Transcription, transliteration, transduction, and translation: a typology of crosslinguistic name representation strategies

Deryle Lonsdale
BYU Linguistics

The processing of proper names is a challenge for text processing in general; the issues become more difficult when the crosslinguistic dimension is added. For example, how sure can we be that CNN's mention of "Yuschenko" refers the same person as Le Monde's "Iouchtchenko"? Or that "بوش" and "bS" are the same as "Bush"? Most genealogists are familiar with Soundex as a simplistic indexing system for American census names; this paper seeks to further investigate and exemplify the issues in finding an ideal language-neutral representation for names across languages and cultures.

We focus on personal names, setting aside other—and related—proper nouns like organization names or country names. To keep the discussion tractable we will focus solely on the orthographic representation of names, ignoring several admittedly crucial linguistic phenomena such as: position and title modifiers; selection and ordering of name components (e.g. surnames, patronymics, lineage indicators); nicknames, abbreviations, and hypocoristics; morphological case marking on names; and coreferential forms, anaphors, and subsequent mentions occuring in connected text.

A substantial literature (not evident in the few references in this short overview) is developing in natural language processing (NLP) research on this topic, given the growing interest in web information extraction, managing transnational migration, and tracking the activities of persons of interest at home and abroad. Though applications involving crosslinguistic treatment of names (e.g. machine translation, text mining, cross-linguistic information retrieval, and speech recognition/synthesis) are well known, genealogical applications have not been widely addressed in the literature, yet certainly they are subject to the same processing issues.

Of course, personal names are closely tied with each language's writing system, and a wide variety of orthographic systems exists across the thousands of human languages. *Alphabetic* languages (e.g. English) are based on a (sometimes very loose) sound-to-symbol correspondence, but sounds and symbols vary greatly across languages. *Syllabic* scripts (e.g. Japanese katakana) tend to match better crosslinguistically, though phonological pecularities which often show at the syllable level introduce complexities. *Abugidic* languages (e.g. Hindi) show characteristics of both alphabetic and syllabic scipts, combining consonants and vowels in linearized but grouped patterns. *Logographic* or *ideographic* scripts (e.g. Chinese Hanzi) abstract away significantly from the actual sounds represented. To this already complex variety of systems are added many complications: alphabets often include diacritics, the spelling of personal names in any system is notoriously unstable, orthographic conventions change over time and due to standardization or reform measures, and people creating primary sources, particularly genealogical ones, demonstrate varying levels of competence in writing. In addition, *abjad* languages only represent a portion of the relevant sounds (e.g. Arabic, which may leave out vowels).

In view of these problems, a basic question arises: Is it possible to represent a person's name in such a way that it can be easily converted from any language to any language? Though this paper does not purport to introduce a comprehensive solution, it is still instructive to survey the typology of approaches that have been explored in the NLP literature. The effort may serve to incite genealogists to participate more actively in what has proven to be a complex undertaking for computationalists.

Perhaps the lowest level of conversion is *transcoding*, the rote conversion, character-by-character, of a string of symbols from one character set to another. For example, the GNU tool "recode" is specifically designed to transcode files between several dozen possible character sets. This method, while appealing at first glance, is wholly inadequate for large-scale genealogical applications since one-to-one mappings, even if they do exist for a given pair of character sets, rarely suffice for either the source or target language's orthography. For example, the surname "Bush" has three different Chinese Hanzi realizations (Gao et al., 2004) based on where the text originates and therefore which character set is used: 布什 (Mainland China), 布希 (Taiwan), 布殊 (Hong Kong); the name "Osama bin Laden" has at least 10 frequently-used variants in Hanzi-based publications (Halpern, 2002). The world of character set representation of linguistic material has historically been extremely chaotic, and though the Unicode effort is rectifying some of the problems many still persist.

Another possibility is to *transcribe* the spoken sounds of the name into some phonemic representation. Some standard representations exist (e.g. the ASCII SAMPA standard), and perhaps the most widely used among linguists worldwide is the International Phonetic Alphabet (IPA). Unfortunately, the IPA requires considerable linguistic expertise to use, and few end-user tools (even editors) fully support the IPA character set. For example, the American pronunciation of "Bush" would be transcribed into IPA as "bʊʃ". On the other hand, the French pronunciation of the same surname would have a different vowel specification. This fact highlights a further issue with transcription—that it is usually done within a language, and is not frequently used as a crosslinguistic representation. Still, some research has borrowed from speech recognition and text-to-speech technologies; a typical approach maps English name pronunciation sounds to a phonemic sequence which in turn is mapped to the romanized form of Hanzi glyphs (Virga & Khudanpur, 2003).

In terms of complexity but also of usefulness, the next possible processing modality is *transliteration*. This involves rewriting the sound symbols of one language in another language's writing system. Transliteration is generally more workable than just transcoding, since it does not involve strict one-to-one mappings. On the other hand, there are scores (if not hundreds) of transliteration schemes in use for working between the most widely used languages, causing almost as much indeterminism as transcoding itself. For example, researchers have noted at least 32 different spellings in the English press for the name of "Muammar Gaddafi", and 6 different transliterations in Arabic news sources for "Clinton". Korean Hangul, often thought to be quite regular, in fact admits significant variation; "Clinton" has the transliterated variants "keulrinteon" and "keulrinton".

Another approach, *transduction*, is used to morph a name into some canonical representation, often by the application of rules. For example, Soundex is a scheme familiar to genealogists that transduces names into an alphanumeric sequence, neutralizing putative variant forms of names and collapsing them together. Several researchers have applied formal computational techniques such as statistically weighted finite-state automata for transducing names into such standard representations or directly into another target language (e.g. Knight & Graehl, 1998). One problem that arises is that the standard itself is often underspecified; Soundex, for example, is alphabetic-based, Anglocentric, and cannot distinguish between names that are clearly should not be collapsed (e.g. Sri Lankan names Sivaramakrishnarao, Sivaramakrishnan, and Sivaramarao).

When previously mentioned strategies prove inadequate, the *translation* strategy is applied, even for proper names. This is often used for name conversion from alphabets or syllabaries into logographic systems, where fidelity to the sound sequences is judged to be too confining. For

example, the place name "Great Salt Lake" is written in Chinese as 大鹽湖 (da yan hu), a literal rendering of each word into a logograph (Chen et al., 2003). Personal names are sometimes rendered in this fashion.

In spite of all of the techniques described above, the most common approach to crosslinguistic name representation involves *lexical lookup*, where comprehensive lists of names in source and target language correspondences are kept. This entails a never-ending, arduous, and expensive task of assuring continual dictionary updates. Though the process can be automated somewhat with current statistical techniques and bitext alignment algorithms, a considerable amount of human involvement is still required. Indeed, some commercial ventures exist involving literally hundreds of millions of personal name entries and large-scale tools to manipulate them, just for work between the European languages. Similar ambitious efforts also exist for CJK (Chinese/Japanese/Korean) names.

It should be mentioned at this point that these techniques are often combined to meet the needs of specific implementations. The result is a wide range of complex, application-dependent systems that offer limited hope for generalization across languages. One might wonder, given the possibilities above, whether a unified architecture is possible, and if so what the best overall approach might be. Borrowing from the history of machine translation development, two architectural solutions seem possible: the direct approach and the pivot approach (see Figure 1).

Following the direct approach would mean developing a mapping module that would mediate specifically between a given pair of languages (e.g. Simplified Traditional Chinese → French Canadian). Clearly the specificity of the module would permit efficient mapping between source and target. Unfortunately, the number of necessary modules (and the concomitant effort in thir development) would explode as the number of source and target languages to be treated grows. The never-ending lexical lookup approach is an example of direct mapping. Huge resources are required, and developing new source-target correspondences requires bilingual expertise that may be as hard to find as the task is itself.

The direct approach's unwieldiness is mitigated somewhat by the pivot approach. A pivot is language-neutral sort of interlingua that is designed to represent the content of any language. An analysis module would need to be developed for each language whose content needs to be expressed by the interlingua, and a generation module would be needed to re-express interlingua content in the target language. What properties would be necessary for the interlingual pivot to be successful? It would have to be an entirely neutral representation scheme, not privileging any particular language's writing system. It would need to encode all the necessary information needed by any language's orthography. It would have to assure as lossless a conversion as possible from the source language and then to the target language, including all dialects of each. Finally, its design would need to be principled enough to allow algorithmic implementation in the relevant modules. One could question whether this approach is even feasible, given these constraints; there seems to be no current implementation that operates in this fashion. It should be mentioned, though, that some pivot-like implementations have been developed for languages with similar writing systems and linguistic principles; the ISCII transliteration scheme is used for many South Asian languages and acts conceptually like a pivot for those languages.

Of course, it should be mentioned that a wide array of processing techniques are being explored in the course of addressing these issues. Machine learning approaches ranging from rule-based transformation to statistical techniques to information-theoretic modeling serve as infrastructure for these tasks; often hybrid approaches incorporate together several of these different types of processing. Various string matching algorithms such as Levenshtein edit distance and associated

dynamic programming techniques can efficiently rate how similar two names might be. The availability of massive online corpora, often translated into more than one language, allows for harvesting of name correspondences and automatic identification of variant forms. Even increasing hardware capabilities enable the use of ever-greater lexical resources. Recent efforts in collaborative annotation environments might provide volunteer armies of enthusiasts the ability to annotate linguistic variants of names of interest in some standardized format.

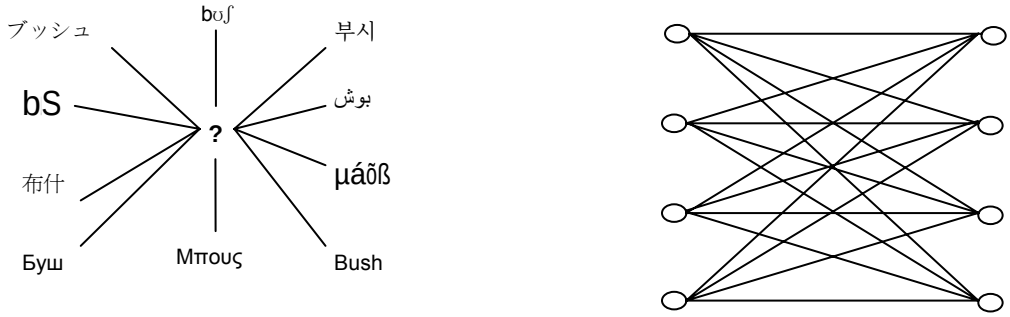

Figure 1: Two possible architectures for crosslinguistic name conversion: the pivot approach where all languages map to a common interlingual pivot (left), and the direct approach where a mapping between each language pair is encoded directly (right).

In summary, the representation and treatment of names is a complex computational and linguistic problem that has not yet been comprehensively solved. Though several implementations exist between high-demand languages (e.g. English, Chinese, Arabic) and within language families (e.g. the Devanagari script languages), they target specific types of applications and have not yet addressed the most complex issues that genealogists are all too familiar with. Genealogical expertise will be needed in the future to inform the design and implementation of next-generation tools for crosslinguistic name processing from both theoretical and practical points of view.

Knight, K. and Graehl, J. (1998). Machine Transliteration; Computational Lingiustcs 24(4).

Chen, H.-H., Yang, C. and Lin, Y. (2003) Learning Formulation and Transformation Rules for Multilingual Named Entities; Proceedings of the Workshop on Multilingual and Mixed-language Named Entity Recognition (ACL 2003); Edmonton, Canada.

Gao, W., Wong, K-F. and Lam, W. (2004) Phoneme-based Transliteration of Foreign Names for OOV Problem; Proceedings of the First International Joint Conference on Natural Language Processing (IJCNLP-2004); Hainan, China.

Halpern, J. (2002). Lexicon-based Orthographic Disambiguation in CJK Intelligent Information Retrieval; Proceedings of the 3rd Workshop on Asian Language Resources and International Standardization (*C*OLING 2002)*;* Taipei.

Virga, P. and Khudanpur, R. (2003). Transliteration of Proper Names in Cross-Lingual Information Retrieval; Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition; Edmonton, Canada.