A NEW METHOD FOR KOREAN TRANSLITERATION

A comparison of current practices and the KORDA system for computer users

PETER KANG

Making information that is written in non-Latin scripts readable in the English-speaking world requires that proper names — personal, geographical and otherwise — be transliterated, or romanized, while all other words and the syntax are translated. Names are often transliterated based on how they are pronounced instead of how they are spelled. Such a phonetic or *euphonic* system is generally complex in structure, leads to ambiguous or multiple results, and is a hindrance to computer applications such as data retrieval and automated transliteration.

While numerous technological innovations have taken place in recent decades in dealing with foreign languages, including advancements in automatic translation, optical character recognition and multilingual word processing, the importance of resolving the ambiguities caused by euphonic transliteration systems has been largely overlooked.

An ideal for the information age would be a spelling-based or *transliteral* system in which a given foreign language character is always transliterated into the same Latin expression (one-to-one mapping), and reverse transliteration always recovers the original expression. Transliteration would not involve human choice, but would eliminate variation based on personal judgment or preference. Only Latin letters — no diacritics, superscripts, numeral representations or special characters — would appear in the transliteration. The system would also provide the best possible representation of the pronunciation of the original language while conforming to the other conditions.

This article describes current practices in transliterating the Korean language, introduces KORDA (Korean Romanization for Data Application), a transliteral system designed to fit this universal transliteration model, and shows the extent to which the various systems satisfy the requirements of the model and the needs of computer users.

THE M-R SYSTEM

The McCune-Reischauer (M-R) system of Korean transliteration, which has become the standard for the governments of the United States and the Republic of Korea, is a prime example of a pronunciation-based system. In their 1939 paper introducing the system, *The Romanization of the Korean Language Based Upon Its Phonetic Structure*, McCune and Reischauer wrote, "We have devised our romanization with the purpose of providing a comprehensible guide to the standard modern pronunciation of Korean." Users of M-R encounter several problems, most of which are caused by the phonetic nature of the system.

Euphonic variation. The M-R system does not offer one-to-one mapping between hangul letters and their Latin counterparts. A given hangul letter can be transliterated into two or more Latin expressions, depending on its pronunciation. For example, the first consonant of the Korean alphabet (ㄱ) is transliterated in three different ways, and the seventh consonant (ㅅ) yields seven possible results.

Reverse ambiguity. The process of finding the original Korean name or word starting from an M-R expression also does not give a unique result. The romanized letter *K* can be identified with not just one but three original hangul letters: ㄱ, ㅅ and ㅋ. And other factors make reverse transliteration of words or names difficult or impossible. A manual prepared by the US Board on Geographic Names (BGN) states that because of these variations, the M-R conversion tables are elaborate, and reverse conversion from Latin script to hangul is *not* possible.

Space and hyphen. Under M-R, it is especially difficult to express geographic names consistently because there are no set rules provided by the system as to when to use spaces and hyphens within a place name. Transliterating the North Korean province name 평안북도 produces results as varied as *Pyongan Puk-To*, *Pyongan-buk To*, *Pyongan Pukto*, *Pyonganbuk-to*, *Pyongan-bukto* and more.

Pronunciation. M-R does not represent the pronunciation of Korean letters or syllables adequately, even though the system was designed basically as a pronunciation guide. Therefore, it is difficult for a non-Korean-speaking person to properly pronounce an M-R expression.

Diacritics. The authors of the M-R system used the diacritical symbols ˘ (breve) and ′ (apostrophe) in order to differentiate pronunciations. The typewriter does not have the breve key; it was evidently meant to be drawn by hand. The breve is also not on the computer keyboard. It can be entered into the computer using Microsoft Word, but the process involves invoking a special character or symbol set. Combining the problem of diacritics with the euphonic variations and the space and hyphen problem dramatically increases variability. If diacritically modified expressions are included, the variations in transliterating 평안북도 will now number at least 192.

Human factor. Some of these ambiguities are incorporated within the rules of the M-R system itself. In many cases, however, the variations are more widespread and are user dependent. Pronunciation of a word can vary significantly from person to person, from organization to organization and from one geographic location to another, as in the case of provincial dialects. Even if the pronunciation is given, how to spell it out in characters may

Comparison of Korean Transliteration Systems

Hangul	M-R (Original Version)	Morse	AOS NK	MOE SK	Yale US	Unified SK-NK	NAKL SK	KORDA
ㄱ	K, G, NG	L	K	G	K	K	G, K, NG	G
ㄴ	N, L, [nt]	F	N	N	N	N	N, L	N
ㄷ	T, D, CH, J	B	T	D	T	T	D, T	D
ㄹ (i)	R, L, N, [nt]	V	R	R	L	R	R, L	R
ㄹ (f)	R, L	V	R	L	L	L	L	L
ㅁ	M	M	M	M	M	M	M	M
ㅂ	P, B, M	W	P	B	P	P	B, P	B
ㅅ	S, SH,D,T, N, P, K	G	S	S	S	S	S, D, N	S
ㅇ (i)	[nt]	K	[nt]	[nt]	[nt]	[nt]	[nt]	[nt]
ㅇ (f)	NG	K	NG	NG	NG	NG	NG	NG
ㅈ	CH, J, T	P	TS	J	C	C	J	J
ㅊ	CH´, T, (N)	C	TSH	CH	CH	CH	CH, T, N	CH
ㅋ	K´, K	X	KH	K	KH	KH	K	K
ㅌ	T´, CH´, T, (D,N)	Z	TH	T	TH	TH	T, CH	T
ㅍ	P´, P	O	PH	P	PH	PH	P	P
ㅎ	H, [nt]	J	H	H	H	H	H, [nt]	H
ㅏ	A	E	A	A	A	A	A	A
ㅑ	YA, A	I	YA	YA	YA	YA	YA	YA
ㅓ	Ŏ	T	Ŏ	EO	E	EO	EO	U
ㅕ	YŎ, Ŏ	S	YŎ	YEO	YE	YEO	YEO	YU
ㅗ	O	A	O	O	O	O	O	O
ㅛ	YO, O	N	YO	YO	YO	YO	YO	YO
ㅜ	U	H	U	U	WU	U	U	OO
ㅠ	YU, U	R	YU	YU	YU	YU	YU	YOO
ㅡ	Ŭ, U	D	Ŭ	EU	U	EU	EU	EU
ㅣ	I	U	I	I	I	I	I	I
ㅐ	AE	EU	AI	AE	AY	AE	AE	AE
ㅒ	YAE	IU	YAI	YAE	YAY	YAE	YAE	YAE
ㅔ	E	TU	E	E	EY	E	E	E
ㅖ	YE, E	SU	YE	YE	YEY	YE	YE	YE
ㅘ	WA	AE	WA	WA	WA	WA	WA	WA
ㅙ	WAE	AEU	WAI	WAE	WAY	WAE	WAE	WAE
ㅚ	OE	AU	OE	OE	OY	OE	OE	OE
ㅝ	WŎ	HT	WŎ	WEO	WE	WEO	WO	WO
ㅞ	WE	HTU	WE	WE	WEY	WE	WE	WE
ㅟ	WI, I	HU	WI	WI	WI	WI	WI	WI
ㅢ	ŬI, I	DU	ŬI	EUI	UY	YI	UI	UI

(i) initial consonant (f) final consonant [nt] not transliterated

also depend on personal preference. In addition, the same person often transliterates the same word differently from one time to another because it is difficult to remember which one of the scores of possible euphonic variants was used before.

Different M-R systems. The original M-R system with the diacritics ˘ and ´ is the official US government standard. Yet, because of the difficulty of using these symbols with typewriters and computers, most government agencies have adopted various modifications of M-R by simplifying or eliminating the use of diacritics. Three such modified versions are in current use. The only exceptions are the Library of Congress and the BGN, which decided to maintain the original M-R system. Thus, additional confusion arises from the mismatch among four different M-R versions.

Search and retrieval. The inherent variability of the M-R system poses a constant challenge in searching a database for romanized Korean words or names. If you want to look up the name of the North Korean ballistic missile that was fired over Japan in 1998, you must type expressions one by one until a hit is obtained, for example, *Taepo-dong*, *Taep'o-Dong*, *Taep'odong*, *Taepodong*, *Daepodong*, *Daepo-dong*. And since few people would have the patience to search through all the variations of a geographic term such as *Pyongan-bukto*, it is understandable that geographic names are often considered "non-searchable data."

In such a search, the term may be "virtually lost" even if it is in the computer because the person searching cannot replicate the spellings that were entered into the computer and an adequate search engine is not available. Personal names are at least searchable, though with difficulty, by use of specially developed search engines. With these tools, the user can enter any one variant of a personal name into the computer, which will then produce all the names with similar sounds or spellings, including all the varied uses of the space, hyphen and diacritic. Then the user has to sift through many incorrect spellings to find the correct one. A non-Korean-speaking reader might better appreciate the agony of searching a database in the M-R environment by imagining an English-language computer database made up of pronunciation-based expressions such as *Linkan*, *Lingcon*, *Ling-con* and *Ling Kon* instead of *Lincoln*.

Special rules. Recognizing weaknesses in their system, McCune and Reischauer provided a number of special rules for personal and geographic names in an effort to preserve hangul spelling and minimize confusion. One of these rules says, "It seems best for romanization purposes to disregard euphonic changes between surnames and given names or titles." Without these special provisions, the M-R system could be much more difficult to use.

Other Transliteration Systems

A number of Korean transliteration systems have emerged since M-R was introduced. The six most prominent systems are Morse, AOS (Korean Academy of Sciences, North Korea, 1956), MOE (Ministry of Education, South Korea, 1959), Yale (Yale University, 1967), Unified (North-South Unified System, 1992) and NAKL (National Academy of the Korean Language, South Korea, 2000). All but NAKL are transliteral systems and certainly much easier to use than M-R. All six systems offer simplified transliteration by doing away with diacritics.

Morse. The Morse system, also called SKATS (Standard Korean Alphabetized Transliteration System), was created by comparing the Morse codes for the English and Korean alphabets. The first Korean consonant ㄱ is transliterated as the Latin character *L* because both of these characters have the Morse code value of ·—·· (dot dash dot dot).

Morse is the only one of these six systems that is free of all the ambiguity problems of M-R and thus assures 100% data accuracy. Transliteration of a given name can be done only one way, and starting from the transliterated expression you can always recover the exact spelling of the original word. It is, however, totally devoid of relevance to the pronunciation of Korean words being transliterated, making a Morse expression completely unreadable to anyone except specially trained personnel. As an example, the name of

the South Korean President would be LUM BEU PHK according to Morse and the North Korean capital would be OSK KIK.

Morse is obviously not applicable for general communication because very few people in the world can read it. However, because of its data accuracy, Morse is used for database maintenance in critical areas in the US government where data ambiguity cannot be tolerated. Interestingly, Morse is also used as a tool to teach spoken Korean to military personnel.

Readable transliteral systems. Because of one-to-one mapping, the AOS, MOE, Yale and Unified systems are much more accurate than M-R and, unlike Morse, are also readable. However, they have not resolved the arbitrary use of space and hyphen and the inadequate representation of pronunciation.

As in the M-R system, spaces and hyphens can be employed more or less at will in these systems. As for pronunciation, AOS, Yale and Unified yield *Tsung*, *Cwung* and *Cung* respectively in Latin script for the last hangul syllable (중) of the South Korean president's name when it should be correctly represented by *Joong* instead. *Joong* closely emulates the original hangul sounds, which include *J* as in *jacket* and *oo* as in *food*.

The NAKL system. The newest South Korean government system was adopted and proclaimed as the South Korean standard by the Ministry of Culture and Tourism in July 2000. In terms of pronunciation representation, NAKL is undoubtedly the best existing system. ㅓ= *eo* and ㅕ= *yeo* are the only flawed relationships, as in *Gim Il Seong* and *Pyeongyang*. There was considerable public objection in Korea over these two relationships soon after the system was announced.

NAKL is actually "a compromise between phonetic and transliteral systems," as pointed out by the Minister of Culture and Tourism. However, as a mostly phonetic system, NAKL carries the shortcomings associated with M-R: euphonic variations, reverse ambiguity, the space and hyphen problem, transliteration varying with user opinion and consequent difficulty in adapting to computers.

Under NAKL it is often up to the user to decide on how to interconnect syllables. Regarding the handling of hyphens, one instruction says, "When there is the possibility of confusion in pronunciation, a hyphen may be used, for example, *Jung-ang* or *Hae-undae*." The decision is left up to the user.

In addition, NAKL has provided deviations from its usual phonetic rules. One of the exceptions specifically tells the user to follow the spelling instead of pronunciation: "When it is necessary to convert romanized Korean back to hangul in special cases such as in academic articles, romanization is done according to hangul spelling and not pronunciation."

KORDA

The author set out in 1983 to create a system that overcomes the inadequacies of the traditional pronunciation-based systems and makes transliteration of Korean easy, accurate and computerizable. The KORDA system was developed over a 10-year period with extensive research in Washington and Seoul and with contributions from Korean linguists and analysts from various US government agencies. The result was published in 1992, and the system has been thoroughly tested for data accuracy and pronunciation.

KORDA accomplishes 100% data accuracy by strictly adhering to the one-to-one principle and by not allowing arbitrary use of spaces and hyphens. For example, *Pyung An Boog Do* is the only possible expression for 평안북도, and *Dae Po Dong* is the only way the name of the North Korean missile can be expressed in Latin script. One can unequivocally recover the original hangul spelling from any transliterated expression.

Under KORDA, a space is always placed between transliterated syllables. The ambiguity problem of *Jungang* and *Haeundae* is resolved by writing the syllables separately as *Joong Ang* and *Hae Oon Dae*. As long as syllable separation is enforced, there is no risk of losing the hangul spelling information. No other Korean transliteration system but Morse seems to provide this feature.

In addition to data accuracy, KORDA transliterations closely emulate the sounds of the corresponding hangul syllables — perhaps even better than M-R, especially in pronunciation of personal names.

KORDA is simple enough to be used by non-Korean speakers. Users do not have to understand what a word means or know how to pronounce it. All that is needed is the KORDA table of one-to-one mapping relationships and minimal training to recognize the hangul characters.

How Names Are Transliterated Under Several Systems

Accepted Spelling	M-R Original	Morse	AOS	MOE	Yale	Unified	NAKL	KORDA
Kim Dae Jung	Kim **Tae-Chung**	**lum beu phk**	Kim **Tai Tsung**	Gim Dae Jung	Kim **Day Cwung**	Kim **Tae Cung**	Gim Dae Jung	Gim Dae Joong
Kim Jong Il	Kim **Chŏng**-Il	**lum ptk kuv**	Kim **Tsŏng Ir**	Gim **Jeong** Il	Kim **Ceng** Il	Kim **Ceong** Il	Gim **Jeong** Il	Gim Jung Il
Kim Il Sung	Kim Il-**Sŏng**	**lum kuv gtk**	Kim **Ir Sŏng**	Gim Il **Seong**	Kim Il **Seng**	Kim Il **Seong**	Gim Il **Seong**	Gim Il Sung
Pyongyang	P´**yŏng**yang	**osk kik**	**Phyŏng**yang	**Pyeong**yang	**Phyeng**yang	**Phyeong**yang	**Pyeong**yang	Pyung Yang
Samsung	Sam**sŏng**	**gem gtk**	Sam**sŏng**	Sam**seong**	Sam**seng**	Sam**seong**	Sam**seong**	Sam Sung
Hyundai	**Hyŏn**dae	**jsf beu**	**Hyŏntai**	**Hyeon**dae	**Hyen**day	**Hyeon**dae	**Hyeon**dae	Hyun Dae
Daewoo	**Tae**u	**beu kh**	**Taiu**	Daeu	**Tay**wu	**Tae**u	Daeu	Dae Oo

Bold type indicates inadequate representation of the Korean pronunciation

AUTOMATIC TRANSLITERATION

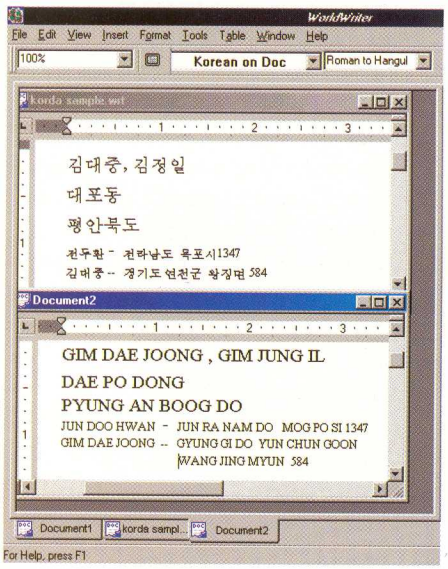
Meaningful automatic transliteration is impossible with M-R or NAKL because hundreds of phonetic rules, many of them variable and subjective, would have to be incorporated into the program. Computers cannot accommodate all the subtleties of the human factors that any phonetic system entails.

A KORDA-based transliteration function was implemented by BlueShoe Technologies as a special Korean language feature for WorldWriter, the company's existing multilingual word processor. It handles automatic transliteration of documents, pure text files and the row/column contents of Microsoft Word or Excel files. The simple hangul-to-Latin mapping relationship can be easily incorporated into any other Korean word processor.

The user selects Korean words or sentences from the screen and clicks the Transliterate button. The English rendering of the hangul text appears in a split screen.

This technology can also be directly applied to the map-making process. Starting with an electronic map of Korea in the Korean language, one can automatically generate a romanized map as a computer converts each hangul place name into Latin script. The process currently requires a professional translator in a mapping facility to create the Latin-script equivalent of each Korean place name by examining its phonetic structure.

Automatic transliteration will be especially important when handling large volumes of data or when processing of material is urgent.



Automatic transliteration

WORKING WITH TRANSLATION TOOLS

Ambiguity-free transliteration of names would clearly enhance the output quality of automatic translation software. SYSTRAN Software, Inc., is incorporating the KORDA system in future releases of its Korean automatic translation program. The user will have the option of seeing personal names in KORDA spellings. SYSTRAN will, however, continue to provide accepted spellings for well-known names in its dictionary.

For geographical names, the existing SYSTRAN tool provides a combination of translation and transliteration. For example, the name 평안북도 has been converted to *North Pyongan Province*. *Pyongan* is an M-R transliteration; the other two words are translation. While retaining this mixed convention, the future SYSTRAN tool will be able to provide a full and accurate transliteration of an entire multisyllabic name such as *Pyung An Boog Do* (*Boog* = *north*; *Do* = *province*), enabling the user to maintain and recover the original spelling.

Three Korean Input Methods

Accepted Spelling	Korean Standard Input (Microsoft)	Hangul Wordian-97	Phonetic Method of KORDA	한글 Hangul
Kim Dae Jung	RLA EO WND	GIM DAI JUNG	GIM DAI JOONG	김대중
Kim Il Sung	RLA DLF TJD	GIM XIL SEONG	GIM IL SUNG	김일성
Kim Jong Il	RLA WJD DLF	GIM JEONG XIL	GIM JUNG IL	김정일
Samsung	TKA TJD	SAM SEONG	SAM SUNG	삼성
Daewoo	DO DN	DAI U	DAE OO	대우
Hyundai	GUS EO	HYEON DAI	HYUN DAE	현대

Relationship between keys entered on a Latin keyboard and corresponding hangul output

PHONETIC INPUT METHOD

The Korean input method most commonly used in the western world is Korean Standard Input (KSI), which is based on the hangul keyboard used in Korea. This input method was adopted by Microsoft as part of the CJK (Chinese-Japanese-Korean) Input Method Editor (IME) for Windows 2000. KSI is the obvious choice if you are using a Korean-made computer. Without a Korean keyboard, however, hangul characters must be clicked with the mouse, one at a time, on the virtual keyboard. Typing without the virtual keyboard is possible, but there is no meaningful relationship between the Korean character to be generated on the screen and the corresponding Latin letter on the keyboard. For example, to create the expression 김대중, you must type *rlaeownd* on the English keyboard. Some people use stickers with native characters to deal with this problem.

BlueShoe programmers have implemented a new phonetic-equivalent hangul input method based on KORDA which allows a user to create a Korean syllable on the screen by typing an English syllable that closely emulates the pronunciation of the hangul syllable. To create the hangul syllables 김대중, you would type *gim dae joong*, separating the syllables with spaces.

This phonetic method, with the option of viewing the KORDA transliteration table for use as a typing guide, has been incorporated in WorldWriter along with other Korean language input methods such as KSI.

Hangul Wordian, a Korean word processor, provides two user-selectable phonetic input methods. One of them closely follows the transliteration tables of MOE, NAKL and KORDA, and is similar to the KORDA input method. The major shortcoming of this method is the use of the relationships ㅓ= *eo* and ㅕ= *yeo*. Wordian's second phonetic input method fails more severely in terms of phonetic equivalence. For example, to enter the hangul name of North Korea's Great Leader, the user types *Gim Xil Seong* (compared with *Gim Il Sung* in KORDA).

IN SUMMARY

Transliterating the Korean language by using a phonetic system such as M-R or NAKL results in several types of ambiguity and is not well suited for information technology. The current alternative, Morse, can accurately keep track of data, but a Morse expression is unreadable except to a specially trained linguist. A transliteration system completely free of variability will vastly streamline data search/retrieval processes.

KORDA, being mechanical in nature with no dependence on the user's knowledge or judgment, offers a solution to a number of complex problems encountered in transliterating Korean and provides cost- and time-saving possibilities for all who work with Korean information.

Peter Kang is an IT program manager working for the US government. He can be reached at p.kang@verizon.net