To appear in the International Journal of Computer Processing of Oriental Languages, Vol. 13, No. 4, pp. 351-375.

EXPERIENCE IN DESIGNING DATABASES FOR LEARNING CHINESE CHARACTERS

H.C. LAM, W.W. KI, A.L.S. CHUNG and P.Y. KO

Department of Curriculum Studies, The University of Hong Kong,

Pokfulam Road, Hong Kong

Web site: `http://www.dragonwise.hku.hk`

Email: `chincal@hkusua.hku.hk`

There is much regularity in Chinese characters. For example, the characters **車**(car), **輪**(wheel), **軌**(rail), **輸**(transport), **載**(carry), **轉**(turn) and **轟**(the roar of train) are all in relation to "transportation" and at the same time share the component **車**. An interactive database that can retrieve these characters as linked together can enable students to explore and discover the relationship among the characters. This paper reports our many years of works in designing and developing such databases for learning Chinese characters. Several interesting issues on the implementation are raised, including intuitive interaction, demand of high accuracy, codification of components, decomposition of characters, handling multiple acceptable values, irregularity of Chinese characters and application-oriented development. Unlike most technology-centric research, much of our experience has been gained through an iterative development of the databases with actual contents and in collaboration with the Chinese educators. In this paper, we hope to propose a more humanistic and contextual approach into the study of Chinese computing.

Keywords: Learning Chinese Characters, Character Component, Computer Assisted Language Learning, Knowledge Representation, Chinese Database and Information Retrieval

1. Introduction

Like many other researchers, we were first fascinated by the incredible and non-trivial regularity that exists among Chinese characters. As an example, relating to water, the characters for pond(**池**), lake(**湖**), stream(**江**), river(**河**), sea(**海**) and ocean(**洋**) all contain a component in common - a three-dot component denoting the meaning of "water". But most researchers have only used Chinese character as a way of illustrating the potential of the technology, rather than tackling some domain specific problems with the use of the technology. In one of the studies, [2] has experimented with the application of available Artificial Intelligence techniques, like Semantic Networks, to analyze and interpret the hierarchical structures of Chinese characters. Along the same line, the Chinese dictionary system in [3] applies Fuzzy Logic technique into the handling of uncertain information to

help students to search for desirable characters. A further example is [4], which attempts to use a heuristic approach to generate Chinese character glyph from components automatically, based upon the rules of Chinese calligraphy. The limitation of these studies is an over-simplification of the contextual domain due to an excessive focus on demonstrating the power of a general-purpose technological principle, thus failing to look into the intricate needs of the domain [5]. As contrast to this, we started off with the study of pedagogy to improve the students' learning of Chinese characters, regarding the technology as only one of the means to achieve our goal. Narrowing to this specific context about learning, we look back and examine the implications upon the design of the technology.

Commonly used Chinese characters amount to a few thousands. Learning the characters by rote memorizing is not only boring but also ineffective. Instead Chinese characters are themselves interrelated. Several characters may share a common component and thus sound alike, for example, the characters **青**(cing1), **清**(cing1), **蜻**(cing1), **請**(cing2), **情**(cing4), **晴**(cing4) and **睛**(zing1). This kind of picto-phonetic composition **形聲字** actually covers 85% of all modern Chinese characters. Making this regularity of the characters explicit to students would certainly enhance their learning of the characters. This can be achieved through the provision of an exploratory environment, in which the students can explore the relationships among the characters on their own. By so doing, the students can learn a particular character by connecting the character to their preexisting knowledge of other characters. The students can also expand their knowledge via the links from the particular character to other unfamiliar characters. Simply put, the more the relationship to a character the students know about, the more firmly they can remember the character.

Besides this, teachers with the help of such a connection of characters can also easily highlight to the students the similarity among the characters that the students have already learned. Using this as a consolidation activity would probably help the students to memorize the characters through understanding the functions and meanings of the components inside the characters. This way of teaching and learning can clearly be supported by an interactive database of Chinese characters, which can retrieve related characters together.

With this in mind, we began to design and develop such databases since the late eighties [6][7][8]. Over the years, many generations of prototypes have been produced, revamped and re-developed [9][10][11]. The recent generation of product has been published and distributed to all primary schools in Hong Kong [12][13][14][15][16]. This paper is a reflection on the lessons we have learned during this many years of iterative development. We will report the problems we encountered and the ways to deal with them. The rest of the paper will be organized as follows: next section will describe the basic functions of the database, setting the scene for the subsequent discussion; then the succeeding section will discuss many of the issues brought up during the development of the databases. After that, the database structure that we have experimented with to store five hundred Chinese characters

will be illustrated.

2. An Overview of the Chinese Character Database

As one of the Dragonwise Series **現龍系列** software, the Chinese Character Database is designed for pre-school and early primary school students learning Chinese as the first language. Inside the database is stored the first five hundred Chinese characters listed in the curriculum. The properties of the characters have been determined, including the structures **結構**, stroke sequences **筆順**, radicals **部首**, pronunciations **讀音**, liushu's **六書**, morphological **形符**, phonological components **聲符**, the meanings **意義** and the usage **運用** of the characters. These characters are further linked up together to form a network via their common properties. For example, with the same speech sound, homophones are connected together such that students can go from one to all of the others. Similarly desirable characters can be located by other properties as shown below.

(pronunciation "cing1") ↓ **青** and **清**

(radical "**口**") ↓ **口**, **可**, **右**, **叫**, **只**, **句**, **同**, **吃**

(Customized cluster) ↓ **林**, **本**, **休**, **困**, **採**, **相**, **森**, **集**

Figure 1. Locating characters by sound, radical and customizable clusters

The characters containing a particular component can also be retrieved. As shown below, selecting **門** as a component **構件** will produce a list of characters: **門**, **們**, **閃**, **問**, **悶**, **開**, **閒** and **闊**. All of them contain **門** as a part of the characters. The students can further specify the character to be a picto-phonetic composition **形聲字**. Only the five characters **們**, **問**, **悶**, **開** and **闊** will then remain in the list. In addition to this, the students can further require the component **門** to be morphological in the characters. The list will be reduced to **開** and **闊**, which are all related to **門** (door) in meaning as **開**(open) and **闊**(wide). The search of the component **門** can also be changed into being phonological. The list will then become **們**, **問** and **悶**, which are all spoken like **門** (mun4) as **們**(mun4), **問**(man6) and **悶**(mun6) respectively.

To appear in the International Journal of Computer Processing of Oriental Languages, Vol. 13, No. 4, pp. 351-375.

(**門** as one component)

↓

門, **們**, **閃**, **問**, **悶**, **開**, **閒** and **闊**

↓

(“picto-phonetic composition **形聲字**)

↓

們, **問**, **悶**, **開** and **闊**

↙↘

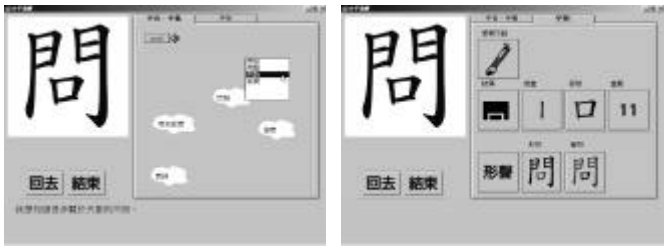
(morphological component **形符**) (phonological component **聲符**)

↓ ↓

開 and **闊** **們**, **問** and **悶**

Figure 2. Selecting characters by different criteria on the components

Clicking one of the characters in the list, say **問**, will transport the students to another window that details all the particulars of the character **問**. The left window of the figure below shows the four different possible meanings and senses of the character **問**, namely, **問答**(question and answer), **學問** (academic attainment), **問候**(showing concern) and **明知故問**(ask even knowing). Since most ancient Chinese characters have evolved from a single-character word into the modern multiple-character word, the meanings of a character should better be illustrated with the uses of the character in words. Thus, moving over one sense gives a list of words, in which the character is used in the same sense. For example, the list starting with **問答**(question and answer) includes the words **問號** (question mark), **問題** (question) and **疑問** (query), in all of which the character **問** means “questioning”. Selecting a word will pronounce the word and will give an example of a sentence using the selected word as shown below.



To appear in the International Journal of Computer Processing of Oriental Languages, Vol. 13, No. 4, pp. 351-375.

問
- ↗ **問答**, **問號**, **問題** and **疑問**
- ↗ **學問**
- ↘ **問候** and **慰問**
- ↘ **明知故問**, **問**, **問起**, **問路**, **訪問**, **發問**, **詢問** and **請問**

An example sentence: **我寫信問候姑母。**

Figure 3. Exposition of the character **問**

The morphological and phonological components **口** and **門** of the character **問** are shown with different highlights in the right window above. Clicking the component **口**, the students can return to the first window with the morphological component **口** pre-selected, producing another list of characters **問**(ask), **啼**(cry), **喊**(shout), **喝**(drink) and others, most of them are actions with the mouth(**口**). Moving back and forth, the students can hence explore lists of related characters, for example, from **門** to **問** to **嗎** to **媽** to **妹** and so on so forth.

門 → **問** → **嗎** → **媽** → **妹** → …

Figure 4. Exploring a list of related characters

3. Discussion

Designing such a database involves making many decisions, judging and balancing the pros and cons of various approaches. The following reports our decisions and the underlying rationale leading to the decisions, which have been boiled down into seven principles.

3.1 The Principle of Intuitive Interaction

Since most teachers using the database are not going to be highly technically competent, ease of use has become one of the foremost important guiding principles in designing the databases. Over the many generations of the developed prototypes, much effort has been put to simplify the operation of the databases, especially in the way students interact with the structural components of the characters. The early prototypes have adopted a rather technology-centric Binary Tree approach, in which each component of a character is divided into at most two sub-components. Students can traverse the structure of the character by choosing to go to the Parent, Left Child and Right Child component by selecting the corresponding buttons as shown below.

To appear in the International Journal of Computer Processing of Oriental Languages, Vol. 13, No. 4, pp. 351-375.

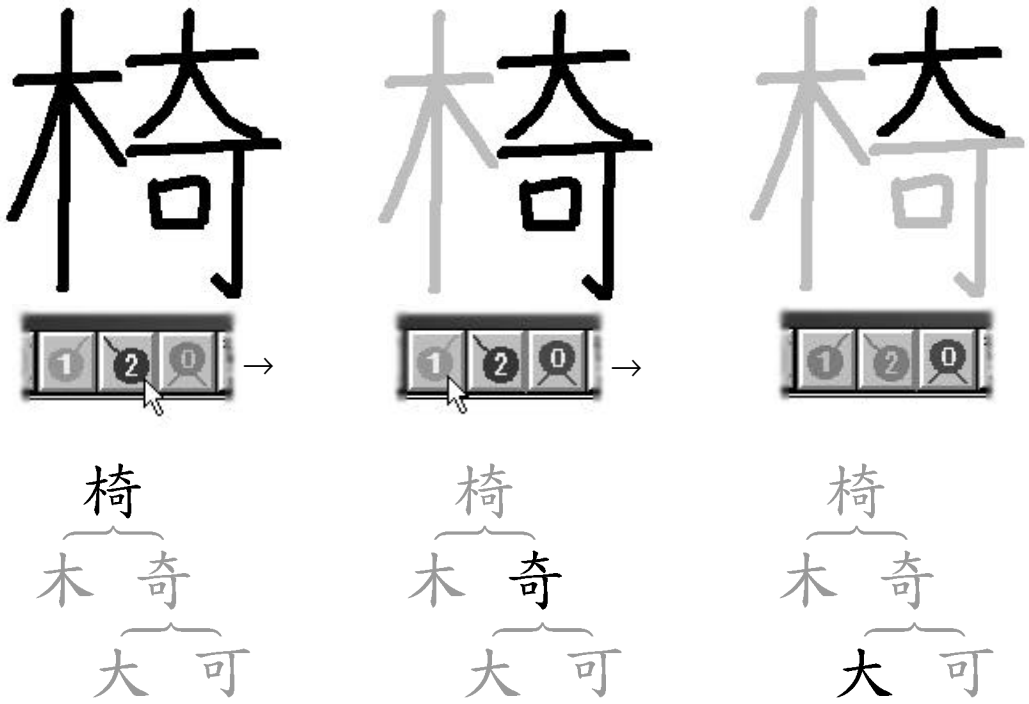


Figure 5. Traversing the components of the character in the early prototype

However, this interface is difficult to use because the represented character on the screen is notably different from the manner of controlling it [17]. For example, to go to the component **大**, the students have to pass through some kind of mental conversion to translate the desirable goal into a sequence of operations, namely, to the Left Child and then the Right Child. This is especially difficult for those students without a conceptual model of a tree in mind. As such, to unload the burden of the conversion, we attempt to build an interface in which the students can more directly manipulate [18] the structure of the character.

As illustrated below, in the Stroke Sequence exercise, a more direct and intuitive way of interacting with the character has been used for manipulating the structure at the stroke level. The students can click the strokes of the character right on the screen. When the correct starting end of the first stroke is clicked, the stroke will be written out immediately. The writing then proceeds stroke by stroke as 丨, 冂, 日 and 日 in the left component and then 明, 明, 明 and 明. This follows the commonly accepted writing practice of stroke sequence in Chinese, namely from left to right and from top to bottom. But if the wrong end or the wrong stroke is being selected, the stroke will still be drawn but then be immediately cleared, giving feedback to the student that he has successfully clicked a stroke but unfortunately on the wrong one. The animated drawing of the strokes will also reveal the direction of writing the strokes, thus the students would know how to improve themselves.

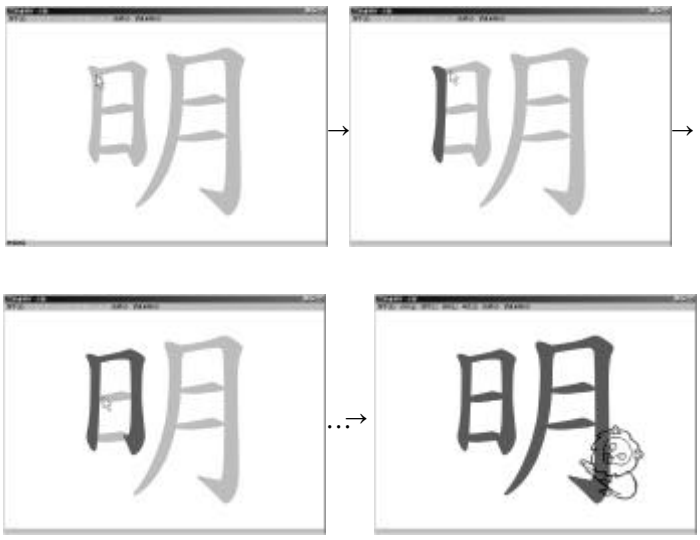


Figure 6. Writing a Chinese character with strokes in sequence

Sometimes the students may select a stroke in the wrong component such as first clicking a stroke in 月 instead of 日. In so doing, the component 月 on the screen will be dimmed, leaving only the component 日 remain visible. This draws the attention of the student back to the component that the students ought to finish with before proceeding. In this way the students are scaffolded to write the components one by one and in the right sequence.

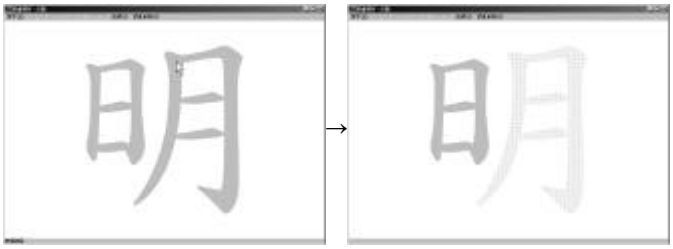


Figure 7. Writing a Chinese character with components in sequence

This sophisticated interaction however involves much difficulty in implementation. The difficulty lies in a lack of functions for processing the strokes of Chinese characters in most programming tools. Because of this, we have to build our own Font Engine, rastering each stroke of the characters primitively from a closed loop of continuous Bezier curves [19]. The Font Engine has not only made possible a more intuitive interaction but also improved the artistic looking of the characters over the monotonous strokes produced in the earlier prototypes. This reflects our decision to assign more weight to the sense of beauty of the characters over realizing such technical possibility as automatic generation of the glyph of the characters recursively out of the components [4].

3.2 The Principle of High Accuracy and Correctness

Teachers are often fully occupied with the busy schedule of teaching practice in schools, thereby

unlikely to verify every detail of the contents in the database themselves. The database, once published, will be in no doubt referred by the teachers as some kind of an authoritative "standard". From a technology-centric perspective, an error may simply be regarded as an erroneous record in the database, which can be easily deleted. However when we consider a student actually learning the characters from the database, this could become detrimental. Because of this, a high accuracy of contents is a must to us. As an example, we have carefully revised the form of each Chinese character following the correct standard as recommended in [20]. The character font commonly available in many operating systems, such as Mingli **細明體**, has sometimes been slightly modified to make the characters more eligible in print. But most Chinese educators will definitely not accept these modified characters for the purpose of teaching the characters.

Figure 8. Difference between the system fonts and the recommended standard

Though this slight difference may not be observable to most people, children will probably imitate this incorrect way of writing as the "correct" one. This seemingly unimportant matter should not be simply overlooked.

The accuracy of the contents will also be threatened by the tedious task of typing the data into the database. To minimize the chance of errors, we have not only conducted meticulous proof reading on our own, but has also convened a group of ten in-service primary school teachers to evaluate the database. Each of the teachers had to review a portion of the contents, and crosscheck with one another. Finally a report had to be turned in, commenting on the database in three perspectives: the accuracy of the contents, the design of the database and the application of the database in classroom. This is the quality assurance procedure we have imposed to ensure accuracy.

Despite all these efforts, certain inaccuracy is intricate and cannot be easily determined. For example, at first glance the shared four-dot component ⺣ of the characters **煎** and **燕** may look identical. But careful scrutiny will reveal the subtle difference. The four-dot component in the character **煎**(fry) is morphological, denoting the relationship of the character to the meaning of "fire"; while the

character **燕**(swallow) is a pictograph derived from the picture of a swallow with the four dots being the tail of the swallow. Hence the two components are actually different in terms of their functions and meanings, and it would be inappropriate to link the characters **煎** and **燕** together.

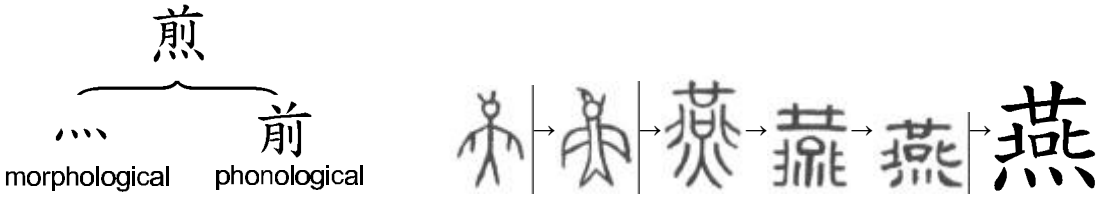


Figure 9. The difference between the four-dot components of the characters **煎** and **燕**

The reason for this complication is that some different orthographic components have been merged to the same writing in Kai **楷書** through historical evolution. For example, the seemingly the same component **月** of the characters **朗**(bright and clear), **臂**(arm) and **服**(submit) actually originates from the different components **月**(moon), **肉**(flesh) and **舟**(boat) respectively. The character **服** possesses the component **舟** because the submitted were sent by the boat at the ancient time. Differentiating the difference between these components will require close examination of the historical development of the characters. Because of this, the ongoing work is now being conducted in collaboration with the renowned scholars at the Beijing Normal University, hoping to leverage their expertise in the etymology of Chinese characters [21][22].

3.3 The Principle of Codifying the Components with Variations

There are now a number of standard schemes for the internal coding of the *whole characters*. However, not until recently [23][24], there is still no widely accepted coding scheme for the *components* inside the characters. For this reason, we have to codify the components in our own way. The codification process is found to be much more difficult than it apparently is.

The difficulty in codifying the components comes from the existence of many slight distortions in their form when they are fitted into different characters. The various forms of the components **人**(man), **心**(heart), **土**(land) and **食**(eat) are shown below. Notice that the base forms are sometimes notably different from the variant forms. For example, the character **食** has nine strokes in total on its own, but has only eight strokes as a component of the character **飯**(rice).

To appear in the International Journal of Computer Processing of Oriental Languages, Vol. 13, No. 4, pp. 351-375.

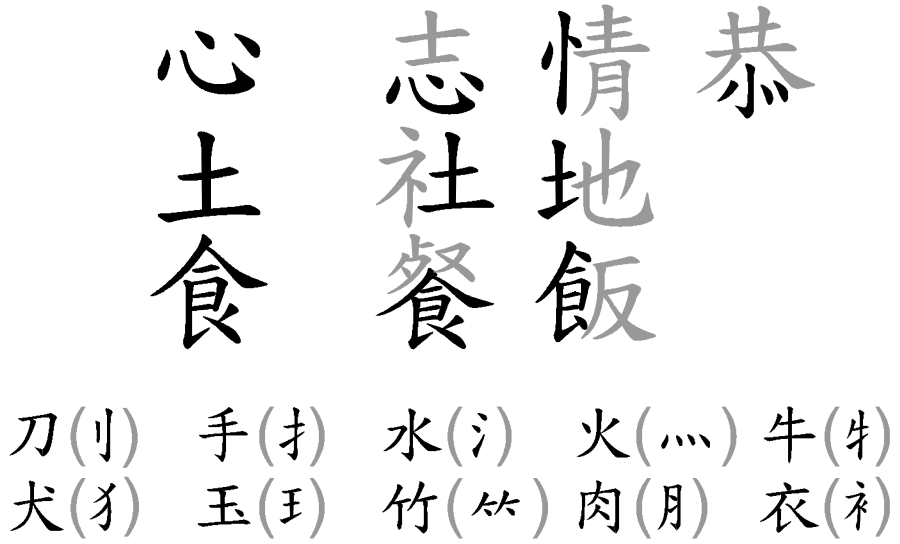
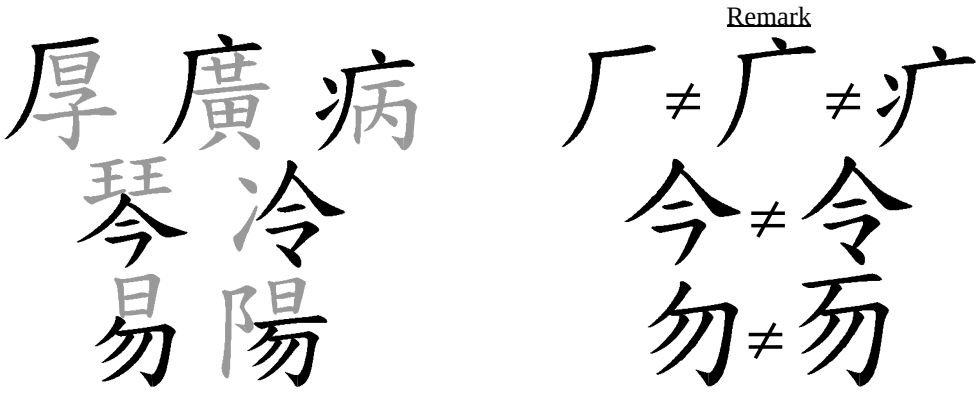


Figure 10. Variants of the same components

If these variant forms are treated as different based on these surface features, the characters will not be linked up together through the component, and many rich relationships among the characters would then be lost. This means that the students will not be able to relate the character **坐**(sit) via the component **人**(man) to the character **休**(rest). But the variants actually carry the same functional meaning in the different characters. All of the variants of the component **心**(heart) are morphological in the characters **志**(will), **情**(feeling) and **恭**(respect), thus their relationships are clearly definite. As such, to deal with this problem, we have to codify the variants of the components as well, but keep a certain order in which the database will return the list of related characters. For instance, via the component 忄 of the character **情**, related characters such as **快**(happy) and **怕**(fear) will be listed out first; followed by the characters **心** and **志** and others, and finally the characters **恭** and the rest.

Though variant forms of components are related, the readers should not be confused with other similarly looking components that are actually treated as completely different in the Chinese writing system. For example, the following components should be codified differently even though the components resemble each others except with only a few more or less strokes.



To appear in the International Journal of Computer Processing of Oriental Languages, Vol. 13, No. 4, pp. 351-375.

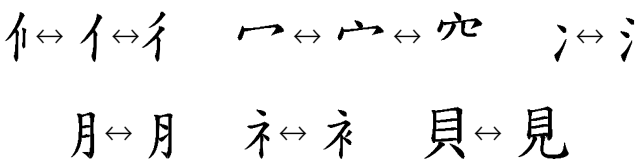


Figure 11. Components regarded as different despite the small variation in form

One way of handling these components is to further break down the components into a more simpler component along with a few detached strokes. Consider decomposing the component 疒 into the components 广 and 冫. By so doing, the total number of codifying components will be reduced, thus saving plenty of coding space. However we have decided not to take this approach because as for the learning purpose, the small difference between the components is considered as useful and important for the students to tell one component from another. Students should be taught to distinguish the two components as totally unrelated with connections to different meanings. For example, if the character **病**(ill) was inappropriately regarded as a combination of the three components 冫, 广 and **丙**, a link from the component 广 to the character would be superfluous even though the character apparently contained the component. It is because the component 广 is the picture of a "house" as illustrated in **店**(shop), **府**(government office) and **庫**(warehouse). It would be wrong for the students to connect the character **病** to the meaning of "house" rather than that of "sickness". As such, students ought not to mix up the components 疒 with 广.

3.4 The Principle of Decomposition of Characters

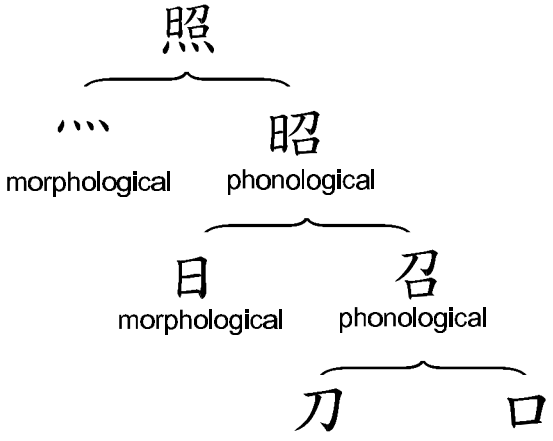
It is our belief that the decomposition of the characters into components can improve the students' effectiveness in learning the characters. But over-decomposing the characters may turn the characters into bits and pieces, ending up with a meaningless collection of strokes. Hence to what extent should the characters be decomposed should be examined with care. We found that it depends for what purpose to be achieved.

When the purpose is to develop a general coding scheme [25][26][27] to code the forms of Chinese characters for input or data storage, we should find a minimal set of basic components. Such components are usually "small" and reusable, which can be used in a uniform way to cover as many characters as possible.

In contrast to this, when the purpose is for learning the characters, the breakdown of the characters should be considered on a character by character basis, only to a level useful for interpreting the character. Simply separating a picto-phonetic composition into the morphological and phonological components is a typical example. Our approach is basically similar to this because our purpose is to identify the components of a character that can help the students to recognize and memorize the

character. As an example, the component 月(muscle) of the character 腳(foot) should be taken as a component since obviously the foot is made up of muscle. However the small piece of 口 inside the component 卻 will not be considered since no student is likely to recall the character 腳 from the meaningless piece of 口. On the contrary, in the context of learning the character 谷(ravine), the component 口 should be made salient because the upper component 八 denotes "water"; while the component 口 is the mouth of the spring, where the water comes out. Taking another example, in the character 湖, the decomposition will only give 氵(the meaning "water") and 胡(the sound "wu4"). However, 胡 is in itself a Chinese character. When taken as a character, 胡 is decomposed into the components 古 and 月. So in a sense, the decomposition is context dependent. Using this approach, the resulting set of components will be large in number, along with many "big" and "small" components across the whole database.

One might think the decomposition issue is primarily about how much and how deep we open up the decomposition to the learners; thus internally we can freely decompose the characters. This however is not always the case. Whether to use a hierarchical or a plain decomposition should also be carefully considered. For example, the picto-phonetic character 照(shine) is decomposed at the first level into the morphological component 火(fire) and the phonological component 昭. 昭, as a character itself, is then decomposed again into the morphological and phonological components 日 and 召 at the second level. Finally the character 召 is also decomposable into the components 刀 and 口 at the third level. Here the hierarchical organization is in line with the literature in etymological studies. However, if we perform the same hierarchical decomposition with the character 解(divide), it will be etymologically wrong. In contrast, the decomposition of the character 解 does not have such hierarchy. The meanings of all the components 角(horn), 刀(knife) and 牛(cow) come in at the same level to contribute to the overall meaning of the resultant character. The meaning of the character 解 originates from that of using a knife to separate the a cow's horn from the cow.



(a) Proper hierarchical decomposition

To appear in the International Journal of Computer Processing of Oriental Languages, Vol. 13, No. 4, pp. 351-375.

(b) Improper hierarchical decomposition (c) Proper plain decomposition

Figure 12. Hierarchical and plain decomposition of the characters **照** and **解** respectively

When using a component to search characters, the characters found should be retrieved in an order according to the level of the component inside the characters. For example, using the component **口** (mouth) should first find the list of characters **吃**(eat), **唱**(sing), **員**(member) and others with the component **口** at the first level; then the characters **圓**(round), **聖**(sacred), **路**(road) and others with **口** at the second level; lastly followed by the characters **落**(down), **湖**(lake), **照**(shine) and others. Notice that usually the more surface the component **口** appear, the more relevant to the meaning of "mouth" the characters are.

So far we discuss about breaking down a character into components for better understanding the meaning of the character. Learning to write the characters should however require a different approach in the breaking down of the characters. In this case, a character is decomposed into components on a flat plane so that the students can learn to write the components one by one. For example, the character **照** is written in the sequence of the components as shown below.

照 = 日 刀 口 灬

Figure 13. Decomposing the character **照** as a linear sequence of components

Sometimes, when writing a character for calligraphic reasons, an intact component may even need to be broken up. For example, the picto-phonetic character **圓** clearly consists of the morphological component 口 and the phonological component **員**. But the outer component 口 has to be separated into 冂 and 一 in writing. It is because the common writing of the character is to finish the inside before closing the last stroke at the bottom, namely, 丨 冂 冂 冂 圓 圓. Similarly, those characters with overlapping components also need to have some components decomposed. For example, the character **東**(east), as explained as the sun(**日**) behind the tree(**木**), will have the component **木** divided into two parts.

To appear in the International Journal of Computer Processing of Oriental Languages, Vol. 13, No. 4, pp. 351-375.

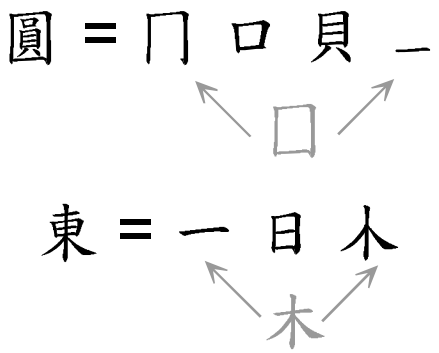


Figure 14. Breaking down an intact component for teaching the writing

3.5 The Principle of Handling Multiple Acceptable Values

Some characters can take on multiple acceptable values. This has caused much complication to the implementation of the database. For example, sometimes one written form can denote more than one characters (**一形多字**). For example, the form **只** is itself a character as in the word **只要**. But it is also the simplified form of the character **隻** as well. Some characters may also have more than one pronunciations (**一字多音**). For example, the character **行** can be pronounced as at least hang4, hong4 and hang6 in the words **步行**(walk), **行業**(business) and **品行**(conduct) respectively. Even a word composed of several characters can have more than one meanings or senses (**一詞多義**). As examples, the word **東西** can mean “east & west” and “something”; while the word **左右** can mean “left & right”, “around” and “obstruct”. These are the unavoidable cases which the database must be able to handle.

However, there are other cases of multiple acceptable values, which can be handled more easily. An obvious example is the existence of more than one way in ordering the strokes of a character. Though the stroke sequences of the characters are governed by many widely recognized writing rules, there are many cases where the rules actually compete with each other. As an example, different people will probably come up with a different stroke sequence for the character **火**. Some would prefer to simply write the character from the left to the right; while the others finish the sides before the center, the latter of which has been accepted in the official standard in mainland China [28].

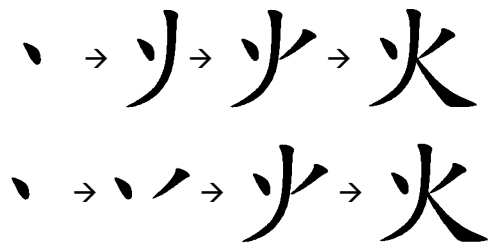


Figure 15. Multiple acceptable ways of writing the character **火**

To appear in the International Journal of Computer Processing of Oriental Languages, Vol. 13, No. 4, pp. 351-375.

How to select the "correct" sequence of writing the character will depend upon how one see the priority of the rules in a slightly different way [29]. However, supporting these variations in writing will give rise to much complexity in the implementation. As such, we have decided not to accept more than one alternative. After checking up the references, we have picked up one sequence as the "recommended" that we want the students to follow. After all, we believe this reduction of complication in programming is worthwhile.

Another case of accepting multiple values is in searching and relating the characters in the database. This has been dealt with differently. Searching for the characters should allow as many variants as possible to ease the students in locating the desirable characters. For example, any one of the sounds hang4, hong4 and hang6 should find the character **行**. This is actually the way many developers of the Chinese input methods have adopted to make the typing of the characters more convenient. In some cases even non-standard variants or the simplified forms of the characters are accepted. For example, entering the keystrokes for the varied forms 烟 and 双 can successfully find the characters **煙**(smoke) and **雙**(pair) respectively.

3.6 The Principle of Irregularity and Redundancy

The regularity existing within Chinese characters is fascinating. However, when the characters were being orderly put into a systematic and restrictive scheme, it was found that some characters are not that regular. The following indicates some of the relations in the database, which apparently exist among the characters but have later been completely modified to accommodate the irregularities.

From the outset, a character is often assumed to contain a number of components, which are in turn consisted of a number of strokes. This seemingly unquestionable relation is depicted as follows.

Figure 16. Apparently "correct" relations of the characters in the database

However, over the long period of history, the evolution of some of the characters has resulted in deviation from the rule. For example, to simplify the writing, the strokes of different components of a character may have been joined together, and some of the strokes may have been fused. As an example, the character **釜**, which has been formed by the morphological component **金** and the phonological component **父**, has now some of the strokes shared together by the two components.

To appear in the International Journal of Computer Processing of Oriental Languages, Vol. 13, No. 4, pp. 351-375.

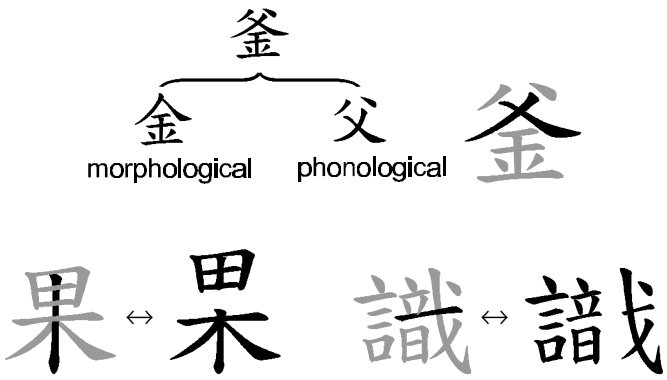


Figure 17. Irregular characters with merged strokes

These unusual cases obviously cannot conform to the relation claimed above since the shared strokes now belong to more than one components, violating the 1 to n relation between the component and the stroke. To adapt to this, we have actually stored the characters as many redundant representations rather than a generic set of logical rules. This approach obviously takes up more storage and running time, but is more malleable and flexible to meet the requirements of the real data. Actually many aspects of natural languages are implicit and even metaphorical [30], we actually feel that the database is only imitating and mimicking the reality of the language. The language clearly cannot be, and ought not to be, governed by the computer.

3.7 The Principle of Application

Since no student can learn the characters solely from looking up the dictionary, we have also provided a variety of Interactive Exercises along with the database for the students to work on. The database has actually been designed with the intention to provide teachers with ease to generate learning exercises that retrieve contents from the database. In a sense, we have followed a scenario-based design approach [31], developing the database with a set of applications in mind. Every time we picked up one learning exercise and then analyzed the functions and data needed to be supported by the database. In this way, we can prioritize and delimit the scope of the database development.

One of the applications alongside the database is the Morphological Component exercise. The purpose of the exercise is to raise the sensitivity of the students on what kind of morphological component a character is likely to possess. For example, the characters related to female will most likely have the morphological component **女**. As shown below, after hearing the sound of the character **妹**(young sister), the student has to think about and choose among the options of **人**(human) and **女**(female) a morphological component for the character.

To appear in the International Journal of Computer Processing of Oriental Languages, Vol. 13, No. 4, pp. 351-375.

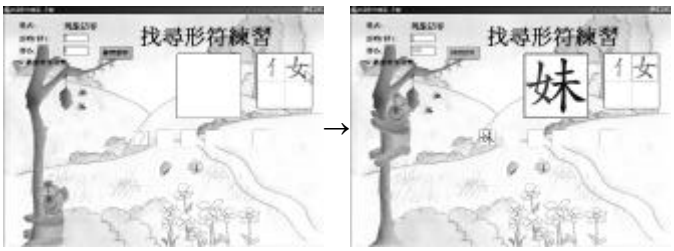


Figure 18. Identifying morphological components

Beyond exercises on computer, some of the exercises can also be delivered as ordinary paper worksheet. Switching between the worksheet and the database, the students can enjoy more variations in the materials and media of instruction, which can certainly maintain the interest and motivation of the students. This is also in line with the findings of our classroom implementation research [32], in which successful lessons have probably been conducted with various classroom activities alongside the computer. To support the above exercise on identifying morphological component, the database can be enhanced with the addition of a function to print out the characters as a stack of paper cards. Through the activity of sorting and categorizing the paper cards, the students can easily induce the relationship between the morphological component and the kind of characters that possess the components.

4. A Scheme of the Database

Figure 19 depicts the schematic structure of the database that we have used to store the data of the first five hundred characters specified in the curriculum. The essential elements of scheme will be briefly described below, with the intention that the readers can have a taste of the intricacy of the database.

Character ***字*** *and Form* ***字形***. The form of a character is purposely made different from the character itself because it is possible for a character to have more than one form. For example, the character **後** has at least the Traditional Form **繁體** of **後** and the Simplified Form **簡體** of **后**. Other kinds of varied forms include the many debatable variant forms **異體** of the character, such as 峯 and 峰 as well as 啟 and 啓. Many properties commonly associated with the character itself should instead be attributed to the form. It is because these properties, like the radical **部首**, may differ in the different forms. For example, the radicals of the character **後** are **彳** and **口** respectively for the two forms **後** and **后**.

Liu Shu ***六書*** *and Origin* ***由來***. Liu Shu is the one of the six ways that the creation of the character forms has been classified. When being Picto-Phonetic Composition **形聲字**, the liu shu of a form is stored together with the indications of the morphological **形符** and phonological components **聲符**, such as the components **耳** and **呈** respectively for the form **聖**. In the cases of Imitative Drafts **象形**, Logical Aggregates **會意** or Indicative Symbols **指事**, the creation of the forms can be explained

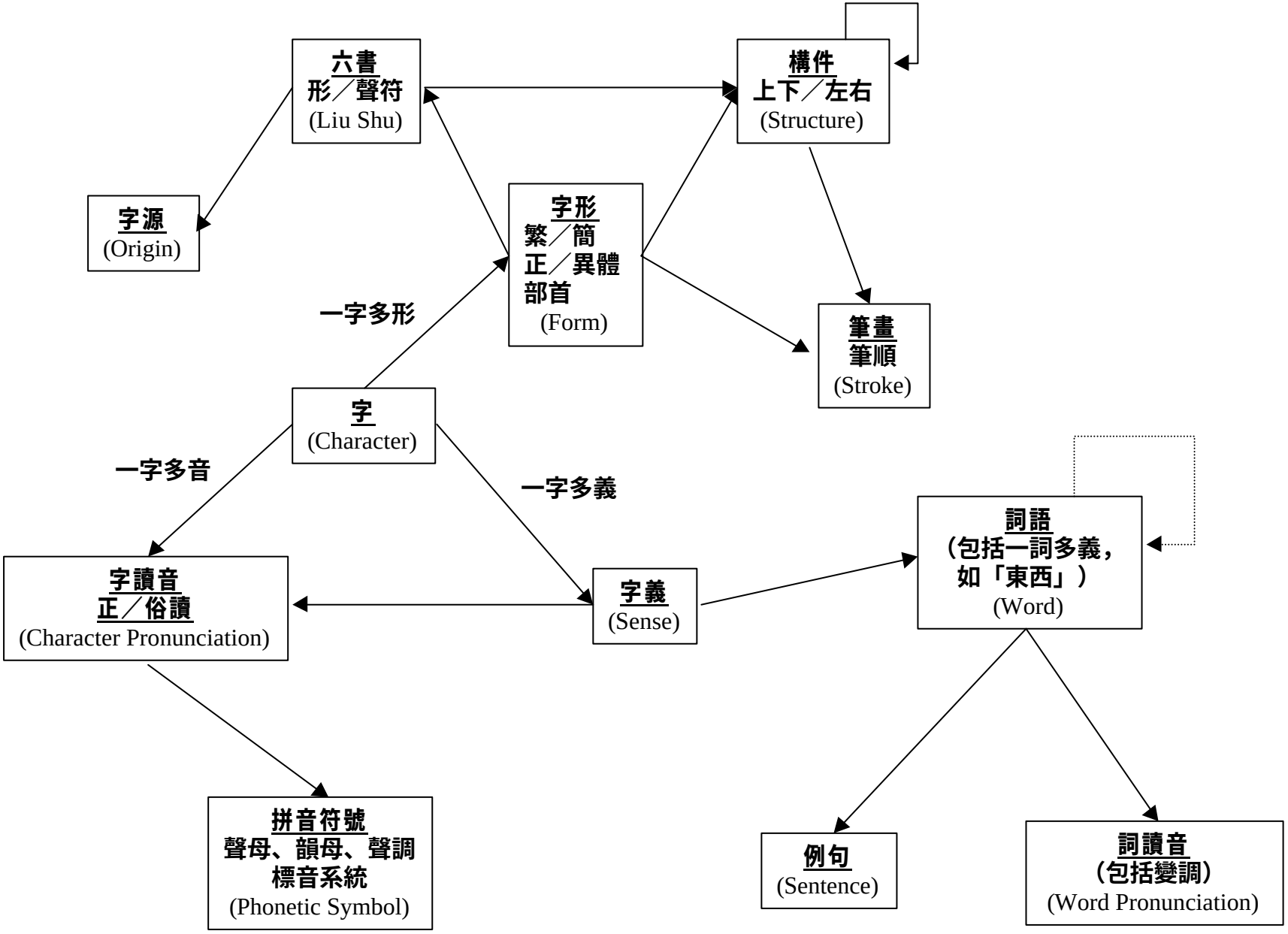


Figure 19. Database Structure for representing the knowledge of Chinese Characters

by their pictorial origins, thus animation about the evolution of the forms will be provided.

Structure ***構件***. The breaking down of the form of a character is represented recursively in Structure. For example, the form **唱** is broken down into the components **口** and **昌**, which is in turn separated into the components **日** and **曰**. Multiple decompositions of a form can also be stored for the different purposes. For example, to depict the radical **豕** of the form **象**, the **豕** is taken as one component; while to teach students to recognize the character **象**, the whole form is kept intact because the form **象** is actually an inseparable pictograph. In addition to this, a sequence number has been introduced into the naming of the components inside a form. It is because the same component may appear several time in a form such as the three components **木**' s of the form **森**. To differentiate the components, a different sequence number has to be assigned to each of them. This conforms to the sequence of writing the first stroke of the components in the common writing practice of the form.

Phonetic Symbol ***拼音符號****, Character Pronunciation* ***字讀音*** *and Sense* ***字義***. Different standards of phonetic alphabets can be used to represent the same sound in the database. We adopt the system advocated by the Linguistic Society of Hong Kong **香港語言學學會** for romanizing the Cantonese sounds of the characters. A character can have more than one pronunciation. For example, three pronunciations are possible for the character **會**, depending upon the referred meanings and senses of the character. When meaning "association", "capable of" and "gathering", the character **會** is pronounced as wui2, wui5 and wui6 respectively, as in **社會**(se2wui2, society), **不會**(bat1wui5, incapable of) and **再會**(zoi3wui6, see you again).

Word ***詞語****, Word Pronunciation and Sentence* ***例句***. Over 4,500 words, which are formed by one or more of the 500 characters, are stored in the database. The meaning and usage of a word are illustrated with two sample sentences using the word. The words are primarily pronounced using a character-by-character approach; but the sound of the character inside the word can also be overridden to accommodate the inflection of the sound at the word level. For example, both meaning "younger sister", the character **妹** is pronounced differently in the words **兄妹**(hing1mui2, elder brother and sister) and **弟妹**(dai6mui6, younger brother and sister). In this case, the tone of the character has been changed. In addition, some words may have the same appearance but of different meanings, such as **學會**(hok6wui2, association) and **學會**(hok6wui5, learn to). Again a sequence number has been used to differentiate the words.

5. Evaluation

The Dragonwise Series software has been introduced to more than 600 teachers in various public lectures. The feedback from the teachers, as indicated in the questionnaires administrated during the lectures, is generally positive. Many teachers think the software will be helpful to them and are also

willing to try out the software in the classroom. Because of this, we have actually conducted a qualitative study on the classroom implementation of the software. The study has been carried out in three primary schools in Hong Kong. During the study, the teachers in these schools have used the software to teach primary one and two students for a number of lessons. Twelve of the lessons have been observed and videotaped for subsequent analysis. The teachers and the students have also been formally and informally interviewed after the lessons. Simply put, the results indicate that the implementation of the software in classroom is feasible, and the use of the software as consolidation activity is achievable. After the lessons, the students have become aware of the regularity of the Chinese characters. This can be clearly seen in one of the lessons, in which the students could successfully identify and distinguish the characters with adjacent pronunciations, namely, the **睛** of **眼睛**(eye), the **晴** of **天氣晴**(sunny sky), the **青** of **青蛙**(frog), the **蜻** of **蜻蜓**(dragonfly), the **清** of **清新**(clear) and the **請** of **請你**(invite you). Further to these observational data are discussed in detail in [16][32] together with the issues on pedagogy.

6. Summary

Chinese characters bear many relations among one another. When students are empowered to proactively explore these relations in a supportive database, the learning of Chinese characters can be greatly enhanced. This paper reports our experience in designing such a database for children to learn Chinese characters. The basic functions of the database are to search, to expound and to relate the characters, as described at the beginning of the paper. Then several issues arisen from the development of the database have been discussed. We have made the following recommendations: (1) The students should be able to interact with the characters in an intuitive way. (2) The quality and accuracy in the contents of the database should be assured. (3) The component to be codified should be distinguished from the variants. (4) The characters should be decomposed in such a way to aid the students to remember the characters. (5) During the design of the database, the multiplicity of the acceptable values of the characters should be taken into consideration. (6) The database should be malleable enough to accommodate the irregularity of Chinese characters. (7) The database should be developed in conjunction with conceived teaching and learning activities. Many findings of this paper have resulted from a decade of experience in implementing the databases in actual contents; while we are working closely together with many Chinese language teacher educators. As such, we hope this research could throw light to a more humanistic and contextual approach towards Chinese computing.

Acknowledgements

The research reported here was supported by the Language Fund sponsorship C/023/95A and the Research Grants Council grants HKU33/91 and HKU168/95H. Our thanks go to those who worked alongside us: S.K. Tse made invaluable comments and suggestions on the pedagogic methods for

teaching Chinese characters. Vincent Lau gave us advice on resolving technical problems related to Chinese computing. Adela Lau, May Mak and Eric Tam implemented most of the software programs. Isis Lee painstakingly prepared the contents in the database. Sunny Wong, Kelvin Yu and Simon Lam drew, touched up and animated the graphics. Angela Chow produced the music. Rex Ng edited the recorded sound. We would also like to acknowledge all the students, teachers and principals who have helped us in many ways.

References

[1] ***粵語拼音字表***，香港語言學學會，香港，1997 年。

[2] P.S.P. Wang, "Knowledge Pattern Representation of Chinese Characters," *Computer Processing of Chinese and Oriental Languages*, Vol. 3, Nos. 3 & 4, 1988, pp. 331-349.

[3] K.S. Leung, Y. Fan and F.Y. Young, "A Chinese Dictionary System based on Fuzzy Logic and Object-oriented Approach," *Computer Processing of Chinese and Oriental Languages*, Vol. 6, No. 2, 1992, pp. 205-219.

[4] P.K. Lai, D.Y. Yeung and M.C. Pong, "A Heuristic Search Approach to Chinese Glyph Generation Using Hierarchical Character Composition," *Computer Processing of Oriental Languages*, Vol. 10, No. 3, 1997, pp. 281-297.

[5] Donald A. Norman, *the Invisible Computer: Why Good Products Can Fail, the Personal Computer Is So Complex, and Information Appliances Are the Solution,* The MIT Press, Cambridge, Massachusetts, 1999.

[6] H.C. Lam, K.H. Pun, S.T. Leung, S.K. Tse and W.W. Ki, "Computer-Assisted-Learning for Learning Chinese Characters," *Communications of COLIPS*, Vol. 3, No. 1, 1993, pp. 31-44.

[7] W. Ki, S.K. Tse, N. Law, F. Lau, K.H. Pun, "A Knowledge-based Multimedia System to Support the Teaching and Learning of Chinese Characters," in *Proceedings of ED-MEDIA 94 – World Conference on Educational Multimedia and Hypermedia*, Vancouver, B.C., Canada, June 25-30, 1994, pp. 323-328.

[8] 謝錫金、祁永華、羅陸慧英、鍾嶺崇，"電腦輔助學習漢字和閱讀理解，"***中文教育論文集第二輯（上）***，1994 年，171-186 頁。

[9] H.C. Lam and S.T. Leung, "Han: Computer-Assisted-Learning for Learning Chinese Characters," Final Year Project Report, supervised by K.H. Pun, Department of Computer Science, The University of Hong Kong, 1991.

[10] Y.T. Lee and W.H. Chau, "A Database for Chinese Characters," Summer Training Report, supervised by W.W. Ki, Department of Electrical and Electronic Engineering, The University of Hong Kong, 1993.

[11] C.W.M. Wong and B.C.C. Tsang, "Multimedia System for Constructing a Chinese Character Database," Final Year Project Report, supervised by K.H. Pun, Department of Computer Science,

The University of Hong Kong, 1995.

[12] ***現龍系列《漢字學習系統》之《現龍漢字資料庫及互動練習》*，香港大學課程學系，香港，**1999 **年。**

[13] ***現龍系列《漢字學習系統》之《童歌字趣》*，香港大學課程學系，香港，**1998 **年。**

[14] ***現龍系列《漢字學習系統》之《小兔子找食物》*，香港大學課程學系，香港，**1998 **年。**

[15] ***現龍系列《漢字學習系統》之《文字國》*，香港大學課程學系，香港，**2000 **年。**

[16] H.C. Lam, W.W. Ki, N. Law, A.L.S. Chung, P.Y. Ko, A.H.S. Ho and S.W. Pun, "Designing CALL for learning Chinese characters," to appear in the Journal of Computer Assisted Learning, Vol. 17, No. 1, 2001.

[17] D.A. Norman, "Design Principles for Cognitive Artifacts," *Research in Engineering Design*, Vol. 4, 1992, pp. 43-50.

[18] B. Shneiderman, "Direct manipulation: A Step Beyond Programming Languages," *Computer*, IEEE Computer Society, Vol. 16, No. 8, 1983, pp. 57-69.

[19] *The TrueType Book*, Apple Computer, Inc., Addison-Wesley Publishing Company, Inc., 1991.

[20] **李學銘主編，*常用字字形表*，香港教育學院，香港，**1997 **年。**

[21] **王寧、鄒曉麗主編，*漢字*，和平圖書，海峰出版社，香港，**1999 **年。**

[22] **王寧主編，*漢字漢語基礎*，科學出版社，北京，**1997 **年。**

[23] ***香港增補字符集*，香港特別行政區政府，資訊科技署及法定語文事務署，香港，**1999 **年。**

[24] **國家語言文字工作委員會，"信息處理用** GB 13000.1 **字符集漢字部件規範，"語文出版社，北京，**1997 **年。**

[25] C.Y. Suen and E.M. Huang, "Computational Analysis of the Structural Compositions of Frequently Used Chinese Characters," *Computer Processing of Chinese and Oriental Languages*, Vol. 1, No. 3, 1984, pp. 163-176.

[26] **陳愛文、周靜梓 、陳尚農，*漢字字形學和表形符號編碼*，光明日報出版社，北京，**1987**。**

[27] **盧紹昌，"漢字部件的研究，"*第三屆國際漢語教學討論會論文選*，第三屆國際漢語教學討論會會務工作委員會編，北京語言學院出版，北京，**1990 **年** 8 **月，**587-597 **頁。**

[28] **國家語言文字工作委員會，*現代漢語通用字筆順規範*，語文出版社，北京，**1997 **年。**

[29] N. Law, W.W. Ki, A.L.S. Chung, P.Y. Ko and H.C. Lam, "Children's stroke sequence errors in writing Chinese characters," in *Reading and Writing: An Interdisciplinary Journal*, Vol. 10, pp. 267-292, C.K. Leong & K. Tamaoka (eds), *Cognitive Processing of the Chinese and the Japanese Languages*, 1998, pp. 113-138.

[30] C.A. Bowers, *The Cultural Dimensions of Educational Computing: Understanding the Non-Neutrality of Technology*, Teachers College Press, New York, 1988.

[31] John M. Carroll and Mary Beth Rosson, "Getting Around the Task-Artifact Cycle: How to Make Claims and Design by Scenario", *ACM Transactions on Information Systems*, Vol. 10, No. 2, April 1992, pp. 181-212.

[32] **祁永華、鍾嶺崇、高寶玉、林浩昌、周燕、巢偉儀，*多媒體電腦軟件的課堂應用研究：中文識字教學*，香港大學課程學系，香港，**1999 **年。**

To appear in the International Journal of Computer Processing of Oriental Languages, Vol. 13, No. 4, pp. 351-375.