Creating a smart database of Chinese character features

**Keywords:** Chinese characters; database;

**Introduction**

With the advent of information age, Chinese characters, which have served as a carrier to transmit information for thousands of years, are now treated as information themselves. Computers and cell phones have also become extremely popular, enabling access to online information. Under such background, a computerized database of Chinese characters can be very helpful.

Han (1993) introduced a database of Chinese character feature information and demonstrated its application in designing stimuli for psycholinguistic experiments. Jin and Chen (2006) created a database of Chinese character features, including 21 features. However, the databases are not accessible. Popular accessible Chinese character databases such as *handian* and *zhongwen* are dictionary-based. After entering a character, the database will return the metainformation of the character such as pronunciation, radicals, and meanings. Those databases have two disadvantages. First, although they are good at providing some metainformation of a single character, the metainformation is insufficient (usually under ten features). Second, they usually cannot generate a list of characters with similar features. Research and teaching are not limited to a single character, the association between characters is also important. Therefore, dictionary-based character databases need improvements in presenting the connection between characters.

**Literature review**

Recently, researchers have conducted a series of studies on Chinese character naming and publicized their data in digital form. Liu et al. (2007) investigated the correlation between naming latencies and 15 features of simplified Chinese characters and found that frequency, semantics, visual features, and consistency of Chinese characters are the major factors in character naming speed. Sze et al. (2014) conducted a similar naming experiment on simplified Chinese characters and released the naming latencies. Chang (2016) conducted another naming experiment on traditional Chinese characters using 12 features and found frequency, consistency, regularity, familiarity, semantic ambiguity, and the interaction between frequency and consistency affect naming latencies. All these character naming studies not only included many features of a character, but also revealed correlation between features of Chinese characters and naming latencies. Most importantly, the data were made publicly accessible. However, the databases are still focusing on single character.

Lo and Hue (2008) developed the Character-Component Analysis Toolkit (C-CAT) software. It can display the features of a character as well as search for a list of characters with user-defined features. The software enables researchers to control their stimuli when designing experiments. The powerful searching function of C-CAT is nevertheless limited to research purpose. A good database of Chinese character should also cater the needs of teachers and students.
Current study
The goal of the current proposal is to create a smart database of Chinese characters. After setting the parameters, the database can automatically select a list of Chinese characters. The database is also multi-purpose, meeting the needs of researchers, teachers, and learners. Researchers can utilize the database to create stimuli for psycholinguistic experiments, teachers can find out the correlation between students’ performance and character features, students can use the database as a vocabulary builder and an assessment tool.

The database will include a sample of 2000 simplified Chinese characters from Liu et al. (2007). Features of the characters include frequency, phonological frequency, homophone density, regularity, initial, final, tone, number of strokes, number of components, number of meanings, and concreteness. The features are categorized into four groups: frequency (frequency, phonological frequency, homophone density), pronunciation (regularity, initial, final, tone), visual form (number of strokes, number of components), and meaning (number of meanings, concreteness). Following Lo and Hue (2008), a matching algorithm is implemented to match the subcategories in frequency, pronunciation, visual form, and meaning according to the user-define requirements.

The interface will allow three types of operations: search, analysis, and generation. In the search operation, users can search for features of specific characters or search for a list of characters with shared features. The analysis operation allows users to see the correlation between character features and other variables that the users are interested in. The generation operation can generate a list of characters for learning and assessment.

Potential applications
Researchers scenarios
Deng et al. (2003) conducted an experiment on third tone sandhi and they wanted to know if lexicality can affect the application of third tone sandhi rule. They had a real word yushui (雨水) with third tone sandhi and they needed a non-word 雨 X. The character X must have the same tone as shui (水) and its frequency and number of strokes should be as close to shui (水) as possible. With the help of the database, we can easily find out suitable candidates for character X.

Language teachers’ scenarios
Chinese characters have been considered as one of the most difficult parts in learning Chinese. Students inevitably make errors in writing Chinese characters. Xiao (2002) summarized common errors in writing Chinese characters: radical replacement, radical addition, radical deletion, and radical transformation. Teachers can enter the characters that cause high error rate in the database and analyze the correlation between errors in writing characters and character features.

Language learners’ scenarios
The database can help learners increase their exposure to characters in a relevant way. The characters they see will not be random, but related to their knowledge and study objectives. For example, students can use the database to create a pronunciation assessment. The database can
create a list of words with same pronunciation, with same syllable but different tones, with same initial, or with same final from the vocabulary list in a textbook. Students can also be exposed to a list of characters with same radicals to enhance their orthographic awareness.

**Conclusion**

The current proposal aims to design a multi-purpose database of Chinese character features. The database can create a list of characters according to the users’ needs. It can help researchers, teachers, and students better work with Chinese characters.

**Reference**


