The Linguistic Foundation of Input Method Editors

The Case of R^2

Ming Hsu*

Division of Humanities of Social Sciences

California Institute of Technology

Pasadena, CA 91125

mhsu@hss.caltech.edu

Yun Hsu

Department of Mechanical Engineering

California Institute of Technology

Pasadena, CA 91125

Introduction

A Nokian Parable

Imagine that in an alternative universe, there exist the Nokian people. Though mute, the Nokians are nevertheless a culturally and scientifically advanced people. Currently, the Nokians are in the midst of a great technological revolution, powered by the invention of what we would call the computer. Productivity has boomed, unemployment is down, and prosperity abounds. There is one serious problem, however. The Nokians have had to resort to using an artificially generated language to interact with their computers.

* Indicates contacting author.

The "problem" lies with the Nokian written language. By rearranging the letters in different positions, the words can potentially have many different meanings, even in words that consist of the same letters. They are what we would call *logographs*. A common form of greeting, written slightly different, becomes a curse. Clearly, a linear typing system, such is the form of the new language created to converse with computers, cannot be used directly for Nokian.

Hi, a form of greeting Die, to be dead

Soon, the Nokian leaders began to worry about the negative effects of this new invention. They claim that this new written system is exacerbating the divide between rich and poor. Some have even begun to decry the effects the computer is having on the rich Nokian cultural and literary heritage. It would seem that the invention of an input system for the native Nokian language is of great importance. But with none in sight, O! what is a good Nokian to do?

The Dilemma

This dilemma, in a similar (albeit less exaggerated) form, is precisely the one faced by Chinese speakers today[1]. In this case, the multiplicity of homonyms, rather than an inability to speak, is the obstacle to a phonetic system.

The impact of a less than ideal input system is difficult to quantify statistically. The existence, in contrast, can be seen in many places. We give one example here. Whereas

[1] This is also the problem faced by phonetic languages when the set of inputs of the input device is limited, such as a cell phone. Systems such as i-Tap and T9 can be seen as methods to get around these limitations.

Chinese speakers account for 14.1% of Internet users, Chinese content accounts for only 2.4% of webpages [1]. Putting it differently, there are more Chinese *users* than French and German combined, but the amount of Chinese *content* is less than either.

This problem persists certainly not from a lack of effort. More than 1,000 input methods have been invented by 2003, according to one estimate [2]. Some notables include the phonetic system *pinyin* and its Taiwanese cousin *zhuyin*. Another variant parses the Chinese characters into elements called *radicals*—e.g., *Wu-bi* and *Cangjie*. There are also those that parse Chinese characters into basic *strokes*—e.g., *Q9*. The list is ever expanding.

The Conceptual System

Necessary Conditions

In this paper, we present a new input method, which we call R^2. We designed it with special attention to the unique linguistic properties of the Chinese language. This method solves a number of outstanding issues with existing solutions. Perhaps even more importantly, we also present a conceptual framework with which to judge current and future methods. A conceptual framework is vital because it offers a way to organize and assess the various systems of in the face of a seemingly endless parade of potential solutions.

Previous Attempts

In this endeavor we follow the seminal work published more than 15 years ago in this journal by Qiao et al. [3]. In it, the authors proposed a series of four criteria that all Chinese input systems should possess:

1. Versatility: There have been many forms of the Chinese language over the years. The most prominent being the simplified/traditional divide that mirrors the PRC/ROC divide. A good input system should be equally adept at dealing with any of these widely used forms.

2. Standard encoding method: There should be a common encoding method for the language, irrespective of font style or other textual representation on screen. This problem has been solved by the development of Unicode.

3. One Code, One Character (OCOC): There should be a one-to-one mapping between the set of inputs to the set of outputs.

4. See Character, Know Code (SCKC): A good encoding system should be intuitive without requiring the user to memorize or possess a vast corpus of information prior to usage.

Of the four criteria, (3) and (4) have been the most difficult to satisfy simultaneously. *Wu-bi* and *Cangjie* satisfy (3), but fail (4). Phonetic systems satisfy (4), but fail (3) due to the homonym problem. *Q9* satisfy (4) but fail (3) as well, but for different reasons.

On the basis of these four conditions, Qiao et al. proposed the *6-Digit-Coding-System*. For a variety of reasons, it did not become widely adopted. Here, we synthesize some lessons learnt with the benefit of 15 years of hindsight, and discuss some of the difficulties of the *6-Digit* as well as other methods. First however, we reorganize and extend Qiao et al.'s conditions. Unlike their conditions, ours were created with the aim of encompassing *all* input systems.

The Current System

1. Monotonic Uniqueness: Define the input system as the domain, and the characters the range, and the input system a set-valued function (correspondence) that maps from the domain to the range. A system is said to exhibit monotonic uniqueness if, for each element y in the range, there exists at least one n-tuple $x = (x_1, x_2, \ldots, x_n)$ s.t. $f(x) = A$, where A = {y}. Furthermore, for each (n-k)-tuple, where $x_k = (x_1, x_2, \ldots, x_{n-k})$ and $k \leq n$, $f(x_k) = A_k$, where $A_k \supseteq A_{k'}$ if $k < k'$.

   In words, this condition implies that, with each additional input, the set of feasible words/characters is a (weak) subset of the previous set, culminating in a singleton in a finite number of steps. This condition is in the spirit of the OCOC condition of Qiao et al.

   Finally, the "speed" of the language can be measured by the rate of convergence to the singleton.

2. Reflexivity (One key-one representation): This requires that each input have a unique output. The keypad on telephones, for example, violates this condition.

3. Naturalism (Follow natural divide of language): for phonetic input, this consists of phonemes or morphemes. For structure-based systems, it may be strokes or radicals. Order of input should follow those of the language. Phonemes that come later in speech should be inputted later. Strokes or radicals that are written later should be inputted later.

4. Compactness (A "reasonably" set of input units): This condition merely ensures that one does not have an unreasonably large number of input units, such as one key per word/character.

Violation Under Current Methods

Current QWERTY keyboards satisfy these four criteria for all Indo-European, and other phonetically based languages. On the other hand, none of the current Chinese input methods satisfy all four. Table 1 lists the compatibility of some notable methods with our conditions.

System	***Natural***	***Reflexive***	***Unique***	***Popularity***
Pinyin	✓	✓	×	High
Q9	✓	✓	×	Medium-High
Wu-bi	✓	×	✓	Low-Medium
Cangjie	✓	×	✓	Low-Medium
Six-Digit	×	✓	✓	Low
Four-Corner	×	✓	×	Low

Table 1: Compatibility of various input systems with our necessary conditions.

Our conditions can be thought of as conditions on the standard concept of model/view/controller in computer science. Here, the model is language itself, the controller the input system, and the view the output on the screen.

1. Uniqueness is a statement about the result of the mapping from the controller to the view. Namely, it states that the result of the mapping should be a singleton.

2. Reflexivity concerns the *state* of the controller. It states that the elements of the controller are to be composed of singletons.

3. Naturalism states that the mapping between the model and the view should obey the linguistic structure of the model underlying the particular input method.

4. Compactness is a constraint on the controller, most frequently imposed externally

Figure 1 depicts the above four points in graphical form.

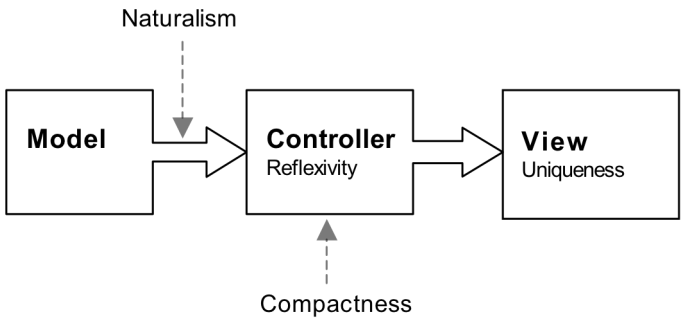

Figure 1: Relation of our conditions to model/view/controller.

Linguistic Foundations

It should now be apparent that the linguistic properties of a language play a crucial role in how one might want to design an input system. What is "natural" for English may well not be natural for Chinese. Similarly, a system that is Unique under English may well not

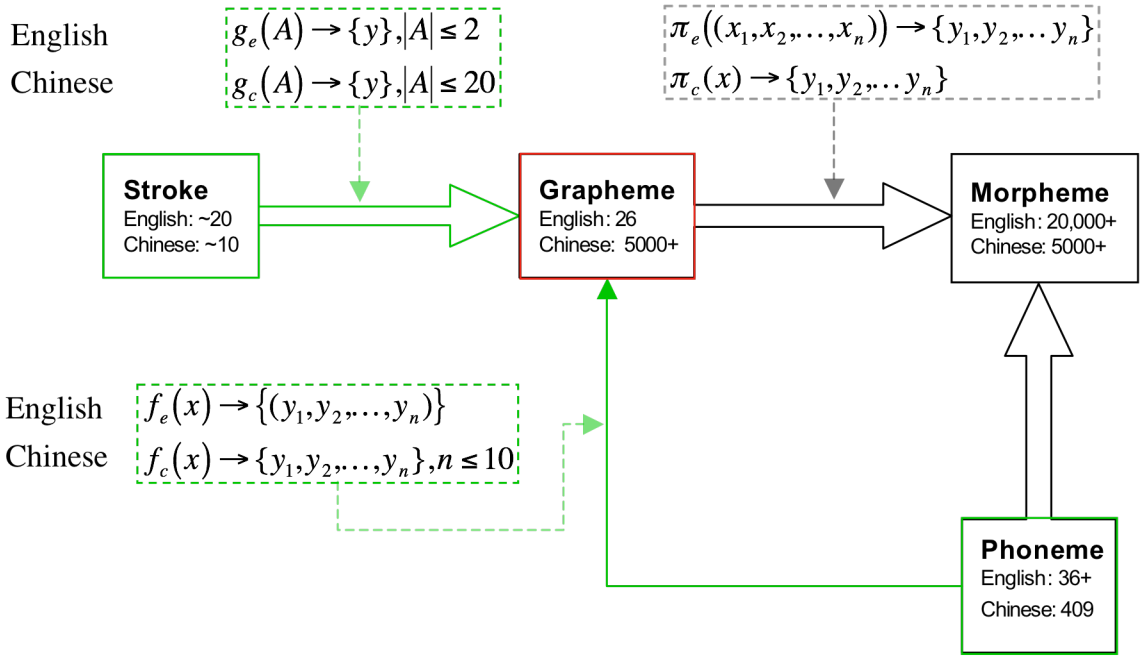

Figure 2: Solid boxes represent the different linguistic components of a language. The number of elements of a component is given inside each box. Large arrows denote the relationship between the components. Dashed boxes contain definitions of the mapping between the components.

be in Chinese. In this section we explore the unique properties of Chinese and the constraints it places on the input system.

Definitions

We begin with definitions. A language can be processed by its phonetic and visual properties. The basic unit of sound is the *phoneme*. An example of a phoneme is "*a*", or "*ba*". Phonemes combine to produce a *morpheme*—the smallest unit of language that carries meaning.

In writing, the basic unit is the *grapheme*. This, and only this, is what input systems need to produce. Together, the graphemes form to make morphemes. Finally, and apparently ignored by the linguists, as it falls under the domain of typography, each grapheme is composed of *strokes*. Figure 2 depicts the relationships between the various linguistic components. The sets and the mappings between them illustrate the commonalities and difference between Chinese and English.

Comparison between Chinese and English

In both languages (and most others), a morpheme consists of an *n-tuple* of phonemes. In contrast, a grapheme consists of a *net* of strokes. A net is a generalization of an n-tuple, and is considerably more complicated. It requires multiple relations to describe the spatial arrangement between its elements. For example, the letter *x* is composed of two intersecting strokes. In Chinese, the number of strokes varies widely, from 1 to easily over 20, with a frequency-weighted mean of 9.1 for traditional Chinese characters (used in Taiwan and Hong Kong), and considerably less for simplified characters (used in China proper) [5].

From the perspective of an input system, one crucial distinction between the two languages lies in the immense difference in the size of the set of graphemes. As Figure 2 shows, the set of graphemes in English are the letters of the alphabet—26 in total. In contrast, Chinese has more than 3000 *commonly used* graphemes, and many thousands more. This is due to the one-to-one relation between graphemes and morphemes in Chinese.

The difficulty is now clear. Whereas in English it is feasible to work *directly* with graphemes, in Chinese, one is forced to choose between the *indirect* route of phonemes, or the direct, but difficult route, of nets. These two routes are represented by the green arrows in Figure 2.

Challenges of Chinese Input

The indirect route from phonemes to graphemes is illustrated by the function $f(\cdot)$ in Figure 2. In a phonetic language this process is relatively simply: each phoneme maps to a unique n-tuple of graphemes. Recall however, in Chinese, each grapheme corresponds to one morpheme. This results in a much larger set of graphemes than there exists phonemes, which numbers 403 [6]. Thus the elements of the range are not singletons, and can in fact be quite large sets, as the distribution across phonemes is highly skewed. This is the *homonym problem*. It is the source of the violation of the Uniqueness condition, and is an inherent limitation of *any* phonetic entry of Chinese graphemes.

The direct route from strokes to graphemes is illustrated by the function $g(\cdot)$ in Figure 2. This mapping is complicated by the nature of graphemes—sets of strokes with up to three relations and their complements (up/down, left/right, inside/outside). So far as we know, no Chinese input systems implements these relations. Qiao et al. comes the closest by

considering the spatial arrangement of Chinese graphemes. Their solution, however, imposes a structure unsuited for the Chinese grapheme. They require input to follow a grid of 6 squares in a pre-specified order, which often violates the linguistic order of input—a violation of Naturalism. One can see this in the example given in their paper.

Most other input methods either do not recognize this aspect, or shy away from it. Instead, they transform the nets to n-tuples, resulting in a loss of information. They then must resort to *ad hoc* methods that try to retrieve back this information, which inevitably leads to a violation of either Uniqueness of Reflexivity. This is the main contribution of our system and input method, R^2. By modeling explicitly the relations amongst the strokes, we retain the information about the structure of the set without resorting to *ad hoceries*.

The R^2 system

Description

Here we describe our new input method, which we call R^2 to emphasize the two dimensional nature of the Chinese grapheme. Denote the three relations (Up, Left, Inside) respectively, as $(\uparrow, \succ, \otimes)$. All three relations are complete ($\forall x, \forall y, x \uparrow y$ or $y \uparrow x$), anti-symmetric ($x \uparrow y \Rightarrow \sim y \uparrow x$), and transitive ($x \uparrow y$ and $y \uparrow z \Rightarrow x \uparrow z$)[2].

The Rules of R^2

In practice, however, it is easier to work with both the relations and their complements. That is (Up/Down, Left/Right, Inside/Outside).

- **Right** creates a new radical to the right of the existing writing. Move everything that had been written to the left.

[2] Intersections are defined as UP and ~UP, LEFT and ~LEFT, or INSIDE and ~INSIDE.

- **Down** creates a new radical to the bottom of the existing writing. Move everything that had been written up.
- **Up** creates a nested level within the radical.
- **Left** creates a nested level to the right of the radical
- **Inside** goes inside the radical
- **Outside** moves outside the radical
- **Intersection** Many Chinese characters involve intersection of two or more strokes. In R^2 this is achieved through simultaneous pressing of two or more strokes.

These functions mirror the basic rules of the Chinese written system (CITE). They are, in order of precedence,

1. Horizontal before vertical
2. Left before right
3. Top before bottom
4. Outside before inside
5. Top right before bottom left
6. Dot in top right is written last

We implement this system in Figure 3.

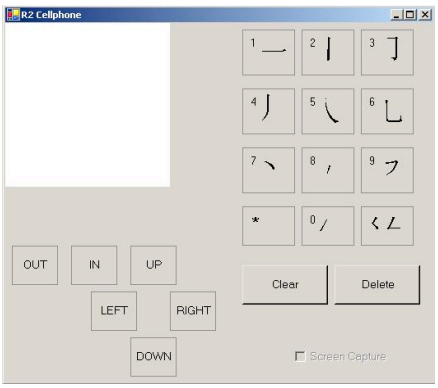

Figure 3: Example program implementing the R^2 language.

It is trivial to see that R^2 satisfies Naturalism, Reflexivity and Compactness. Uniqueness is satisfied for virtually all characters.[3]

An Example

We write the grapheme 成 (success). We chose it because, although it consists of only six strokes, it requires all three relations and their complements to determine it completely.

Start with 横	Push up	Put in 撇
Go inside, the 横 and 撇 are completely determined to form the 厂 radical.	Put in 横折	Push 横折 to the left
Put in 捺 and simultaneously push it	Put in 撇, and simultaneously push it	Go outside and put in the 点。The character

[3] There are a handful of characters that are can be distinguished only by the relative length of the strokes. For example, 土 and 土. This, as one can imagine, is a tiny minority.

up to make it intersect the 横 above	up to make it intersect the 捺 above.	is completely determined to be 成

Table 2 Sequence of stroke and location to construct the logograph 成.

Comparison of R^2 and Previous Methods

At first glance, it may appear Naturalism is the most controversial of the four conditions. After all, human beings are adaptive, and are likely to be able to learn whatever input system if it is required of them.

Here, we clearly benefit from the 15 years separating Qiao et al. and us. Note that the most widely used systems (*pinyin*, *Q9*) are Natural, Reflexive and Compact, but not Unique. Systems that are Unique, Reflexive, and Compact, but not Natural, have either failed to catch on, or limping along with a small user group. This group includes Qiao et al's *6-Digit* method.

This should not be surprising. Naturalism is the sole connection between a language and the input system. Any input system that ignores the linguistic structure of the language is unlikely to be intuitive for the user—a point we have taken pains to emphasize throughout this paper.

On the other hand, systems that violate either reflexivity or compactness tend to be complicated and difficult to learn. A good example of this is the *Wu-bi* method, which is natural and unique, but not reflexive. The fact that this method is still used by quite a few is a testament to the human ability to master complicated and difficult algorithms.

Finally, a user's choice of input system can be rationalized by the ordering of her preferences over the four conditions. This is both interesting and unforeseen result.

There are those who exhibit $\text{Naturalism} \succ \text{Reflexivity} \succ \text{Uniqueness}$[4] ($x \succ y$ iff x is strictly preferred to y). This is the majority of users, and they use either *pinyin* or *Q9*[5].

Another, less common preference is $\text{Naturalism} \succ \text{Reflexivity}$ and $\text{Uniqueness} \succ \text{Reflexivity}$. The best system for them in the set is one that is Natural and Unique. *Wu-bi* and *Cangjie* falls into this category.

A small minority show the preference profile $\text{Uniquenss} \succ \text{Reflexivity}$ and $\text{Uniqueness} \succ \text{Naturalism}$. These are the users who would use *6-Digits*.

Conclusion

Textual input is arguably *the* most fundamental aspect of the human-computer interaction process. The lack of an adequate input method is worrying because of its impact on the adoption and progression of Chinese computing and Internet. At a time when China is growing rapidly to catch up with the developed nations of the world, this is a weighty and unnecessary hindrance.

We believe this problem has persisted so long because of a lack of understanding of the unique linguistic properties of the Chinese language. In this paper, we provide such a basis. This allows us to develop a general conceptual framework with which to assess *all* input systems. We then relate this to concepts in both computer science and linguistics.

[4] We ignore compactness as it is imposed exogenously on the system.
[5] Note that *pinyin*, a phonetic system, and *Q9*, a structured system, are fundamentally different. Under our conditions, however, we see that they are popular for the same reasons.

Only after this were we able to develop a new method that reflects the special challenges of the input of the Chinese language.

We leave the final word to Donald Knuth, who pioneered what is perhaps the Western equivalent of the Chinese input system—free and transparent typesetting:

> I do strongly think that people, when they start throwing computers at something, they think that it's a whole new ballgame, so why should they study the past. I think that is a terrible mistake… But I don't think responsible computer scientists should be unaware of hundreds of years of history that went before us.

Reference

1. Global Internet Statistics. http://global-reach.biz/globstats
2. Wang, Jian (2003) Human-computer interaction research and practice in China. Interactions 10:2, 88-96.
3. Qiao, J., Qiao, Y. and Qiao, S. (1990) Six-Digit Coding Method. Communications of the ACM 33:5, 491-494.
4. Bett, Steve (1999) Can we pin down the number of phonemes In English. Simpl Speling Newsletter. p. 7.
5. Tsai, C.H. (1996) Frequency and Stroke Counts of Chinese Characters. http://technology.chtsai.org/charfreq.
6. Xin Hua Dictionary, 6th ed. Shang Wu Press, Beijing, 1987.
7. Knuth, Donald (2000) Interview: Donald Knuth. http://www.advogato.org/article/28.html.