all corpora for both metrics except *MWA* on *E*7. This implies that ranking system performance on the same corpus most likely yields a system ranking that is transferable to other corpora. To further investigate this, we randomly extracted 100 corpora of 500 word pairs from EDA_7 and ran the two systems on them and evaluated the results using both *MWA* and *UWA*. Both of the measures ranked the systems consistently using all these corpora (Figure 2).

As expected, the *UWA* metric is consistently higher than the *MWA* metric; it allows for the top transliteration to appear in any of the possible variants for that word in the corpus, unlike the *MWA* metric which insists upon a single target word. For example, for the E_7 corpus using the SYS-2 approach, *UWA* is 76.4% and *MWA* is 47.0%.

Each of the three sub-corpora can be further divided based on the seven individual transliterators, in different combinations. That is, construct a sub-corpus from T1's transliterations, T2's, and so on; then take all combinations of two transliterators, then three, and so on. In general we can construct ^7C_r such corpora from r transliterators in this fashion, all of which have 500 source words, but may have between one to seven different transliterations for each of those words.

Figure 3 shows the *MWA* for these sub-corpora. The x-axis shows the number of transliterators used to form the sub-corpora. For example, when $x = 3$, the performance figures plotted are achieved on corpora when taking all triples of the seven transliterator's transliterations.

From the boxplots it can be seen that performance varies considerably when the number of transliterators used to determine a majority vote is varied.

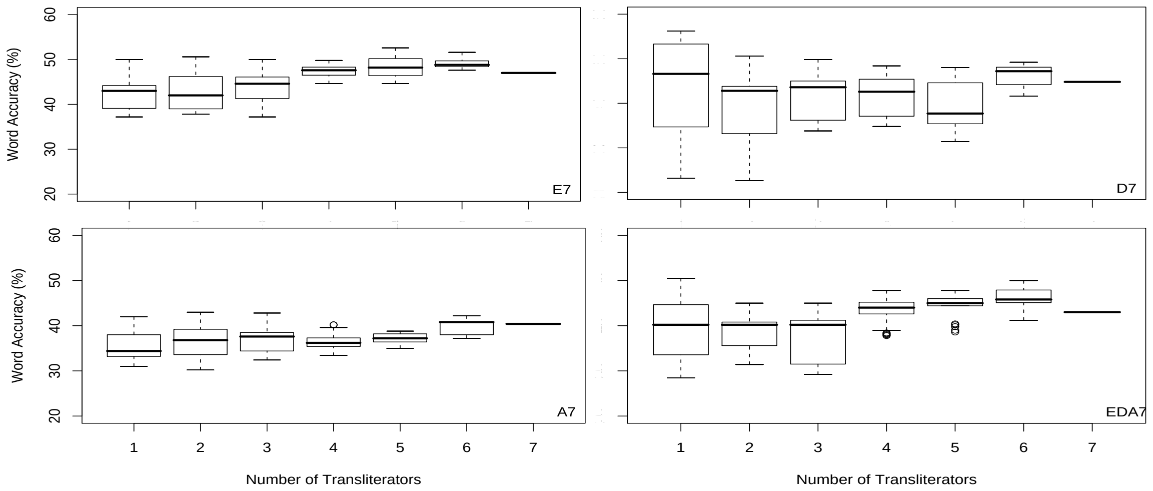

Figure 3: Performance on sub-corpora derived by combining the number of transliterators shown on the x-axis. Boxes show the 25th and 75th percentile of the *MWA* for all ${}^{7}C_{x}$ combinations of transliterators using SYS-2, with whiskers showing extreme values.

However, the changes do not follow a fixed trend across the languages. For E_7, the range of accuracies achieved is high when only two or three transliterators are involved, ranging from 37.0% to 50.6% in SYS-2 method and from 33.8% to 48.0% in SYS-1 (not shown) when only two transliterators' data are available. When more than three transliterators are used, the range of performance is noticeably smaller. Hence if at least four transliterators are used, then it is more likely that a system's *MWA* will be stable. This finding is supported by Papineni et al. (2002) who recommend that four people should be used for collecting judgments for machine translation experiments.

The corpora derived from A_7 show consistent median increases as the number of transliterators increases, but the median accuracy is lower than for other languages. The D_7 collection does not show any stable results until at least six transliterator's are used.

The results indicate that creating a collection used for the evaluation of transliteration systems, based on a "gold standard" created by only one human transliterator may lead to word accuracy results that could show a 10% absolute difference compared to results on a corpus derived using a different transliterator. This is evidenced by the leftmost box in each panel of the figure which has a wide range of results.

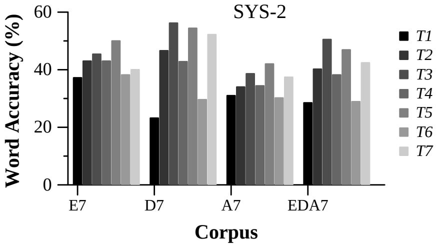

Figure 4: Word accuracy on the sub-corpora using only a single transliterator's transliterations.

Figure 4 shows this box in more detail for each collection, plotting the word accuracy for each user for all sub-corpora for SYS-2. The accuracy achieved varies significantly between transliterators; for example, for E_7 collections, word accuracy varies from 37.2% for T1 to 50.0% for T5. This variance is more obvious for the D_7 dataset where the difference ranges from 23.2% for $T1$ to 56.2% for $T3$. Origin language also has an effect: accuracy for the Arabic collection (A_7) is generally less than that of English (E_7). The Dutch collection (D_7), shows an unstable trend across transliterators. In other words, accuracy differs in a narrower range for Arabic and English, but in wider range for Dutch.

This is likely due to the fact that most transliterators found Dutch a difficult language to work with, as reported in Table 1.

3.1 Transliterator Consistency

To investigate the effect of invididual transliterator consistency on system accuracy, we consider the number of Persian characters used by each transliterator on each sub-corpus, and the average number of rules generated by SYS-2 on the ten training sets derived in the ten-fold cross validation process, which are shown in Table 2. For example, when transliterating words from E_7 into Persian, T3 only ever used 21 out of 32 characters available in the Persian alphabet; T7, on the other hand, used 24 different Persian characters. It is expected that an increase in number of characters or rules provides more "noise" for the automated system, hence may lead to lower accuracy. Superficially the opposite seems true for rules: the mean number of rules generated by SYS-2 is much higher for the EDA_7 corpus than for the A_7 corpus, and yet Figure 1 shows that word accuracy is higher on the EDA_7 corpus. A correlation test, however, reveals that there is no significant relationship between either the number of characters used, nor the number of rules generated, and the resulting word accuracy of SYS-2 (Spearman correlation, $p = 0.09$ (characters) and $p = 0.98$ (rules)).

A better indication of "noise" in the corpus may be given by the consistency with which a transliterator applies a certain rule. For example, a large number of rules generated from a particular transliterator's corpus may not be problematic if many of the rules get applied with a low probability. If, on the other hand, there were many rules with approximately equal probabilities, the system may have difficulty distinguishing when to apply some rules, and not others. One way to quantify this effect is to compute the *self entropy* of the rule distribution for each segment in the corpus for an individual. If p_{ij} is the probability of applying rule $1 \leq j \leq m$ when confronted with source segment i, then $H_i = -\sum_{j=1}^{m} p_{ij} \log_2 p_{ij}$ is the entropy of the probability distribution for that rule. H is maximized when the probabilities p_{ij} are all equal, and minimized when the probabilities are very skewed (Shannon, 1948). As an example, consider the rules: $t \rightarrow <$ت$,0.5 >$, $t \rightarrow <$ط$,0.3 >$ and $t \rightarrow <$د$,0.2 >$; for which $H_{\mathrm{t}} = 0.79$.

The expected entropy can be used to obtain a single entropy value over the whole corpus,

$$E = -\sum_{i=1}^{R} \frac{f_i}{S} H_i,$$

where H_i is the entropy of the rule probabilities for segment i, R is the total number of segments, f_i is the frequency with which segment i occurs at any position in all source words in the corpus, and S is the sum of all f_i.

The expected entropy for each transliterator is shown in Figure 5, separated by corpus. Comparison of this graph with Figure 4 shows that generally transliterators that have used rules inconsistently generate a corpus that leads to low accuracy for the systems. For example, T1 who has the lowest accuracy for all the collections in both methods, also has the highest expected entropy of rules for all the collections. For the E_7 collection, the maximum accuracy of 50.0%, belongs to $T5$ who has the minimum expected entropy. The same applies to the D_7 collection, where the maximum accuracy of 56.2% and the minimum expected entropy both belong to $T3$. These observations are confirmed by a statistically significant Spearman correlation between expected rule entropy and word accuracy ($r = -0.54, p = 0.003$). Therefore, the consistency with which transliterators employ their own internal rules in developing a corpus has a direct effect on system performance measures.

3.2 Inter-Transliterator Agreement and Perceived Difficulty

Here we present various agreement proportions (P_A from Section 2.2), which give a measure of consistency in the corpora across all users, as opposed to the entropy measure which gives a consistency measure for a single user. For E_7, P_A was 33.6%, for A_7 it was 33.3% and for D_7, agreement was 15.5%. In general, humans agree less than 33% of the time when transliterating English to Persian.

In addition, we examined agreement among transliterators based on their perception of the task difficulty shown in Table 1. For A_7, agreement among those who found the task *easy* was higher (22.3%) than those who found it in *medium* level

	E_7		D_7		A_7		EDA_7	
	Char	Rules	Char	Rules	Char	Rules	Char	Rules
T1	23	523	23	623	28	330	31	1075
T2	22	487	25	550	29	304	32	956
T3	21	466	20	500	28	280	31	870
T4	23	497	22	524	28	307	30	956
T5	21	492	22	508	28	296	29	896
T6	24	493	21	563	25	313	29	968
T7	24	495	21	529	28	299	30	952
Mean	23	493	22	542	28	304	30	953

Table 2: Number of characters used and rules generated using SYS-2, per transliterator.

(18.8%). P_A is 12.0% for those who found the D_7 collection *hard* to transliterate; while the six transliterators who found the E_7 collection difficulty *medium* had $P_A = 30.2\%$. Hence, the harder participants rated the transliteration task, the lower the agreement scores tend to be for the derived corpus.

Finally, in Table 3 we show word accuracy results for the two systems on corpora derived from transliterators grouped by perceived level of difficulty on A_7. It is readily apparent that SYS-2 outperforms SYS-1 on the corpus comprised of human transliterations from people who saw the task as easy with both word accuracy metrics; the relative improvement of over 50% is statistically significant (paired t-test on ten-fold cross validation runs). However, on the corpus composed of transliterations that were perceived as more difficult, "Medium", the advantage of SYS-2 is significantly eroded, but is still statistically significant for *UWA*. Here again, using only one transliteration, *MWA*, did not distinguish the performance of each system.

4 Discussion

We have evaluated two English to Persian transliteration systems on a variety of controlled corpora using evaluation metrics that appear in previous transliteration studies. Varying the evaluation corpus in a controlled fashion has revealed several interesting facts.

We report that human agreement on the English to Persian transliteration task is about 33%. The effect that this level of disagreement on the evaluation of systems has, can be seen in Figure 4, where word accuracy is computed on corpora derived from single transliterators. Accuracy can vary by up to 30% in absolute terms depending on the transliterator chosen. To our knowledge, this is the first paper to report human agreement, and examine its effects on transliteration accuracy.

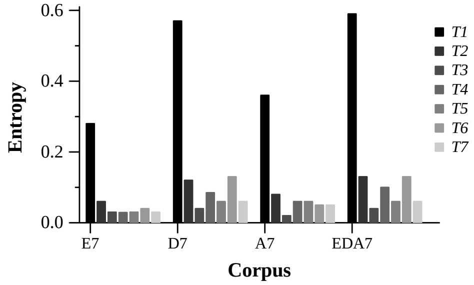

Figure 5: Entropy of the generated segments based on the collections created by different transliterators.

In order to alleviate some of these effects on the stability of word accuracy measures across corpora, we recommend that at least four transliterators are used to construct a corpus. Figure 3 shows that constructing a corpus with four or more transliterators, the range of possible word accuracies achieved is less than that of using fewer transliterators.

Some past studies do not use more than a single target word for every source word in the corpus (Bilac and Tanaka, 2005; Oh and Choi, 2006). Our results indicate that it is unlikely that these results would translate onto a corpus other than the one used in these studies, except in rare cases where human transliterators are in 100% agreement for a given language pair.

Given the nature of the English language, an English corpus can contain English words from a variety of different origins. In this study we have used English words from an Arabic and Dutch origin to show that word accuracy of the systems can vary by up to 25% (in absolute terms) depending on the origin of English words in the corpus, as demonstrated in Figure 1.

In addition to computing agreement, we also in-

				Relative	
	Perception	SYS-1	SYS-2	Improvement (%)	
UWA	Easy	33.4	55.4	54.4	($p < 0.001$)
	Medium	44.6	48.4	8.52	($p < 0.001$)
MWA	Easy	23.2	36.2	56.0	($p < 0.001$)
	Medium	30.6	37.4	22.2	($p = 0.038$)

Table 3: System performance when A_7 is split into sub-corpora based on transliterators perception of the task (Easy or Medium).

vestigated the transliterator's perception of difficulty of the transliteration task with the ensuing word accuracy of the systems. Interestingly, when using corpora built from transliterators that perceive the task to be easy, there is a large difference in the word accuracy between the two systems, but on corpora built from transliterators who perceive the task to be more difficult, the gap between the systems narrows. Hence, a corpus applied for evaluation of transliteration should either be made carefully with transliterators with a variety of backgrounds, or should be large enough and be gathered from various sources so as to simulate different expectations of its expected non-homogeneous users.

The self entropy of rule probability distributions derived by the automated transliteration system can be used to measure the consistency with which individual transliterators apply their own rules in constructing a corpus. It was demonstrated that when systems are evaluated on corpora built by transliterators who are less consistent in their application of transliteration rules, word accuracy is reduced.

Given the large variations in system accuracy that are demonstrated by the varying corpora used in this study, we recommend that extreme care be taken when constructing corpora for evaluating transliteration systems. Studies should also give details of their corpora that would allow any of the effects observed in this paper to be taken into account.

Acknowledgments

This work was supported in part by the Australian government IPRS program (SK).

References

Nasreen AbdulJaleel and Leah S. Larkey. 2003. Statistical transliteration for English-Arabic cross-language information retrieval. In *Conference on Information and Knowledge Management*, pages 139–146.

Yaser Al-Onaizan and Kevin Knight. 2002. Machine transliteration of names in Arabic text. In *Proceedings of the ACL-02 workshop on Computational approaches to semitic languages*, pages 1–13.

Slaven Bilac and Hozumi Tanaka. 2005. Direct combination of spelling and pronunciation information for robust back-transliteration. In *Conference on Computational Linguistics and Intelligent Text Processing*, pages 413–424.

Patrick A. V. Hall and Geoff R. Dowling. 1980. Approximate string matching. *ACM Computing Survey*, 12(4):381–402.

Sung Young Jung, Sung Lim Hong, and Eunok Paek. 2000. An English to Korean transliteration model of extended Markov window. In *Conference on Computational Linguistics*, pages 383–389.

Sarvnaz Karimi, Andrew Turpin, and Falk Scholer. 2006. English to Persian transliteration. In *String Processing and Information Retrieval*, pages 255–266.

Krister Lindén. 2005. Multilingual modeling of cross-lingual spelling variants. *Information Retrieval*, 9(3):295–310.

Eun Young Mun and Alexander Von Eye, 2004. *Analyzing Rater Agreement: Manifest Variable Methods*. Lawrence Erlbaum Associates.

Jong-Hoon Oh and Key-Sun Choi. 2006. An ensemble of transliteration models for information retrieval. *Information Processing Management*, 42(4):980–1002.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *The 40th Annual Meeting of Association for Computational Linguistics*, pages 311–318.

Ari Pirkola, Jarmo Toivonen, Heikki Keskustalo, and Kalervo Järvelin. 2006. FITE-TRT: a high quality translation technique for OOV words. In *Proceedings of the 2006 ACM Symposium on Applied Computing*, pages 1043–1049.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423.

Paola Virga and Sanjeev Khudanpur. 2003. Transliteration of proper names in cross-language applications. In *ACM SIGIR Conference on Research and Development on Information Retrieval*, pages 365–366.

Dmitry Zelenko and Chinatsu Aone. 2006. Discriminative methods for transliteration. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 612–617.