LACUS FORUM XXXII

Networks

Dartmouth

Reprint

LACUS FORUM XXXII

Networks

Edited by

Shin Ja J. Hwang,
William J. Sullivan &
Arle R. Lommel

THE LINGUISTIC ASSOCIATION OF CANADA AND THE UNITED STATES

EME-WARE: FROM EMIC ANALYSIS TO PRACTICAL INPUT SYSTEMS[1,2]

Daniel S. Mailman
MindStride Corporation

In the twenty-first century, for good or ill, the personal computer (PC) will replace pen and paper as the medium for the average, educated participant in global society and commerce to record, explore, manipulate, and communicate thought. So-called productivity applications—such as word processors, e-mail, and instant messengers—will increasingly be used as the medium for written (sic) communication in all the world's languages, and the symbols for communications will expand beyond language (e.g. the use of emoticons).

Unfortunately, from the 1940s through the 1980s, the PC was designed for use almost exclusively for the English language. Although recent decades have seen software makers begin to address non-English languages, it is still the case that—for use in any language (or with any input set) other than English—there is an additional language-dependent degree of difficulty for entering basic elements into productivity applications. The degree of deficit is dependent on a measure we at MindStride informally call the **distance** between English and the (target) language or input set. The concept of **distance from English** can be made more rigorous, but since we focus on input methods production and not their categorization, we are generally satisfied to relate it to the number of separate scripts in the language[3] as well as the number of non-keyboard symbols in common use. We then use the measure to estimate that the following input sets are increasingly **far from English** (with commensurately increasing **input efficiency deficits**):

1. Spanish (**Figure 2**)
2. Vietnamese (**Figure 5**)
3. 'European Languages' (**Figure 8**)
4. Japanese (**Figure 9**)
5. Korean (**Figure 10**)
6. Chinese (**Figure 11**)

Given the impracticality of reducing input efficiency deficits with input-set-specific keyboards,[4] the most common tactic—that taken by MindStride and other organizations—is to develop software (**input methods**)[5] that enables users to send symbols to productivity applications. Input methods are more or less **elegant** to the degree that they make the input task:

- Easier (fewer, rapidly learned, simple gestures producing more input symbols)[6]
- Faster (greater count of input symbols per unit of time)
- More accurate (fewer gestures wasted on unintended symbols)

Reprinted from ***LACUS Forum 32: Networks***, edited by Shin Ja Hwang, William J. Sullivan & Arle R. Lommel. 2006. Houston TX, LACUS.

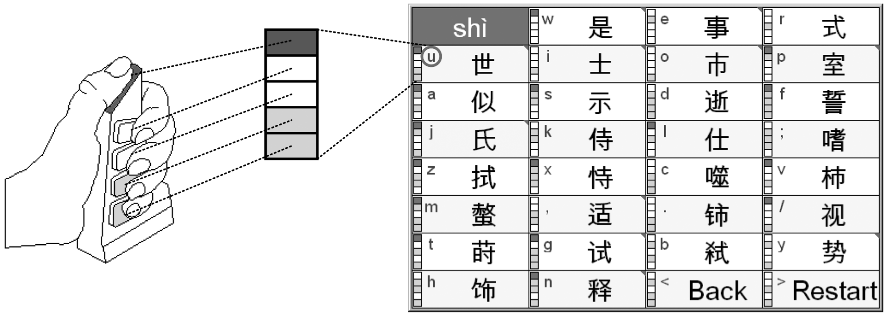

Figure 1. *Chordal input device with Chinese input method.*

The asymptotes of an ideal (maximally elegant) input method are indicated by a *reductio ad absurdum* specification: it takes *no* time to learn to *instantaneously* input a *maximal* sequence of desired symbols from a *maximal* range of symbols using a *minimal* set of unique gestures to select from *minimal* sets of choices arranged in *minimal* sequences.

It is interesting, possibly useful, and certainly maddening to speculate that the practical limits to input method elegance are user languages for ranges of symbols and associated sequences of phonemes for speed and gesture range size. The implications of such speculation together with observations of competing phonemic and linguistic analyses for specific languages is a strong indication that no system (MindStride's included) will automatically result in the single best input method for a given language, but also an indication that the limits of elegance can be approached.

MindStride input methods are based on **emic** (e.g. graphemic, phonemic) analyses of input sets, as well as an understanding of the average, educated, input set user. Each analysis is optimized to take best advantage of the features of the MindStride Input Engine to maximize gestural efficiency and learnability. Input set analyses for Spanish, for Vietnamese, and for European languages taken as an aggregate illustrate that, as with linguistic analyses in general, input set analysis (and therefore input method development) is as much art as science.

1. MINDSTRIDE INPUT ENGINE. The MindStride Input Engine displays grids of choices (**Figure 1**) that indicate how to use input devices to select input choices.

Users make selections with the devices in the standard way:

- **Mouse/Stylus**: Click anywhere on the cell containing the choice.
- **Keyboard**: Type the key indicated in the upper left corner of the cell.
- **Chordal Devices**: Press and release the finger combination in the cell.

Selections either:

- Cause input to be sent to (e.g.) productivity applications for simple inputs, or
- Cause display of subsequent choice grids in sequences leading to complex inputs.

Reprinted from *LACUS Forum 32: Networks*, edited by Shin Ja Hwang, William J. Sullivan & Arle R. Lommel. 2006. Houston TX, LACUS.

Because MindStride can be actuated by combinations of five fingers on chordal devices, there is an additional constraint of limiting the number of grid choices to 31 ($= 2^5 - 1$).

MindStride's **Press & Hold (P&H)** feature[7] enables users to access alternative inputs or alternative subsequent grids (indicated by a red tic-mark in the upper right corner of the cell). P&H is activated by:

- **Mouse/Stylus**: Instead of clicking on the cell, sustain the press for a short period.
- **Keyboard**: Instead of releasing the key immediately upon typing it, hold it down.
- **Chordal Devices**: Press and hold the finger combination indicated in the cell.

In all cases, feedback is provided when the sustained activation time is reached. There are three different P&H actions: **P&H Repeat, P&H Grid, and P&H Selection**. **P&H Repeat** sends the selection to productivity applications for as long as the device is actuated. **P&H Grid** causes MindStride to display an alternative successor grid. **P&H Selection** displays new selections in place of current selections; completing the input device gesture causes MindStride to send the selection to the productivity application.

Users change the interval MindStride waits before responding to P&H by changing a configuration value.

2. INPUT DEVICES. Most input devices (mouse, stylus, joysticks, etc.) are unmarked. Using MindStride with these devices is simply a matter of using the device in the standard way for sequence choice selection, and in an easily learned, intuitive way for the expanded selections available via the P&H feature.

As intimated prior, there are 31 possible combinations of 5 fingers taken 1, 2, 3, 4, or 5 at a time. So, for optimal operation with chordal devices, software (including input methods) must offer a maximum of 31 choices to its users at any one time.[8] MindStride has found it most useful to indicate chordal operation by visually associating input sequence choices with **ChordMaps** (**Figure 1** shows vectors of five juxtaposed blocks: filled blocks indicating pressed fingers; empty blocks indicating un-pressed fingers).

Consideration of the field of input devices will reveal why the keyboard, although most common and arguably most useful, is also the most challenging input device to accommodate with a software input method. The keyboard, in that the device itself comprises a multiplicity of 'named' keys, which names are integral to its operation, forces input methods to visually associate keys with non-key inputs and with sequenced sets of choices leading to inputs (e.g. pinyin letters leading to Chinese characters). The choice of key to associate with sequence choice is a major consideration in input method development (upper left hand corner of each grid cell in **Figure 1**). And, in part because keys are often more familiar than the inputs themselves, a secondary issue is the order in which the associations are presented to the user.[9]

3. SPANISH LETTERS INPUT METHOD. Because, by our measure, Spanish is fairly close to English, dEspañol™—MindStride's Spanish letter input method—is a complete example of identifying the emic level(s) (e.g. graphemic, phonemic, morphemic, sememic) at which

á é í ó ú ü ñ Á É Í Ó Ú Ü Ñ ¿ ¡ € « » ª º

Figure 2. *Common non-keyboard symbols for Spanish.*

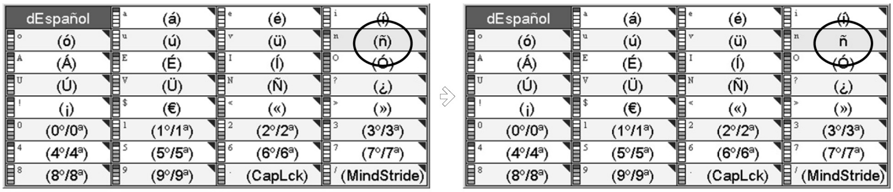

Figure 3. *(dEspañol) Press & Hold of **n** key causes input of the letter **ñ**.*

analysis, science, and art yield optimal input methods for intuitive sequence choice presentation to the average, educated user of an input set.

The problem, amenable to many solutions (of less or more elegance) is to make it easy, fast, and accurate for the user to access non-keyboard characters commonly used in producing written Spanish documents (**Figure 2**).

Visual inspection of the domain indicates that a graphemic approach is possible. Emic analysis is fairly straight-forward; the non-keyboard-key graphemes of the system can be named by:

{'Acute' 'Diaeresis' 'Tilde' 'Inverted' '€' 'Double' 'Superscript'}

With the exception of {'Diaeresis' 'Superscript'}—and with an excursion into the sememic domain by noting that the keyboard symbol **$** and the non-keyboard symbol **€** are both **currency indicators** for the average, educated Spanish speaker—the emes are in one-one correspondence with {'Keyboard Symbol'}, yielding the (Keyboard Symbol:Non Keyboard Symbol) map:

{a:á e:é i:í o:ó u:ú n:ñ A:Á E:É I:Í O:Ó U:Ú N:Ñ ?:¿ !:¡ <:« >:» $:€}

dEspañol uses the P&H feature to enable users to input the non-keyboard symbols from the map by **holding** the associated key; so, for instance, the letter **n** is input by typing the **n** key normally, and the letter **ñ** is input by holding the same key down for a slightly longer period of time. The software gives visual feedback in the form of removing parentheses around the desired input to let the user know when the key has been held sufficiently long (**Figure 3**).

The advantages of the P&H implementation are:

- It doesn't interfere with prior existing skills; specifically, the keyboard and mouse are used in the accustomed manner.
- Additional learning and memorization are extremely minimized; the desired input is already visually (graphemically) associated with the key and the single, additional skill of holding a key down is readily learned.

Reprinted from *LACUS Forum 32: Networks*, edited by Shin Ja Hwang, William J. Sullivan & Arle R. Lommel. 2006. Houston TX, LACUS.

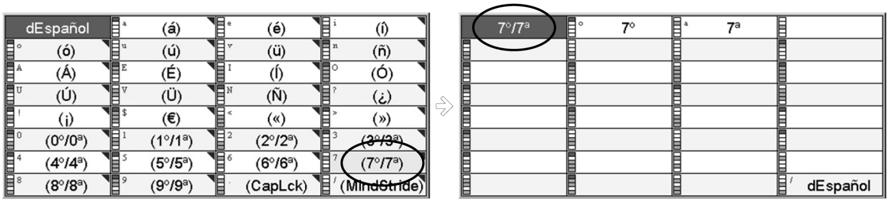

Figure 4. *(dEspañol) P&H of digit keys causes ordinalizing superscript display choice.*

The remaining emes, {'Superscript' 'Diaeresis'}, if implemented with P&H for graphemically associated keyboard symbols, would conflict with existing mappings; for example the key **a** cannot be associated with both **á** and **ª** for efficient use of the P&H feature.

An elegant solution (**Figure 4**) for the ordinalizing superscripts **ª** and **º** takes advantage of the (quasi-morphemic) observation that they occur only after digits **0–9**; that is **0**–**9** are emic contexts for both. So, rather than associating ordinalizers with corresponding keyboard keys, dEspañol presents ordinalizer choices when users hold the digit keys and enables selection of the particular ordinalizer by typing (not holding) the associated key.

The last remaining issue is the eme {'Diaeresis'} associated with the letters **ü** and **Ü**. Interviews with many educated, literate users indicate that overloading the **u** key is less convenient than associating the dieresis on a P&H basis with the letters **v** and **V** (historically used to indicate the letter **U**).

The process for creating dEspañol was:

1. Determining the most useful eme is (primarily) graphic—rather than, say, phonemic.
2. Identifying strategic emes.
3. Applying science (e.g. to symbols that map uniquely to keyboard keys).
4. Applying art (e.g. determining contexts to avoid overloading keyboard symbols).

One result of applying art in this case resulted in making the sacrifice of a single, moderate inconvenience (associating the **v** and **V** keys with the symbols **ü** and **Ü**) to avoid the more egregious inconvenience of overloading two keys. Elements of the same analytic process were used in the development of **dViệt-ngữ**™—MindStride's Vietnamese input system.

4. VIETNAMESE LETTERS INPUT METHOD. The most salient difference between Spanish and Vietnamese (for the purposes of designing an input method) is the larger counts of variations of keyboard symbols (e.g. **Figure 5**, overleaf, contains 18 variations of the letter **a**, including the symbol itself).

As with dEspañol, to preserve intuitiveness and not to interfere with pre-existing typing skills, each set of variations is accessed via P&H from the key (visually/graphemically) associated with the variations. If there is only one variation (e.g. **đ**, **Đ**), P&H Selection is used to input the variation. If there are multiple variations, the P&H Grid feature displays the choices for variations to be selected.

Reprinted from *LACUS Forum 32: Networks*, edited by Shin Ja Hwang, William J. Sullivan & Arle R. Lommel. 2006. Houston TX, LACUS.

		á	Á	à	À	ả	Ả	ã	Ã	ạ	Ạ
ă	Ă	ắ	Ắ	ằ	Ằ	ẳ	Ẳ	ẵ	Ẵ	ặ	Ặ
â	Â	ấ	Ấ	ầ	Ầ	ẩ	Ẩ	ẫ	Ẫ	ậ	Ậ
đ	Đ										
		é	É	è	È	ẻ	Ẻ	ẽ	Ẽ	ẹ	Ẹ
ê	Ê	ế	Ế	ề	Ề	ể	Ể	ễ	Ễ	ệ	Ệ
		í	Í	ì	Ì	ỉ	Ỉ	ĩ	Ĩ	ị	Ị
		ó	Ó	ò	Ò	ỏ	Ỏ	õ	Õ	ọ	Ọ
ô	Ô	ố	Ố	ồ	Ồ	ổ	Ổ	ỗ	Ỗ	ộ	Ộ
ơ	Ơ	ớ	Ớ	ờ	Ờ	ở	Ở	ỡ	Ỡ	ợ	Ợ
		ú	Ú	ù	Ù	ủ	Ủ	ũ	Ũ	ụ	Ụ
ư	Ư	ứ	Ứ	ừ	Ừ	ử	Ử	ữ	Ữ	ự	Ự

Figure 5. *Non-keyboard symbols for Vietnamese.*

The multiplicity of keyboard symbol variations makes overloading keys unavoidable, so the principal remaining issues in the design of dViệt-ngữ are:

- How to order the variations
- How many gestures to associate with the variations
- Which keys to associate with the variations

Visual inspection indicates that the non-keyboard-key graphemes of the input set can be named by

{'Breve' 'Circumflex' 'Horn' 'Stroke' 'Acute' 'Grave' 'Hook' 'Tilde' 'Dot'}[10]

The collating order for the standard Vietnamese alphabet (quốc ngữ)

a ă â b c ch d đ e ê g gi h i k kh l m n ng nh o ô ơ p ph q r s t th tr u ư v x y

yields a useful primary ordering:

{'Breve' 'Circumflex' 'Stroke' 'Horn'}

while the received pedagogical order for tones yields a secondary ordering:

{'Acute' 'Grave' 'Hook' 'Tilde' 'Dot'}[11]

So, dViệt-ngữ presents the multiple key symbol variations in the order of **Figure 5**.

dViệt-ngữ takes two simultaneous approaches to addressing the design issues of gesture count per symbol and key-to-symbol associations.

Reprinted from *LACUS Forum 32: Networks*, edited by Shin Ja Hwang, William J. Sullivan & Arle R. Lommel. 2006. Houston TX, LACUS.

Grapheme	Key	Keyname
Breve	(	Left paren
Circumflex	^	Circumflex
Horn	,	Comma
Acute	'	Apostrophe
Grave	`	Grave
Hook	?	Question mark
Tilde	~	Tilde
Dot	.	Period

Figure 6. *Mnemonics for Vietnamese vowel marks.*

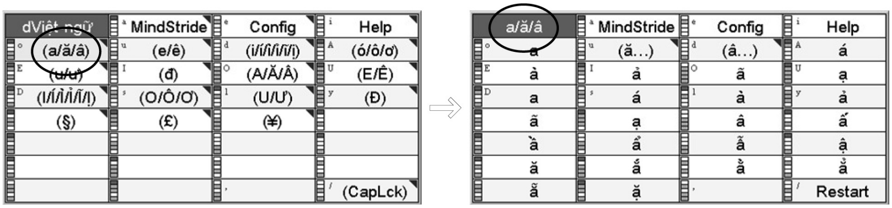

Figure 7. *Two different approaches for overloaded Vietnamese letters.*

The first approach—for users whose primary input device for Vietnamese is not the keyboard or who want to maximize long-term productivity—minimizes gesture count per symbol at the expense of a slight learning curve. This **minimal sequence** approach associates single keys with each of the (maximum 18) vowel variations. Mnemonics are impractical, so, to facilitate learning, the arrangement of associated keys follows their physical layout on the keyboard. The associations remain constant for same variations of different keyboard symbols (e.g. **Ẫ** and **Ỗ**) and are learned, via repetition, in fairly short order (approximately two hours of typing). Because it halves the gesture count for most Vietnamese symbols, the minimal sequence approach also reduces fatigue if the primary input device for Vietnamese is a mouse, stylus, or chordal device.

The second approach—for users willing to sacrifice some speed for virtually no learning curve—associates keys with each of the graphemic components of the keyboard key variations. Since most variations have two separate marks, the strict graphemic approach forces the user to type two keys for many Vietnamese letters, but involves only 8 easily learned, visually mnemonic associations of keyboard symbols with graphemic elements (**Figure 6**).

dViệt-ngữ implements both approaches in the same display so users can easily choose which approach they prefer (**Figure 7**). Rows 2–3 implement the strict graphemic approach and rows 4–8 implement the minimal sequence approach.

Due to the greater distance from English for Vietnamese, the design of dViệt-ngữ involves both more science and more art than the design of dEspañol. Science assists in determining that a graphemic approach is reasonable, in identifying the individual graphemes, and in ordering the graphemes. Art is used to determine presentation ordering principles (by

Reprinted from *LACUS Forum 32: Networks*, edited by Shin Ja Hwang, William J. Sullivan & Arle R. Lommel. 2006. Houston TX, LACUS.

·	ĸ̧	Ǫ	ǫ	ȷ	ϴ	ϵ	ș	ț	Ș	Ț		–	—	ˆ	¡	¦	¨	¯
´	¸	¿	!	˜	ʼ	ˏ	˛	: ;		‥	‘	’	‚	“	”	„	‹	›
±	«	»	×	÷	|	‖	ǂ	¢	£	¤	¥	§	©	¬	®	°	µ	¶
	†	‡	•	…	‰	€	¼	½	¾	¹	²	ƻ	³	Ƽ	ƽ	ª	Á	á
À	à	Ậ	â	Ä	ä	ă	Ă	Ă	ă	Ā	ā	Ã	ã	Å	å	Ą	ą	ǡ
Ǡ	ǻ	Ǻ	ǟ	Ǟ	ȁ	Ȁ	ȃ	Ȃ	Æ	æ	ǽ	Ǽ	ǣ	Ǣ	ƀ	Ɓ	ƃ	Ƃ
ƅ	Ƅ	Ć	ć	Ċ	ċ	Ĉ	ĉ	Č	č	Ç	ç	ƈ	Ƈ	Ɔ	Ď	ď	Đ	đ
ƌ	Ƌ	Ɗ	Ð	ð	ƍ	ǳ	ǲ	Ǳ	ǆ	ǅ	Ǆ	Ɖ	É	é	È	è	Ė	ė
Ê	ê	Ë	ë	Ě	ě	Ĕ	ĕ	Ē	ē	Ę	ę	ȅ	Ȅ	ê	Ȇ	Ɛ	ə	Ǝ
Ə	Ƒ	ƒ	ƒ	ǵ	Ǵ	Ġ	ġ	Ĝ	ĝ	ǧ	Ǧ	Ğ	ğ	Ģ	ģ	ǥ	Ǥ	Ɠ
Ɣ	Ĥ	ĥ	Ħ	ħ	ƕ	ı	Í	í	Ì	ì	İ	Î	î	Ï	ï	ǐ	Ǐ	Ĭ
ĭ	Ī	ī	Ĩ	ĩ	Į	į	Ɨ	ȉ	Ȉ	ȋ	Ȋ	l	Ĳ	ĳ	Ĵ	ĵ	ǰ	ĸ
ǩ	Ǩ	Ķ	ķ	ƙ	Ƙ	Ĺ	ĺ	Ŀ	ŀ	Ľ	ľ	Ļ	ļ	ƚ	Ł	ł	ƛ	ǉ
ǈ	Ǉ	Ɯ	Ń	ń	Ň	ň	Ñ	ñ	Ņ	ņ	Ɲ	ŉ	ƞ	ǌ	ǋ	Ǌ	Ŋ	ŋ
º	Ó	ó	Ò	ò	Ô	ô	Ö	ö	ǒ	Ǒ	Ŏ	ŏ	Ō	ō	Õ	õ	ǫ	Ǫ
Ő	ő	Ɵ	Ø	ø	ǿ	Ǿ	ǭ	Ǭ	ȍ	Ȍ	ȏ	Ȏ	ơ	Ơ	ƣ	Ƣ	Œ	Œ
œ	œ	ƥ	Ƥ	Ŕ	ŕ	Ř	ř	Ŗ	ŗ	ȑ	Ȑ	ȓ	Ȓ	Ʀ	Ś	ś	Ŝ	ŝ
Š	Š	š	š	Ş	ş	Ʃ	ƪ	ƨ	Ƨ	ß	ſ	Ť	ť	Ţ	ţ	ƭ	Ƭ	ţ
Ʈ	Þ	þ	™	Ŧ	ŧ	Ú	ú	Ù	ù	Û	û	Ü	ü	ǔ	Ǔ	Ŭ	ŭ	Ū
ū	Ũ	ũ	Ů	ů	Ų	ų	Ű	ű	ǘ	Ǘ	ǜ	Ǜ	ǚ	Ǚ	ǖ	Ǖ	ȕ	Ȕ
ȗ	Ȗ	ư	Ư	Ʊ	Ʋ	Ŵ	ŵ	ƿ	Ý	ý	Ŷ	ŷ	Ÿ	Ÿ	ÿ	ƴ	Ƴ	Ź
ź	Ż	ż	Ž	Ž	ž	ž	ƶ	Ƶ	Ʒ	ǯ	Ǯ	ƹ	Ƹ	ƺ	ƾ	Α	α	Ά
ά	Β	β	ϐ	Γ	γ	Δ	δ	Ε	ε	Έ	έ	Ζ	ζ	Η	η	Ή	ή	Θ
θ	ϑ	Ι	ι	Ί	ί	Ϊ	ϊ	ΐ	Κ	κ	ϰ	Λ	λ	Μ	μ	Ν	ν	Ξ
ξ	Ο	ο	Ό	ό	Π	π	ϖ	Ρ	ρ	ϱ	Σ	σ	ς	ϲ	Τ	τ	Υ	υ
Ύ	ύ	Ϋ	ϔ	ϋ	ΰ	ϒ	ϓ	Φ	φ	ϕ	Χ	χ	Ψ	ψ	Ω	ω	Ώ	ώ
Ϛ	ϛ	Ϝ	ϝ	Ϟ	ϟ	Ϡ	ϡ	Ϣ	ϣ	Ϥ	ϥ	Ϧ	ϧ	Ϩ	ϩ	Ϫ	ϫ	Ϭ
ϭ	Ϯ	ϯ																

Figure 8. *Representative non-keyboard symbols for 'European languages'.*[12, 13, 14]

stepping outside the analysis to use the Vietnamese alphabet and the received ordering for tones) and also to decide that it is productive to take two simultaneous approaches in presenting variations for overloaded keys.

5. EUROPEAN LETTERS INPUT METHOD. The letters and symbols of European languages considered as a group (**Figure 8**) serves as an example of an input set that is not (strictly) associated with a single language and as a final example of the science and art of designing input methods.

Many of our users often switch among several languages frequently in their day-to-day use of productivity applications. Several factors (e.g. considerable keyboard familiarity, rapidity and frequency of language switching) indicate that explicitly switching languages would be a frustration compared to simply presenting keyboard symbol variations (*à la* dViệt-ngữ).

Using prior explicated methods, dLinguist™—MindStride's European language input method—uses P&H to access keyboard character variations. Even cursory examination of the range of graphemic elements in keyboard symbol variations indicates it unlikely that art can produce universal mnemonics. So, the design of dLinguist relies mainly on science to identify which variations are most frequent across keyboard symbols[15] and on art, where obvious, to associate the variations with the same subsequent key.

Hiragana (平仮名)

ぁ あ ぃ い ぅ う ぇ え ぉ お か が き ぎ く
ぐ け げ こ ご さ ざ し じ す ず せ ぜ そ ぞ た
だ ち ぢ っ つ づ て で と ど な に ぬ ね の は
ば ぱ ひ び ぴ ふ ぶ ぷ へ べ ぺ ほ ぼ ぽ ま み
む め も ゃ や ゅ ゆ ょ よ ら り る れ ろ ゎ わ
ゐ ゑ を ん ゔ ゛ ゜ ゝ ゞ

Count: 90

Katakana (片仮名)

ァ ア ィ イ ゥ ウ ェ エ ォ オ カ ガ キ ギ ク
グ ケ ゲ コ ゴ サ ザ シ ジ ス ズ セ ゼ ソ ゾ タ
ダ チ ヂ ッ ツ ヅ テ デ ト ド ナ ニ ヌ ネ ノ ハ
バ パ ヒ ビ ピ フ ブ プ ヘ ベ ペ ホ ボ ポ マ ミ
ム メ モ ャ ヤ ュ ユ ョ ヨ ラ リ ル レ ロ ヮ ワ
ヰ ヱ ヲ ン ヴ ヵ ヶ ヷ ヸ ヹ ヺ ・ ー ヽ ヾ

Count: 94

Education kanji (kyōiku kanji 教育漢字)

一 二 三 四 五 六 七 八 九 十 百 千 上 下 左 右
数 多 少 万 半 形 太 細 広 長 点 丸 交 光 角 計
主 乗 予 事 仕 他 代 住 使 係 倍 全 具 写 列 助
位 低 例 便 信 倉 候 借 停 健 側 働 億 兆 児 共
個 備 像 再 刊 判 制 券 則 効 務 勢 厚 句 可 営
冊 処 刻 割 創 劇 勤 危 卵 厳 収 后 否 吸 呼 善

Count: 1,006

Daily-use kanji (jōyō kanji 常用漢字)

丈 与 且 丘 丙 丹 乏 乙 乾 了 互 井 亜 享 亭 介
仙 仰 企 伏 伐 伯 伴 伸 伺 但 佐 佳 併 侍 依 侮

Count: 939

Name kanji (jinmeiyō kanji 人名用漢字)

馨 馴 馳 駕 駒 駿 驍 魁 魯 鮎 鯉 鯛 鰯 鱒 鱗 鳩
鳶 鳳 鴨 鴻 鵜 鵬 鶴 鷗 鷲 鷺 鷹 鹿 麒 麓 麟 麿

Count: 983

Total Kana: 184 // Official Kanji: 2928

日本漢字能力検定試験 ('Test of Japanese Kanji Aptitude') ~6000 kanji

Figure 9. *Representative non-keyboard symbols for Japanese.*

For example, each of the **acute** (ˊ) variations of keyboard symbols

{Á Ć É Ǵ Í Ĺ Ń Ó Ŕ Ś Ú Ý Ź á ć é ǵ í ĺ ń ó ŕ ś ú ý ź}

can be input by P&H on the letter, followed by typing the apostrophe.

But examination of the range of variations of (e.g.) the letters **E**, **T** and **S**,

{É € È Ê Ë Ė Ě Ĕ Ē Ȩ Ȅ Ê Ɛ Ǝ E Έ H Ή Ə}
{† ™ Þ Ť Ţ Ŧ Ƭ Ʈ Θ T ƭ}
{Ś § Š ß Ŝ Š Ş Σ ʅ Σ Ϛ ϡ ɰ}

indicates that a comprehensive set of mnemonics, even across just letters, would be so complicated as to defeat the purpose.

Given that any single user is likely to use only very few of the variations in the set, we have found that it is sufficient to rely on repetition to cover the learning curve, so art is neither useful nor necessary, as in the case of dViệt-ngữ.

6. THE LIMITS OF THE GRAPHEMIC APPROACH. Although full analyses and descriptions of input methods for Chinese, Japanese, and Korean are outside the scope of this introduction, it is an interesting, but only seemingly paradoxical, observation that the written languages which graphically vary most are more amenable to input methods based on phonemic (or even morphemic) rather than graphemic analyses.

To understand why phonemic analysis is useful for Chinese, Japanese, and Korean, note that estimates of the number of symbols ranges from 3112 (Official Japanese, **Figure 9**) to 80,000 (every Chinese character ever written, **Figure 11,** overleaf). For each of these ranges, all of the characters are associated with spoken languages that have phoneme counts which

Reprinted from *LACUS Forum 32: Networks*, edited by Shin Ja Hwang, William J. Sullivan & Arle R. Lommel. 2006. Houston TX, LACUS.

	Totals	Examples
Jamo (자모, 字母)	**51**	
Ja-eum (자음, 子音)	30	
Simple	*14*	ㄱ ㄴ ㄷ ㄹ ㅁ ㅂ ㅅ ㅇ ㅈ ㅊ ㅋ ㅌ ㅍ ㅎ
Double	*5*	ㄲ ㄸ ㅃ ㅆ ㅉ
Clusters	*11*	ㄳ ㄵ ㄶ ㄺ ㄻ ㄼ ㄽ ㄾ ㄿ ㅀ ㅄ
Mo-eum (모음, 母音)	21	
Simple	6	ㅏ ㅓ ㅗ ㅜ ㅡ ㅣ
Yotized	4	ㅑ ㅕ ㅛ ㅠ
Diphthongs	11	ㅐ ㅒ ㅔ ㅖ ㅘ ㅙ ㅚ ㅝ ㅞ ㅟ ㅢ
Obsolete Jamo	**43**	
Ja-eum	34	
Simple	*6*	ㅿ ㆁ ㆆ ㅱ ㅸ ㆄ
Double	*4*	ㅥ ㆀ ㆅ ㅹ
Double Cluster	*20*	ㅦ ㅧ ㅨ ㅪ ㅬ ㅭ ㅮ ㅯ ㅰ ㅲ ㅳ ㅶ ㅷ ㅺ ㅻ ㅼ ㅽ ㅾ ㆂ ㆃ
Triple Cluster	*4*	ㅫ ㅩ ㅴ ㅵ
Mo-eum	9	
Simple	*1*	ㆍ
Dipthongs	*8*	ᆜ ᆡ ㆇ ㆈ ㆉ ㆊ ㆋ ㆌ
Hangul	**11,172**	가 각 갂 갃 간 갅 갆 갇 갈 갉 힚 힛 힜 힝 힞 힟 힠 힡 힢 힣
Hanja (Official)	**1,800**	一 二 三 人 口 子 女 下 上 入 大 小 九 十 土 七 八 中 山 不

Overall Total: 13,066

Figure 10. *Representative non-keyboard symbols for Korean (한국어, 조선어).*

are orders of magnitude smaller. For instance, every Chinese character has one or more of approximately 1280 Mandarin syllabic pronunciations (including tone).

Using a phonemic approach to specify syllables before characters, 舞中文 ('WǔZhōngWén')—MindStride's Chinese input method—enables each syllable to be exactly specified in a maximum of three gestures, then individual characters in a maximum fourth, and words and phrases beginning with the character in a maximum fifth (via Press & Hold).

Graphical input methods based on either the (8) basic strokes or the (200+) composite parts (radicals, 部首 'bùshǒu') require many gestures (e.g. one gesture per stroke) or extraordinary learning curves to memorize associations of keys with components (e.g. 五笔字型 *wǔbǐzìxíng*).

7. CHINESE EXEMPLIFIES COMPLEXITIES OF OTHER LANGUAGES AND INPUT SETS. In developing input methods for a considerable number of languages, it has become apparent that the development of an input method for Chinese illustrates most—if not all—of the difficulties likely to occur in the development of input methods for any language or input set. A full discussion of general input method design issues and considerations as epitomized by 舞中文 can be found in MindStride's white paper, 'Opportunity in Work Clothes: The Development of MindStride's Chinese Input Method',[16] a précis for which follows:

The morass of factual and statistical information about the Chinese language (**Figure 11**) make it very difficult to choose which observations are useful in the design of an efficient, **elegant** input method for average, educated computer users, whether they are learning the

Chinese (汉语/漢語 - Hànyǔ, 华语/華語 - Huáyǔ, 中文 - Zhōngwén)

Pīnyīn Vowels (拼音元音)

ā	Ā	á	Á	ǎ	Ǎ	à	À
ē	Ē	é	É	ě	Ě	è	È
ī	Ī	í	Í	ǐ	Ǐ	ì	Ì
ō	Ō	ó	Ó	ǒ	Ǒ	ò	Ò
ū	Ū	ú	Ú	ǔ	Ǔ	ù	Ù
ǖ	Ǖ	ǘ	Ǘ	ǚ	Ǚ	ǜ	Ǜ

Count: 48

Zhùyīn Fúhào (注音符號)

ㄅ	ㄆ	ㄇ	ㄈ	ㄉ	ㄊ	ㄋ	ㄌ
ㄍ	ㄎ	ㄏ	ㄐ	ㄑ	ㄒ		
ㄓ	ㄔ	ㄕ	ㄖ	ㄗ	ㄘ	ㄙ	
ㄧ	ㄨ	ㄩ	ㄚ	ㄛ	ㄜ	ㄝ	
ㄞ	ㄟ	ㄠ	ㄡ	ㄢ	ㄣ	ㄤ	ㄥ
ㄦ	ˉ	ˊ	ˇ	ˋ			

Count: 41

Simplified Chinese characters (简体字 - jiǎntǐzì)
万与丑专业丛东两严丧个丰临为为丽举么义乌乐
Traditional Chinese Characters (繁體字 - fántǐzì)
萬與醜專業叢東兩嚴喪個豐臨為爲麗舉麼義烏樂

Count estimates range from 6,000 to 80,000.

Figure 11. *Representative non-keyboard symbols for Chinese.*

language academically or are native language speakers in China, Taiwan, Hong Kong, or Singapore (each of which presents specific complexities). However, once a set of useful observations is winnowed from those available—and appropriate analyses and inferences made—a number of unique opportunities present themselves:

- An elegant Chinese input method has the potential to invert the current **input efficiency deficit** between Chinese and English.
- Pedagogical tools based on the distinctions inherent in the input method have the potential to shorten learning curves for the written (and to some degree, the spoken) language itself.

In general, the specific arrangement of observations about Chinese determined to be useful in the development of 舞中文 ('WǔZhōngWén')—MindStride's Chinese input method—reduces complexity by an order of magnitude moving from words and phrases to (especially, keyboard) gestures[17]:

1. Each of **many tens-of-thousands** of common multi-character words and phrases begins with a single character from among **several thousands**. The number of words and phrases beginning with a particular single character is generally **very few tens**.
2. Each of the **several thousands** of individual Chinese characters—when used in context—is pronounced with one of **several hundred** unique syllables (if tone is ignored a **very few hundred**). The number of individual Chinese characters with a particular syllabic pronunciation is generally **very few tens**.
3. Each of the **several hundred** unique syllabic pronunciations for Chinese characters begins with one of **very few tens** of (approximately twenty) initials, participates in one of **very few tens** of (approximately twenty) rhymes, and has one of **five**

tone values. The number of syllabic pronunciations for a particular initial or rhyme is generally **very few tens**.

4. Most of the **approximately twenty** syllabic initials (and all of the **five** tone values) are readily associated with keyboard symbols. Most of the **approximately twenty** syllabic rhymes are *not* readily associated with keyboard symbols.

Inferences related to these observations, taken in reverse order, are suggestive and can be taken as input method design goals; it should be generally possible to:

1. Specify a pronunciation with two gestures,
2. Specify a character with a single additional gesture from the pronunciation, and
3. Specify a word or phrase with an optional second single additional gesture from the character,

with each gesture (e.g. keystroke, mouse selection, chord) being selected from a small set.

Applying art and science makes these theoretical design goals practical by user friendly presentation of choices and associations of gestures leading to selection of pronunciations, then characters, then words or phrases.

The central issue in designing for users is the balancing of immediate ease of use with what must be learned for long-term productivity gains. The most salient example in 舞中文 is specification of the syllabic pronunciation with two keyboard gestures: short-term learnability is gained via mnemonic keyboard mappings for syllabic initials (and optional tones, if used), but long term productivity is achieved via learned portmanteau associations of single keys with entire syllabic rhymes where other systems require multiple (but more mnemonic) keys. The learning curve is mitigated by there being only 20 or so non-intuitive portmanteau keystrokes to learn.[18]

By reducing selection of most Chinese words and phrases to four reasonably readily learned gestures, words and phrases can be selected in fewer keystrokes than required on average for English (assuming the average word in an English typed document is between 5 and 7 letters). It is intriguing to speculate that—once the learning curve is climbed—it may be faster to type Chinese than English.

A side note: MindStride is developing a training game that quickly steps users through learning the keys associated with mandarin syllabic rhymes as well as (we hope) making it engaging to rapidly develop motor memory for gesture sequences associated with common characters, words, and phrases. Although further research is necessary, in discussing the game design with Chinese language teachers, we determined that (1) using rhymes instead of individual (e.g. pinyin) letters may be beneficial in rapidly learning standard pronunciations with fewer errors and (2) teaching gesture sequences associated with common characters, the first two of which indicate pronunciation, may have benefits toward character memorization in general. If these results are borne out by further research, then MindStride input methods may be valuable not only as input tools, but also as additions to language teaching curricula.[19]

Reprinted from ***LACUS Forum 32: Networks***, edited by Shin Ja Hwang, William J. Sullivan & Arle R. Lommel. 2006. Houston TX, LACUS.

	Keys	Ratio	Meters
English	101	1	0.25
Spanish	121	1.20	0.30
Viet	225	2.23	0.56
Euro	609	6.03	1.51
Japanese	6000	59.41	14.85
Korean	13066	129.37	32.34
Chinese	50000	495.05	123.76

Figure 12. *Estimated physical length of keyboards for various languages.*

1 A publication and revision history of this document is available on request. Please contact the author at *MindStride@gmail.com* or *dmailman@rice.edu*.

2 Throughout the paper, the term 'emic' is used in its popular sense to indicate something that makes sense to a native speaker and is 'intuitive', rather than its related, but technically more accurate, linguistic meaning.

3 E.g. Japanese (four scripts): Kanji, Hiragana, Katakana and Romaji.

4 The first objection to using input-set-specific keyboards to reduce input efficiency deficits is the increasing tendency for users to require more than one input set. For example, there are keyboards available for the Spanish language that associate keys with its non-(QWERTY)-keyboard symbols. If an additional language—even English—is required, any input efficiency gains due to the keyboard are negated. The other obvious objection is that for many input sets, a keyboard is simply impractical. If the input method problem reduced to making a keyboard for each input set, the result would be keyboards of the lengths indicated in **Figure 12**.

5 Most modern operating systems (e.g. Linux, Windows, and Macintosh OS X) have input methods for non-English languages; and it is these that we use to estimate productivity gaps. There are many resources that can be read for background on input methods; three suggestions (thanks to Arle Lommel) are Cahill (2003), O'Hagan and Ashworth (2002) and Lunde (1999).

6 'Fewer gestures' has two meanings: (1) Fewer junctures in the sequence leading up to an input and (2) Fewer choices at each juncture in a sequence.

7 Many thanks to Diane Campbell, the co-inventor of Press & Hold.

8 Indeed, we view 31 as the upper limit of the number of simultaneous choices from which a user can quickly learn to choose with anything approaching reasonable speed (speed of learning as well as speed of selection).

9 Pinyin initials are a good example: should keyboard mappings be presented in alphabetical order, physical layout order, BPMF order, corpus frequency order, or dictionary frequency order?

10 A stroke is the diacritic that occurs in **đ**; other terms are illustrated in **Figure 6**.

11 'First Tone' is indicated by no mark on the vowel.

12 As nearly as we can tell, the 500+ characters listed, together with the common keyboard characters, comprehensively account for the following languages: Afrikaans, Basque, Breton, Catalan,

Reprinted from ***LACUS Forum 32: Networks***, edited by Shin Ja Hwang, William J. Sullivan & Arle R. Lommel. 2006. Houston TX, LACUS.

Coptic, Croatian, Czech, Danish, Dutch, English, Esperanto, Estonian, Faroese, Finnish, Flemish, French, Frisian, German, Greek, Greenlandic, Hungarian, Icelandic, Irish, Italian, Latin, Latvian, Lithuanian, Maltese, Norwegian, Polish, Portuguese, Provencal, Rhaeto-Romanic, Romanian, Romany, Sami, Slovak, Slovenian, Sorbian, Spanish, Swedish, Turkish, Welsh and a few others. Please email me if you find something missing for any of these languages or if another language should be added to the list.

13 In input methods, we generally elect to leave out blended letters that are composed of two simple keyboard characters because they can be input by typing the letters individually (a skill already developed by average, educated users).

14 Some characters in the table appear to be duplicates, but are in fact treated differently by different productivity applications, so dLinguist enables input of both. One example is the characters:
Đ (Unicode #0110, 'Latin capital letter D with stroke') in Croatian, Sami, and Vietnamese
Ð (Unicode #00D0, 'Latin capital letter eth') in Icelandic, Faroese, Old English, and IPA

15 Given the nature of the problem it seems unlikely that a corpus can be identified to make it possible to do the analysis to determine which variations occur most frequently across languages or are most frequent across users. Please email me if you have any suggestions as to how these may be accomplished.

16 The title is paraphrased from Thomas Edison's comment on invention: 'Opportunity is missed by most people because it comes dressed in overalls and looks like work.'

17 Secondary information—useful in defining an order for presenting characters associated with pronunciations—takes the form of character frequency lists. Information regarding the visual components of characters is ignored.

18 It should be noted that for any device other than a keyboard, 舞中文 requires no long-term learning curve and reduces the number of gestures per word or phrase dramatically.

19 Please email me if you have an interest in, or suggestions for, researching MindStride input methods' or training games' efficacy as a pedagogical tool.

REFERENCES

CAHILL, DAVID. 2003. Lookup methods for Chinese characters: Electronic versus paper media. *LACUS forum* 29:61–72.

O'HAGAN, MINAKO & DAVID ASHWORTH. 2002. *Translation-mediated communication in a digital world: Facing the challenges of globalization and localization.* Clevedon UK: Multilingual Matters.

LUNDE, KEN. 1999. *CJKV information processing*. Beijing: O'Reilly.

Reprinted from ***LACUS Forum 32: Networks***, edited by Shin Ja Hwang, William J. Sullivan & Arle R. Lommel. 2006. Houston TX, LACUS.