Proceedings of the 21st North American Conference on Chinese Linguistics (NACCL-21). 2009. Volume 2.
Edited by Yun Xiao. Smithfield, Rhode Island: Bryant University. Pages 375-387.

Romanization Patterns in Chinese as Evidenced by a Personal Name Corpus

Tom McClive
STG, Inc.

Chinese has a lengthy and often non-uniform history of transliteration and Romanization patterns, from systems such as Wade-Giles and Pinyin to more extemporized attempts. One domain of language severely resists conformity—personal names. The multiple romanized variants of a Chinese name stem from historical source patterns and personal choices. Romanization standards are often inconsistent or unobserved, and may diverge from existing orthographic intuitions. This study shows that a sizable corpus of personal names in romanized form is integral to any attempts at reconciliation and record linkage; its strength is shown in the confluence among statistical methods, human factors, and linguistic knowledge. The results constitute a type of surface form grammar, one based on the corpus romanization patterns rather than underlying forms and sources.

1. Introduction

Record linkage is the term for one of the newer yet now-widespread applied applications of computational linguistics. Through methods including synonym lists and letter comparisons, an algorithm can match personal name records containing variants such as Tom and Thomas, as well as misspellings or previous-unknown variants such Thhomas or Tohmas.

Without a truthed corpus to corroborate the process, the success rate of any linkage method is unverifiable. Conventional wisdom may cause a plurality of agreement, yet opinions will still vary. My own name can be used as an example. If one compares Tom McClive to Thomas Mac Cleavon, those familiar with Western names would agree that Tom and Thomas are closely-used variants of the same name, and that two records using those names could refer to the same person. As to the surname, the Mc and Mac are both a variant of the Scottish-origin prefix loosely meaning "child of", clearly corresponding, and Cleavon can be shown to historically be a variant of Clive.

Record linkage still is not like a mathematical equation where x = y; one cannot say for sure that a Tom and a Thomas are the same person, but we can assign a certain degree of confidence to a yes or no answer. The confidence, difficult to quantify, would still not be without human intuition; those more familiar with the names may feel that the

surname comparison in question is likely not the same name (I would certainly feel this way), while those unfamiliar with the names may find them perfectly acceptable variants.

Questions of sameness in written Chinese names can mostly be solved by looking at the characters, but the task becomes quite complicated when dealing with the romanized forms. Comparing the romanized names of the martial artists Bruce Lee and Jet Li gives one nothing but the representative sounds. Since most of the world does not use Chinese characters, and most computer records do not contain them, their Romanized versions are the forms that are dissembled.

The dialects, and perhaps different languages, that fall under the colloquial categorization umbrella of "Chinese" have a lengthy and often non-uniform history of transliteration and Romanization patterns, from popular, largely accepted systems such as Yale, Wade-Giles, and Pinyin to more extemporized attempts.

Bruce Lee and Jet Li indeed happen to have the same character for their surname (李), but this is not at all evident by their spelling, which clearly comes from two different eras and two different transliteration traditions. LI is more of a pinyin-style construction, while LEE is a more Western-influenced fossilization. The name Robert E. Lee clearly is not connected historically to either men, but also shares the same surface form surname, and any record linkage would start a surname comparison by connecting the group.

2. Challenges of Chinese Romanization

One particular challenge with romanization in monosyllabic East Asian languages such as Chinese is the consistently increased semantic weight each letter carries. By design, a contrived romanization system does not contain any extraneous symbols. Most have no silent letters or adjustments for regional or personal variation. The silent "H" in "Thomas" would not be allowed in a designed system for English, as the TH combination would overlap with the established TH digraph for the voiceless interdental fricative, unless it somehow is needed to contrast with, say, an unaspirated [t] sound.

This semantic weight demands that each letter present in a transliterated surface form be initially accorded an assumed status of deliberateness. That extra H, we first assume, must mean something, though this is certainly not always the case. A difference of one letter between two words can make a lexical distinction in any representative system, but the letters in shorter words carry more weight. One complication for any language's romanization is that there are usually competing systems used, making the letter differences harder to judge. Consider:

ZANG = TZANG
ZANG ≠ ZHANG

The surface forms ZANG and TZANG can mean the same word, through two different romanization systems who represent the phoneme [dz] in different ways, even

though the letter in question, T, would not seem to be incidental. However, another type of one-letter difference between the forms ZANG and ZHANG makes them into two different words even though that letter in question, H, is historically often merely ornamental.

The historical and generally accepted variants on a common name like THOMAS stem from geographic distribution across an area, with some changes coming from efforts to conform to local phonological patterns, and some arbitrary, perhaps even capricious, spelling changes. One could still look at a list of Tomas, Tomash, Tomaj, Tomac, Thoma, Tomaso, Tomaq, Tuomo, Tuomas, Tomek, and Tamhas, along with the nickname and variants rule creations such as Tom, Thom, Tommy, and Tommie, and still perhaps judge them to be the same name, although some geographical variants such as the English John being the Scottish Ian may not be as recognizable. But the variants of a name that has been romanized can come from entirely different sources. The Chinese name CAI may also be realized as Tsai, Zai, Tsay, Tsair, Tzai, Tzay, and Tsae, among other forms.

The variants of CAI listed above have few common attributes; they share a single letter, A, all possess an onset, and most of them are an open syllable. That's little to connect them. Many reference works for Chinese names try to list common variables, but as with the romanization system itself, there is no way to enforce or ensure these lists and the divisions between them. Listing of variations may ignore the human factor, saying that ZHÀO with a fourth tone may have one list of variants, while ZHĀO with a first tone may have a different list.

The process of romanization, or any transliteration in general, has its own set of en suite issues. They include such challenges as:

(1) A different inventory of sounds between two languages.

(2) A common inability to perform a direct A → B type of transliteration. It is often the case that one symbol cannot be replaced by one other symbol. Even if a common pattern exists, the surface forms may differ due to the phonological environment.

One example comes from Korean where the symbol ㄱ is realized as a voiceless velar stop [k] in one environment, as voiced [g] in another, and as nasal [ŋ] in yet another, thus being transliterated as "k", "g", or "ng" depending on its position.

ㄱ = as /k/ in 고려 /g/ in 적용 /ng/ in 직면

Another example comes from Japanese, from a more logographic writing system analogous to Chinese. The character 田 is pronounced [ta] when in the beginning of a word, and as a voiced [da] when at the end, such as in Tanaka and Yamada.

(3) Imperfect alphabet symbol inventory. There is no mass consensus on representation.

(4) Adoption of dormant letters (such as Q and X), digraphs or trigraphs, and diacritics. Sounds that cannot easily be represented in romanization through the most commonly used letters are often assigned such lesser-used letters such as Q and X, or are represented through digraphs or trigraphs, or even diacritics.

One example comes from Thai, where the Royal Thai Government System of transliteration decrees that the Thai vowel เ◌ือะ should be transliterated as UEA, a vowel combination that no native English speaker could correctly pronounce by sight.

Beyond these general linguistic difficulties, there are the human factors that can lead to orthographic variation, the reasons that individual, non-native transliterations will choose certain realizations. Some of these issues, often leading to particular forms with Chinese, are:

(1) Not knowing the phoneme inventory. The difference between the pinyin realizations CH and Q may not be discernable to non-native speakers without a minimal pair, and thus someone may hear QING but write CHING. The same holds true for other pairs such as ZH~J and SH~X.

(2) Trying to represent each sound. With a retroflex consonant and a semi-vowel, the pinyin SHI may sound more like a SHIR to a non-native speaker.

(3) Conforming to native orthography. Even without trying to represent each perceived sound nuance, non-native speakers will often use their own perceived native orthography pattern, especially with vowels, leading to such forms as SHIH.

(4) Wedded to fossilized forms. Anyone who has been to a Chinese restaurant in America has seen such dishes as Szechwan beef or General Tso's Chicken. These forms, like the LEE realization of the name LI, have become fossilized and popularized and are unlikely to go away.

(5) OCR or transcription errors. Instances of a form such as CHANS may be determined to be CHANG, with the G~S switch attributed to either an OCR error or some other type of transcription inaccuracy.

(6) Concatenation and segmentation. The convention of how to write a Chinese given name has changed over the years, and still varies according to location. A given name with two syllables YA and HONG is usually written concatenated as YAHONG in China, as YA-HONG in Taiwan, and as YA HONG in Hong Kong and other Chinese communities such as Singapore. When performing record linkage, it is of course more helpful to have consistency. The form that is preferred is a segmented YA HONG, to be able to work with each element separately.

(7) Forcing non-western names into the canonical western name format. Dr. Sun Yat-Sen might find his name written as YAT S. SUN while living in the West. Many times the second element of the given name is treated in the same way as a Western-style middle name.

(8) Hypercorrection. Many romanization systems have spelling conventions that violate the perceived rules of the target language. An orthographically correct name such as HSIN may be perceived by an English speaker to be a misspelling of SHIN.

(9) Finally, people recording names make the general type of mistakes and typos with Chinese names as they would with any others. The occurrence of mistakes for non-Chinese speakers is likely to be higher, as the letter patterns are not familiar.

What occurs from this list of nine phenomenon is that we are left with a grammar of surface forms. The romanization patterns that occur in Chinese names are their own corpus, without reliable mappings or underlying forms, and without any way to get back to those items. With some form of underlying grammar, HSIN and SHIN can be judged as different lexical entries. With a grammar of surface forms, they cannot. There may still be a high degree of probability for difference, but there is also some probability degree for sameness.

With a grammar of surface forms, even positing an underlying form is problematic, perhaps even unhelpful. Knowing the commonly associated underlying Chinese characters for particular surface forms doesn't conclusively show sameness. All probability judgments must be made based on knowledge of the romanization systems and the human factors.

3. Challenges of Personal Names

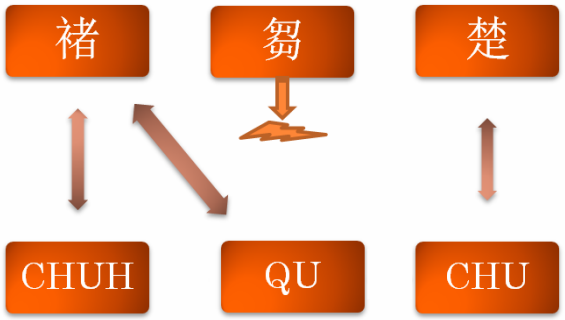
Personal names sit at the intersection of orthography and personal choice. The multiple Romanized variants of a Chinese name, such as Li, Lee, Le, and Yi, stem from historical source patterns and personal choices, much in the same way that English can have Cathy, Kathy, and Kathie. Personal names tend to break the rules of the language, in their spelling conventions and formation. My own surname, McClive, breaks English phonology rules with its sonority-bending four consonants in a row MCCL orthography-bound onset.

The canonical Chinese name has three elements: one element serving as a surname (in other words, a family name that can be passed down through generations), and two serving as the individual-identifying given name, although one-element given names have become more popular in recent generations. Each element corresponds to one written Chinese character and thus one syllable. An adopted Western name is sometimes appended to the given name.

One unintended consequence from romanizing a Chinese name is that the order may be reversed, in accordance with Western conventions. The normal surname-given name order of a name such as LI YAHONG is often written as YAHONG LI. While many, if not most, of these reserved names can be identified as to which elements are the surname and given name, a more ambiguous constructions such as LI ZHANG is not so easily identified. Each element is plausible both as a surname and as a given name.

4. Surface Realization Splits, Mergers, and Variants

To illustrate the surface form grammar, it is not difficult in Chinese to find three characters with very similar phonetic realizations, minimal triplets. Their representative romanization forms, from perhaps different transliteration systems, clearly do not form a one-to-one correlation. The character 褚 may have a surface form of CHUH at times but also appear as QU, a split. The character 楚 may appear as CHU, not overlapping the other characters, while the character 蜀 may not even be traceable to a particular surface form in a corpus. It is also not difficult to imagine a merger of two characters being realized by the same surface form.



According to the parameters set by the Pinyin romanization system, the above three characters should all be written with the letter combination CHU (ignoring tonal diacritics for now), but it is possible that only one character will be traced to a CHU surface form. The many-to-one relationship that the romanization system projects (characters to surface form) is already a deviant from the one-to-one that a general population might imagine in a transliteration system; the imperfect mappings demonstrated above further complicate the issue.

Consider the character 蔡, with a sound pattern [ts^ha^i], transliterated as CAI according to Pinyin. With an initial sound that is not naturally an initial in English, and with a diphthong vowel, its romanized form could vary even more than the relatively simpler CHU above. Even with the same simple syllable structure, the romanized form could vary more. If we assume there could be:

(1) Three onset possibilities: C, TS, TZ

(2) Three vowel possibilities: AI, AY, AE
(3) One possible coda ending: R (thus, two possibilities: R or nothing)

These combinations would create 3*3*2 = 18 possibly variants, with forms such as CAI and TSAER. Moreover, the variants tend to me more untidy in several senses. They have less alphabetic letters in common, which would affect such comparative techniques as edit distance, and they have more substantial consonant variation, which would affect a method such as Soundex keys.

Moreover, the standard four tonal markers from such systems as Pinyin are very often lost in name copra. Though the majority of the world's computers are now able to employ diacritics in their character sets, social practices dictate that they are very often not entered, and once they are lost, their lexical distinction value is gone. Unless the context is clear, it is impossible to tell if CAI is CÀI or CÁI.

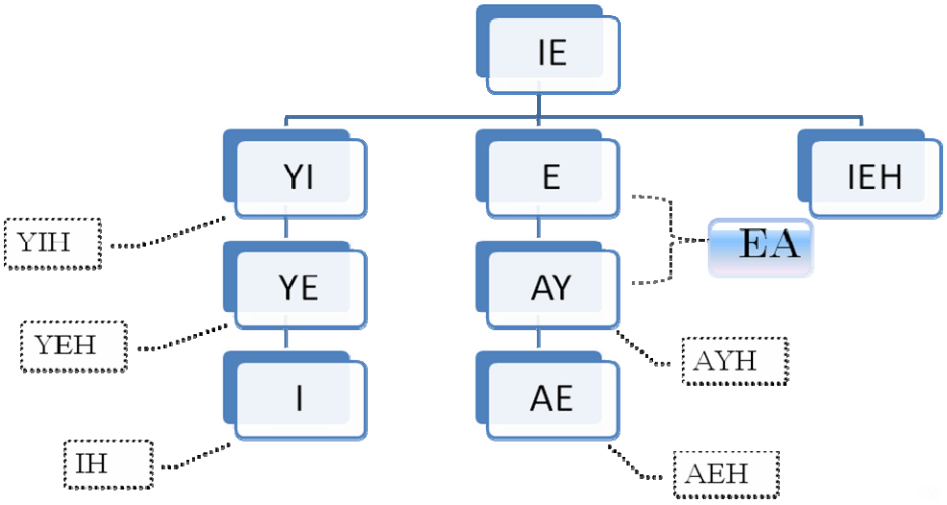
The eighteen possible variants above multiply when a complete personal name (given name and surname) is considered, instead of just a single name segment. Consider a standard-form three-element Chinese name with a syllable structure of CV.CV.CVC that has these qualities:

(1) The initial and final consonants each have two variants.
(2) The internal consonants may or may not be doubled.
(3) Each vowel has two variants.

This creates a pattern like: $[C_1C_2][V_1V_2]CC?[V_1V_2]CC?[V_1V_2][C_1C_2]$

At each of the seven positions, there are two choice points, which yield 2^7, or 128 possibilities. For longer names, or names in which there are more alternations or conditions, the number of variants is even greater.

As an example of how a single variation path can be linked to others, consider the vowel combination IE. It may have a set of three variants {YI, E, IEH}, and some of those variants may have their own set of variants, such as {YI, YE, I} and {E, AY, AE}. Furthermore, there may be an overlapping set {E, AY, EA}, and almost all of them may have an optional H ending. The resulting complicated tree would look like:



Attempts at such mappings may naturally lead to positing rules for linkage of the variant forms. It may be easy to determine that YE and YI are variants, or YI and EI. Yet if we put forward that YI and E are variants, does the same hold true for YI and AE, or for AEH and IEH? If connections are made this way, the suggestion that E and I are variants, from the tree above, would logically extend to minimal name pairs such as XENG and XING, a bold implication.

An inverse method to ferret out larger variant patterns is to look at traditional variants using whole name elements, but this also can lead to the type of overreaching seen above. We could examine two groups of traditional variants, based on known historical variants of common name elements as evidenced by direct character mapping:

WANG, WONG, ONG
HUANG, HWANG, WONG

The first line would suggest that a W initial is compatible with a null initial, and that A is transposable with O. The second group would suggest that H and W initials are interchangeable, and a vowel variant grouping of {UH, A, O}, all suggestions that are also potentially overreaching.

The eventual solution may involve a detailing collecting of each variant grouping, to control exactly how each variant linkage can work. Two groupings could be concocted, labeled group numbers 101 and 102, whereby variants are defined by being intergroup but not crossing group boundaries:

101 SHIH, HSI, SHI, SHII, SHYI, XI
102 SHIH, SHI, SHY, SHYH, SHYR, SHYY

Thus, SHIH can match SHYI and can match XI, but XI and SHI cannot match each other. This would be an exact, but quite tedious, method of defining variants.

5. Use of a Name Corpus

One of the advantages of a potentially large corpus, with hundreds of thousands of personal names, is confidence in the presence of surface forms. If it happens enough in the world, it is probably in the corpus. One can posit surface forms then use the corpus to check for their existence. We are able to return to our CAI example and check for variants by listing possible alternative consonants {TS, TZ, Z} and vowels {AY, AIR, AE}, then checking for their name part frequency. If the occurrence looks somewhat like the chart below:

Variant	Count	Frequency
CAI	5225	0.82963
TSAI	544	0.08638
ZAI	499	0.07923
TSAY	11	0.00174
TSAIR	3	0.00047
TZAI	8	0.00127
TZAY	8	0.00031
TSAE	0	0.00000
CAY	5	0.00079
CAE	1	0.00015
TZAE	0	0.00000

At this juncture, a cutoff point is chosen, perhaps after the third variant or perhaps including the next few most populous variants, and the remainder are discarded as being statistically insignificant to be considered. These name elements of course are representative of surface forms present, and not necessarily equal to each other, yet they show the distribution of possible variation, both in whole form and, possibly considered, in individual phone transliteration. A TZ initial, for instance, may be perceived to be somewhat archaic by today's romanization schemes and standards, yet its presence in the corpus shows that it is not yet entirely absent in the world.

As a practical application, consider the challenge of segmenting Chinese name parts. Most Chinese from China who have a two-element (two character) given name write the romanized version as a concatenated form, such as YAHONG or QINGYING. With record linkage, it is highly advantageous to segment these names back into their two elements before working with them. With a name such as QINGYING, the division seems obvious, QING+YING, but with YAHONG there could be two candidates,

YA+HONG and YAH+ONG. Consider the following list of Chinese given names and their possible segmentation candidates:

a.	XIAOOU	[['XIAO', 'OU'], ['XIA', 'OOU'], ['XIAOOU']]
b.	HAIANG	[['HAI', 'ANG'], ['HA', 'IANG'], ['HAIANG']]
c.	ZHENGAI	[['ZHENG', 'AI'], ['ZHEN', 'GAI'], ['ZHE', 'NGAI']]
d.	CHAKWANG	[['CHAK', 'WANG'], ['CHA', 'KWANG']]
e.	CHAWONG	[['CHAW', 'ONG'], ['CHA', 'WONG']]
f.	CHIAHAO	[['CHIAH', 'AO'], ['CHIA', 'HAO']]
g.	CHIHSIEN	[['CHIH', 'SIEN'], ['CHI', 'HSIEN']]
h.	GUANEN	[['GUAN', 'EN'], ['GUA', 'NEN']]
i.	LAIMUNG	[['LAI', 'MUNG'], ['LAIM', 'UNG']]
j.	MINHAN	[['MINH', 'AN'], ['MIN', 'HAN'], ['MI', 'NHAN']]
k.	SHINAE	[['SHIN', 'AE'], ['SHI', 'NAE']]

The candidates for (g) above include a non-standard CHIH and a possible Wade-Giles produced HSIEN. The strength of a corpus is that it allows us to compile a large list of possible variant candidates, using them in ways such as assigning degrees of probability or confidence. If we check the frequency occurrence of the four element candidates involved in the two segmentation scenarios, we might find that we can support the HSIEN candidacy more strongly than the SIEN. A frequency distribution confidence could also help us lean toward discouraging the XIAOOU and HAIANG candidates in (a) and (b), respectively.

The advantages of a corpus are rarely stand-alone. For a more holistic approach, these frequency confidences would need to be combined with other tools such as knowledge of Chinese syllable structure and of linguistics in general. Our knowledge of Chinese tells us that the NGAI candidate of (c) is unlikely because of its initial, likewise with the NHAM of (j).

Still, while knowledge of Chinese and Linguistics would also help eliminate candidates such as XIAOOU in (a), referenced above as a strength of using a corpus, corpus usage would further lend confidence to preference of segmentation scenarios when the candidates are not distinguished by linguistic form. The third segmentation candidate for (c) may be eliminated because of the NGAI element in the third scenario, but the first two scenarios are both viable in form, syllable structure, and sonority. It may, of course, be impossible to confidentially posit only one segmentation scenario (likely in this case), but the existence of a corpus again may allow us to assign confidence degrees to likely scenarios, by confirming that the ZHENG+AI patterns, or even the ZHENG and AI elements considered separately, occur far more frequently than the ZHEN+GAI pattern and elements.

As another example of the confluence of methods that leads to romanization comparisons, consider an individual case of comparing two name elements, CHWEANG and JWAEN.

Our first setting uses edit distance, a computational linguistics comparative method that compares the strings letter by letter, and seeks to answer the question of how far apart the two strings are by examining the steps needed to change one into the other (Levenshtein, 1966, Wagner and Fischer, 1974). It assigns penalty-type points for operations of letter deletion, insertion, substitution, and reversal (here, all are 1.0 except for reversal at 1.5), then sees how many points must be used to turn one name into the other and normalizes that figure across the lengths of the two strings.

For our two strings of CHWEANG and JWAEN, the resultant grid of the edit distance process would look like this:

Longer string: CHWEANG length: 7
Shorter string: JWAEN length: 5

		-1	0	1	2	3	4	5	6	7
				C	H	W	E	A	N	G
-1		***	***	***	***	***	***	***	***	***
0		***	0	1	2	3	4	5	6	7
1	J	***	1	1	2	3	4	5	6	7
2	W	***	2	2	2	2	3	4	5	6
3	A	***	3	3	3	3	3	3	4	5
4	E	***	4	4	4	4	3	3.5	4	5
5	N	***	5	5	5	5	4	4	3.5	4.5

The edit distance process returns an integer between zero and 1.0. The result in this case is 0.357 (somewhat rounded), a not-good score, and certainly nothing that would pass any system's internal threshold to be considered a viable match.

In other words, some strictly computational methods would fail us in this comparison case. This form of edit distance does not take into account the linguistic structure of the string, the romanization pattern similarities, or the phonetic similarities.

Let us consider a better method that takes into consideration some of the romanization and phonetic properties of the letters, along with the syllable structure. One advantage of an East Asian language such as Mandarin Chinese is that each word is only one syllable, and thus the initials, the vowel cluster, and the finals can be considered separately.

If we use a syllable parser, each element can be compared individually, with a degree of similarity assigned for each, and then either normalized or averaged across the strings:

Beginning Consonants	Glide: W or Y	Vowels	Ending Consonants
CH	**W**	**AE**	**NG**
J	**W**	**EA**	**N**
Pretty good	Same	Perhaps	Not likely

With this method, a CH and J comparison must be considered as the phonemic minimal pair that they are, along with considering the effect of this difference upon Chinese (phonemes, and thus a possible lexical distinction). The glides W are the same. The AE and EA vowels are a reversal, a potential but not probable match. The N and NG endings have one letter in common but are distinct phonemes in Chinese.

To make the operation simple, an arithmetic assignment of 0.8 for the "pretty good" CH~J status, 1.0 for the glide status of "same", 0.5 for the AE~EA "perhaps", and 0.2 for the NG~N "not likely" gives us an 0.625 average result. These scores could be weighed or refined to produce an even more accurate comparison number of course, but it seems clear already that this basic 0.625 result is more appropriate for a CHWAENG~JWEAN comparison than the 0.357 outcome that edit distance alone produces.

6. Conclusions

Size matters. Having a large corpus allows most romanization patterns to become evident; without a critical mass of names, the lack of a particular surface pattern could not be assumed to be significant. With a large enough sampling, there is a certain degree of confidence that if a particular surface form happens in the world, it will likely be present in the corpus. Furthermore, the strength of a corpus is that employing frequency statistics alone on romanization patterns often is more reliable that using linguistic knowledge.

Humans matter. The human factor cannot be discounted in analyzing data. The surface form results of various intuition, guesswork, and imperfect knowledge still show up, factors independent from orthographic patterns or linguistic knowledge.

Linguistics still matters. Despite the advantages of a sizable corpus and perceptions of human nature, we still need linguistic knowledge. Computational methods such as edit distance often fall somewhat short. Many techniques are often based on math or statistics, and we usually find that we need more than that.

Finally, we must still admit that there is no absolute value to surface forms. Without further information, it is impossible to verify that TCHANG and CHANG map to the same underlying sound pattern, much less the same Chinese character, lexical entry, and individual person. Surface forms usually are not accompanied by a truthed corpus. The idea of a variant, and any rules to their usage, is still often left to a human decision.

REFERENCES

Levenshtein, Vladimir. 1966. Binary codes capable of correcting deletions, insertions, and reversals. Cybernetic and Control Theory, 10(8), 707-10.

McClive, Tom, and Arehart, Mark. 2005. Romanization Systems, Phonemic Mapping, and Human Guesswork: The Conflation and Splintering of Computerized Non-English names. Paper presented at American Name Society session at the annual meeting of the Linguistic Society of America. Oakland.

Wagner, Robert and Fischer, Michael. 1974. The string-to-string correction problem. Journal of the Association for Computing Machinery, 21, 168-73.