Visually and Phonologically Similar Characters in Incorrect Chinese Words: Analyses, Identification, and Applications

C.-L. LIU, M.-H. LAI, K.-W. TIEN, and Y.-H. CHUANG, National Chengchi University
S.-H. WU, Chaoyang University of Technology
C.-Y. LEE, Academia Sinica

Information about students' mistakes opens a window to an understanding of their learning processes, and helps us design effective course work to help students avoid replication of the same errors. Learning from mistakes is important not just in human learning activities; it is also a crucial ingredient in techniques for the developments of student models. In this article, we report findings of our study on 4,100 erroneous Chinese words. Seventy-six percent of these errors were related to the phonological similarity between the correct and the incorrect characters, 46% were due to visual similarity, and 29% involved both factors. We propose a computing algorithm that aims at replication of incorrect Chinese words. The algorithm extends the principles of decomposing Chinese characters with the Cangjie codes to judge the visual similarity between Chinese characters. The algorithm also employs empirical rules to determine the degree of similarity between Chinese phonemes. To show its effectiveness, we ran the algorithm to select and rank a list of about 100 candidate characters, from more than 5,100 characters, for the incorrectly written character in each of the 4,100 errors. We inspected whether the incorrect character was indeed included in the candidate list and analyzed whether the incorrect character was ranked at the top of the candidate list. Experimental results show that our algorithm captured 97% of incorrect characters for the 4,100 errors, when the average length of the candidate lists was 104. Further analyses showed that the incorrect characters ranked among the top 10 candidates in 89% of the phonologically similar errors and in 80% of the visually similar errors.

Categories and Subject Descriptors: I.2.7 [**Computing Methodologies**]: Artificial Intelligence—*Natural language processing*; J.5 [**Computer Applications**]: Arts and Humanities—*Linguistics*; K.3.1 [**Computing Milieux**]: Computers and Education—*Computer uses in education; Computer-assisted instruction (CAI)*; H.3.5 [**Information Systems**]: Information Storage and Retrieval—*Online information services; Web-based services*; J.4 [**Computer Applications**]: Social and Behavioral Sciences—*Psychology*

General Terms: Design, Languages

Additional Key Words and Phrases: Error analysis of written Chinese text, student modeling, traditional Chinese, simplified Chinese, computer-assisted language learning, psycholinguistics

ACM Reference Format:
Liu, C.-L., Lai, M.-H., Tien, K.-W., Chuang, Y.-H., Wu, S.-H., and Lee, C.-Y. 2010. Visually and phonologically similar characters in incorrect Chinese words: Analyses, identification, and applications. ACM Trans. Asian Lang. Inform. Process. 10, 2, Article 10 (June 2011), 39 pages.
DOI = 10.1145/1967293.1967297 http://doi.acm.org/10.1145/1967293.1967297

This article was completed while C.-L. Liu visited the Department of Electrical Engineering and Computer Science of the University of Michigan as a visiting scholar.
This research was supported in part by the research contracts 97-2221-E-004-007, 99-2221-E-004-007, and 99-2918-I-004-008 from the National Science Council of Taiwan.
Authors' addresses: C.-L. Liu, M.-H. Lai, K.-W. Tien, and Y.-H. Chuang, Department of Computer Science, College of Science, National Chengchi University, Taipei, Taiwan; email: {chaolin, g9523, g9627, g9804}@cs.nccu.edu.tw; S.-H. Wu, Department of Computer Science and Information Engineering, College of Informatics, Chaoyang University of Technology, Taichung, Taiwan; email: shwu@cyut.edu.tw; C.-Y. Lee, Institute of Linguistics, Academia Sinica, Taipei, Taiwan; email: chiaying@gate.sinica.edu.tw.

DOI 10.1145/1967293.1967297 http://doi.acm.org/10.1145/1967293.1967297

1. INTRODUCTION[1]

The studies about people using incorrect characters in Chinese words are related to the education, perception, recognition, and applications of the Chinese language[2]. Some Chinese words contain just one character, but most words comprise two or more characters. For instance, "好" (hao3)[3] is a word that has just one character and means "good" in English. "語言" (yu3 yan2) is a word that is formed by two characters and means "language" in English. Experience indicates that the two most common causes for writing or typing incorrect Chinese words are due to phonological and visual similarity between the correct and the incorrect characters [Liu et al. 2009a, 2009b, 2009c]. For instance, one might use "素" (su4) in the place of "肅" (su4) in "嚴肅" (yan2 su4) because of the phonological similarity; one might use "施" (shi1) for "旅" (lu3) in "旅途" (lu3 tu2) due to the visual similarity.

Manipulating the similarity between characters has served as an instrumental technique in psycholinguistic studies into how people read and recognize Chinese characters. Researchers in psycholinguistics investigate the cognition processes of Chinese readers [Kuo et al. 2004; Lee et al. 2006; Tsai et al. 2006], by measuring readers' response times to words that have various numbers of "neighbor" words. The neighbors of a Chinese word include phonologically and visually similar characters.

Phonologically and visually similar characters are also useful for computer assisted language learning (CALL). In elementary schools in Taiwan, students may be requested to identify and correct "erroneous words" in test items, where, typically, an "erroneous word" contains an incorrect character that was introduced intentionally when teachers prepared the test items. Such tests are Incorrect Character Correction tests (ICC tests). It takes effort and time to provide incorrect characters that are appropriate for different assessment purposes, and to make sure that the test items do not repeatedly use the same incorrect characters at the same time. We have built an environment for assisting the preparation of such test items [Liu et al. 2009a] by finding a way to offer phonologically and visually similar Chinese characters as candidates to serve as the incorrect characters [Liu and Lin 2008].

In addition, phonologically and visually similar characters can be applied to student modeling, optical character recognition (OCR), and information retrieval (IR) in Chinese. Bug libraries contain students' records of previous errors [Sison and Shimura 1998; Virvou et al. 2000], and are useful for modeling student behavior. Some algorithms for optical character recognition for printed Chinese and for written Chinese try to guess the input images based on confusion sets [Fan et al. 1995; Liu et al. 2004]. Characters in a confusion set are similar to each other visually, and they help the OCR programs to confine the search space for a given image. It would be possible to reduce the computational costs and to increase recognition rates if we can pinpoint the confusion set of a character that is being recognized. The current confusion sets are hand-crafted clusters of visually similar characters. In recent years, it has become a

[1]This article is a significantly extended version, in terms of the depth of discussion and the scale of experiments, of the material reported in Liu and Lin [2008], Liu et al. [2009a, 2009b, 2009c], and Liu et al. [2010].

[2]In this article, we use "Chinese" to refer to Mandarin Chinese.

[3]We show traditional and simplified Chinese characters followed by their Hanyu pinyin (http://en.wikipedia.org/wiki/Pinyin). The Hanyu pinyin of a Chinese character shows the sound of the character by a string of English letters, and the digit that follows the letters is the tone for the character. To simplify our presentation, we will show Chinese text only in either the traditional or the simplified form, but not both. If presented in simplified Chinese, the errors listed in the first paragraph in the Introduction will replace "肅" (su4) in "严 肃" (yan2 su4) by "素" (su4) for phonological similarity and "旅" (lu3) in "旅途" (lu3 tu2) by "施" (shi1) for visual similarity. The traditional and simplified forms of a Chinese character might not differ from each other.

common practice for IR service providers, such as Yahoo! and Google, to offer corrections when users enter queries that contain incorrect words. For English queries, one may apply the Levenshtein distance to compute the edit distance between the spellings and employ the Soundex system to determine the degree of similarity between the pronunciations of words [cf., Croft et al. 2010; Manning et al. 2008]. These methods are not perfect but can catch similar English words in practice. The work reported in this article can be applied to find possible corrections for Chinese queries.

Some researchers state that there are more than 50,000 Chinese characters [HanDict 2010], although only thousands of characters are used in daily lives. In the People's Republic of China, a government agency selected 7,000 popular Chinese characters and highlighted 3,500 characters among these 7,000 characters as the most frequently used characters in 1988[4]. In Taiwan, 5,401 characters were selected to be the most commonly used in daily lives in 1984 when the BIG5 code was formulated [Dict 2010].

Given that Chinese is used in different areas and in different countries in the world, it should not be surprising that not all people speak the "standard" Mandarin. We will focus on the standards that are stated in specific lexicons in this research. Given a specific lexicon, it is relatively easy to judge whether two characters have the same or similar pronunciations based on their records, when we do not consider the phenomena of co-articulation. Although there are thousands of Chinese characters, these characters are pronounced in only 420 different ways (cf., Lee [2009]). Interestingly, there are many fewer Chinese words that are pronounced exactly the same way than the number of Chinese characters that are pronounced exactly the same way. The problem of determining the pronunciation of Chinese characters becomes more complex if we consider tone sandhi [Chen 2000] in Chinese and if we consider the influences of sublanguages in the Chinese language family. We will discuss related issues in Section 2.

In contrast, there were no obvious ways to determine algorithmically whether two Chinese characters are visually similar yet. For instance, "員" (yuan2), "圓" (yuan2), and "勛" (xun1) are similar to each other in some ways, due to the presence of "員" (yuan2). Image processing techniques may be useful but are not perfectly practical, given the size of Chinese characters. A more important factor that affects the applicability of image processing methods is that many of the Chinese characters are similar to each other in subtle ways. "員" (yuan2) is contained in "員" (yuan2), "圓" (yuan2), and "勛" (xun1) in different sizes and at different positions.

We apply an extended version of the Cangjie codes [Cangjie 2010; Chu et al. 2010] to encode the layouts and details of traditional Chinese characters for computing visually similar characters [Liu and Lin 2008; Liu et al. 2009a, 2009b, 2009c], and extend the work to compare similar characters in simplified Chinese characters [Liu 2010]. Evidence observed in psycholinguistic studies [Feldman and Siok 1999; Lee et al. 2006; Yeh and Li 2002] offers a cognition-based support for the design of our approach; namely, the use of shared components to define the visual similarity between Chinese characters.

The proposed method proves to be effective in capturing incorrect words for both traditional [Liu et al. 2009a, 2009b, 2009c] and simplified Chinese [Liu 2010]. We col-

[4]The statistics are available on the following two Wikipedia pages: http://zh.wikipedia.org/zh-tw/%E7%8E%B0%E4%BB%A3%E6%B1%89%E8%AF%AD%E9%80%9A%E7%94%A8%E5%AD%97%E8%A1%A8 (if Chinese is available on your computers: http://zh.wikipedia.org/zh/现代汉语通用字表) and http://en.wikipedia.org/wiki/Xi%C3%A0nd%C3%A0i_H%C3%A0ny%C7%94_Ch%C3%A1ngy%C3%B2ng_Z%C3%ACbi%C7%8Eo (if Chinese is available on your computers: http://zh.wikipedia.org/zh/现代汉语常用字表). The first page is written in Chinese, and the second one is in English. The translations of "现代汉语" (xian4 dai4 han4 yu3) and "字表" (zi4 biao3) are "Modern Chinese" and "character list", respectively. We use "popular" for "通用" (tong1 yong4) and "most frequently used" for "常用" (chang2 yong4).

lected and analyzed approximately 4,100 errors that were reported in published books, found in students' compositions, or posted on the Internet. Each reported error is of a word which will be understood as appearing in its correct form as "嚴肅"; but which in the error may appear as "嚴素", where "素" is used instead of "肅". Namely, writing "嚴肅" as "嚴素" is a reported error. We found that 76% of the errors were related to phonological similarity and that 46% of the errors were related to visual similarity. More significantly, the dominance of the phonological factor was also observed in hand-written text, not just in electronic documents that were directly prepared on computers.

In experiments that aimed at reproducing the collected errors, we ran our programs to select and recommend a list of candidates from more than 5,100 Chinese characters for the correct character, that is, "肅", and we recorded the likelihood that the candidate list actually included the incorrect character. Experimental results show that if the length of the candidate list is about 100, we achieved inclusion rates of about 97% for both traditional and simplified Chinese. If the length of the candidate list was shortened to 10, the average inclusion rates were 89% for the phonologically similar errors and 80% for the visually similar errors. We have also applied our algorithms for reproducing the reported errors to build an environment to assist teachers to prepare test items for ICC tests.

In this article, we integrate and extend the previous reports on the phonologically and visually similar characters in both traditional and simplified Chinese to capture errors in Chinese words. We go over some issues about phonological similarity in Chinese in Section 2, elaborate how we extend and apply the Cangjie codes to judge the visual similarity between Chinese characters in Section 3, explain how we acquired the reported errors and how we analyzed the phonological and visual influences on these errors in Section 4, present details about our experiments and discuss the observations in Section 5, show a real-world application of the proposed techniques to the authoring of test items for the ICC tests in Section 6, and review some of the design issues and experience in Section 7 before we summarize our work in Section 8.

Compared with the previous conference articles [Liu and Lin 2008; Liu et al. 2009a, 2009b, 2009c; Liu et al. 2010], we expanded the scale of experiments and discussions in terms of both depth and coverage. More specifically, we validated the reliability of the Web-based statistics by examining the data that we collected in 2009 and in 2010, compared the contribution of different sources of similar characters, explored the applications of alternative ranking methods, and exhibited the robustness of our approach by running our systems over new data sets.

2. PHONOLOGICALLY SIMILAR CHARACTERS

Chinese characters are single syllable. The pronunciation of a Chinese character involves the nucleus and a tone, where the nucleus contains a vowel that follows an optional consonant. In this article, we use the Hanyu pinyin method to denote the sound of Chinese characters, and show the tone with a digit that follows the symbol string for the sound. In Mandarin Chinese, there are four tones. (Some researchers include the fifth tone.)

Although Chinese is not an alphabetical language, it is shown that the pronunciations of characters affect how people write Chinese [Ziegler et al. 2000]]. The pronunciation of a Chinese character has two parts: sound and tone. Therefore, the phonological similarity between two characters may consider these two aspects, and we consider four categories of phonological similarity between two characters: *s*ame sound and *s*ame tone (SS), *s*ame sound and *d*ifferent tone (SD), si*m*ilar sound and *s*ame tone (MS), and si*m*ilar sound and *d*ifferent tone (MD).

Table I. Samples of the Similar Phonemes with Example Characters

	Original Phoneme	Similar Phoneme	A Character with the Original Phoneme	Examples with the Similar Sound and the Same Tone (MS)
consonant	/s/	/sh/	肅 (**s**u4)	數、樹、恕 (**sh**u4)
	/z/	/zh/	早 (**z**ao3)	找、沼、爪 (**zh**ao3)
	/c/	/ch/	從 (**c**ong2)	重、虫、崇 (**ch**ong2)
vowel	/eng/	/en/	徵 (zh**eng**1)	真、針、貞 (zh**en**1)
	/eng/	/ang/	徵 (zh**eng**1)	張、章、蟑 (zh**ang**1)

We rely on the information provided in a lexicon [Dict 2010] to determine whether two characters have the same sound or the same tone. The judgment of whether two characters have a similar sound should consider the language experience of an individual. An individual who lives in southern China and one who lives in northern China, for instance, might have quite different perceptions of similar sound. In this work, we resort to the confusion sets observed in a psycholinguistic study, conducted at the Academic Sinica in Taiwan, to obtain a list of confusion sets of vowels and consonants in Mandarin Chinese.

Some Chinese characters are heteronyms [cf., Fromkin et al. 2002]. Let C_1 and C_2 be two characters that have multiple pronunciations. If C_1 and C_2 share one of their pronunciations, we consider that C_1 and C_2 belong to the SS category. This principle applies when we consider phonological similarity in other categories.

With a lexicon and the list of confusion sets, our program can select a list of phonologically similar characters for a given character. Consider the example "嚴肅" (yan2 su4) that we mentioned in Section 1. We can find all of the characters that have exactly the same pronunciation with "肅" (su4) based on the information provided by a lexicon. The SS list of "肅" will include characters such as "素" (su4) and "速" (su4). It is also easy to find the SD list for "肅", and it includes "蘇" (su1), "俗" (su2), and other characters. Based on the results of the psycholinguistic studies, we know that one might confuse the consonant /s/ with the consonant /sh/. Hence, the characters "數" (**sh**u4) and "樹" (**sh**u4) are in the MS list for "肅" (**s**u4), and the characters "書" (**sh**u1), "數" (**sh**u3), and "暑" (**sh**u3) are in the MD list for "肅". Notice that the character "數" is in the MS and MD lists for "肅" because it is a heteronym.

Table I shows more pairs of the confusing phonemes that we used in our system. Note that phonological similarity is a symmetric relationship. Namely, when phoneme X is similar to phoneme Y, phoneme Y is similar to phoneme X. To help readers focus on the symbols of the phonemes, we underline the confusing phonemes of the example characters in boldface. Notice also that, although we do not explicitly provide examples in Table I, it is possible to change both the consonant and the vowel for a character to find a phonologically similar character. For instance, "髒" (**z**ang1) is phonologically similar to "徵" (**zh**eng1) because we can replace the consonant /zh/ and vowel /eng/ in "徵" with /z/ and /ang/, respectively, and find "髒".

One challenge in defining phonological similarity between characters is that a Chinese character may be pronounced in more than one way, and the actual pronunciation depends on the context. Tone sandhi [Chen 2000] is a frequently mentioned source of confusion. The most common example of the use of tone sandhi in Chinese is that the first third-tone character in words formed by two adjacent third-tone characters will be pronounced with the second tone. For example, although "你" (ni3) and "好" (hao3) are both third-tone characters, "你" in "你好" is pronounced with the second

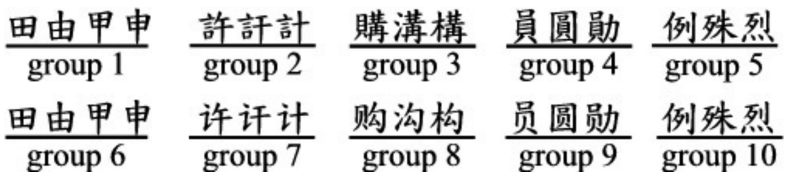

Fig. 1. Examples of visually similar characters in traditional Chinese (groups 1-5) and in simplified Chinese (groups 6-10).

tone in practice. Namely, native speakers usually pronounced "你好" as ni2-hao3. At present, we ignore the influences of context when determining whether two characters are phonologically similar. (As we shall see in Section 5, doing so did not disturb the experimental results.)

Although we have confined our definition of phonological similarity to the context of the Mandarin Chinese, we would like to note that the influence of sublanguages within the Chinese language family will affect the perception of phonological similarity. Dialects used in different areas in China, for example, Shanghai, Min, and Canton, share the same written forms with the Mandarin Chinese, but have quite different though related pronunciation systems. Hence, people living in different areas in China might perceive phonological similarity in different ways. The study in this direction, however, is beyond the scope of the current study.

3. VISUALLY SIMILAR CHARACTERS

Figure 1 shows examples of visually similar Chinese characters. The first row contains five groups of visually similar traditional Chinese characters, and the second row contains five corresponding groups of simplified Chinese characters. The j^{th} character (counted from left to right) in group $(i + 5)$ is the simplified form of the j^{th} character in group i. Notice that the traditional and simplified forms of a character may be exactly the same.

The characters in group 1 differ subtly at the stroke level, as do the characters in group 2. The characters in group 3 share the same components on their right sides. The shared components of the characters in group 4 and group 5 appear at different places within the characters.

Analogously, characters in group 6 differ subtly at the stoke level, as do the simplified characters in group 7. Characters in group 8 share the components on their right sides. The shared components of the characters in group 9 and group 10 appear at different places within the characters.

The radical of a Chinese character carries the main semantic information about the character [cf., Feldman et al. 1999], and lexicographers employ radicals to organize characters in Chinese dictionaries. Characters that belong to the same radicals are placed in the same category, and are listed sequentially by the number of strokes. Hence, it is possible to employ the information about radicals to find visually similar characters. The characters in group 1 and group 2 have the radicals "田" (tian2) and "言" (yan2), respectively. Analogously, the simplified characters in group 6 and group 7 have the radicals "田" and "讠", respectively. ("讠" is the simplified form of "言".) Notice that, although the radicals for group 2 and group 7 are obvious, those for group 1 and group 6 are not because "田" is not a standalone component in these groups.

Although radicals themselves provide information about the shared components of characters, the most saliently shared components of characters might not be the radicals of the characters. This problem occurs in both traditional and simplified Chinese. The shared component of the characters in group 3 is not the radical. The

shared components of the characters in groups 4, 8, and 9 are not the radicals for the characters in the groups either. In these cases, the shared components carry information about the pronunciations of the characters. Hence, those characters are listed under different radicals, though they do look similar in some ways.

In some cases, one may be interested in characters that share small elements in the characters, such as "歹" (dai3) in group 5 and group 10. The shared elements in these two groups do not carry semantic or phonological information, and they are not the radicals either. It is also possible that a radical is written in different ways in the characters that have the same radical in a dictionary, for example, "泉" (quan2) and "泊" (bo2). These two characters are listed under the radical "水" (shui3). The radical appears literally in "泉", but is written as "氵" in "泊".

Therefore, we cannot rely only on the information about radicals of characters in typical lexicons to find visually similar characters, and we will use the extended Cangjie codes as the basis to judge the degree of similarity between Chinese characters.

3.1 Cangjie Codes for Traditional Chinese

The Cangjie input method is one of the most popular methods used for entering Chinese characters into computers. The designer of the Cangjie method selected a set of 24 basic elements occurring in characters, and proposed a set of rules to decompose Chinese characters according to these elements [Chu et al. 2010]. Because the Cangjie system is designed to help people enter Chinese characters into computers, the design of the Cangjie codes had aimed at allowing its users to recall the codes for Chinese characters as easy as possible. Namely, users must be able to easily figure out the Cangjie codes for the characters that they want to enter. Given the popularity of the Cangjije input method in a wide range of Chinese speaking communities, the Cangjie codes of Chinese characters have practically shown their strong links with the formation of Chinese characters. This was an important motivation for us to try to define the similarity between two Chinese characters based on the degree of similarity between their Cangjie codes.

Table II shows the Cangjie codes for the 13 characters listed in groups 1 to 4 in Figure 1 and for five other characters. The "ID" column shows the identification number for the characters, and we will refer to the i^{th} character by c_i, where i is the ID. The "CC" column shows the Chinese characters, and the "Cangjie" column shows the Cangjie codes. Each symbol in the Cangjie codes corresponds to a key on the keyboard, for example, "田" (tian2) and "中" (zhong1) collocate with "W" and "L", respectively. Information about the complete correspondence is available in Wikipedia[5].

Using the Cangjie codes saves us from the need to apply image processing methods to determine the degrees of similarity between characters. Take the Cangjie codes for the characters in group 2 (c_5, c_6, and c_7) for example. See that the characters share a common component based on the shared substrings of the Cangjie codes (shown in boldface), that is, "卜 口" (bu3 kou3). We may also find the shared component "冓" (gou4, encoded by "廿廿月") for the characters in group 3 (c_{10}, c_{11}, and c_{12}), the shared component "力" (li4, encoded by "大尸") in c_{15} and c_{16}, and the shared component "巫" (jing1, encoded by "一一") in c_{16} and c_{17}.

However, the original Cangjie codes are still lacking in some respects, in spite of their perceivable advantages. The Cangjie codes have been limited to contain no more than five keys, in order to maintain efficiency in inputting Chinese characters. Thus,

[5]See http://en.wikipedia.org/wiki/Cangjie_input_method#Keyboard_layout; last visited on 30 September 2010.

Table II. Examples of Cangjie Codes for Traditional Chinese

ID	CC	Cangjie	ID	CC	Cangjie
1	田	田	10	購	月金廿廿月
2	由	中田	11	溝	水廿廿月
3	甲	田中	12	構	木廿廿月
4	申	中田中	13	員	口月山金
5	許	卜口人十	14	圓	田口月金
6	訐	卜口一十	15	勛	口金大尸
7	計	卜口十	16	勁	二二大尸
8	政	二二人大	17	頸	二二一月金
9	元	二二山	18	經	女火一女一

Table III. Examples of Cangjie Codes for Simplified Chinese

ID	CC	Cangjie	ID	CC	Cangjie
19	田	田	28	购	月人心戈
20	由	中田	29	沟	水心戈
21	甲	田中	30	构	木心戈
22	申	中田中	31	员	口月人
23	许	戈女人十	32	圆	田口月人
24	讦	戈女一十	33	勋	口人大尸
25	计	戈女十	34	劲	弓一大尸
26	鲳	弓一日日	35	颈	弓一一月人
27	驹	弓一心口	36	经	女一弓人一

users of the Cangjie input method must familiarize themselves with the principles for simplifying the Cangjie codes. While the simplified codes help to enhance the input efficiency, they also introduce difficulties and ambiguities when we compare the original Cangjie codes for computing similar characters. The shared component "員" (yuan2) is encoded in three different ways in c_{13}, c_{14}, and c_{15}, i.e., "口月山金" (kao2 yue4 shan1 jin1), "口月金", and "口金". The prefix "一一" in c_{16} and c_{17} can represent "巫" (jing1), "正" (zheng4; e.g., in c_8), and "二" (er4; e.g., in c_9). Consequently, characters whose Cangjie codes include "一一" may contain any of these three components, but c_8, c_9, and c_{16}do not really look alike.

3.2 Cangjie Codes for Simplified Chinese

Not surprisingly, the Cangjie codes are also useful for capturing the similarities between simplified Chinese characters. Using a structure similar to Table I, Table III shows the Cangjie codes for the characters listed in groups 6 to 9 in Figure 1 and five other characters.

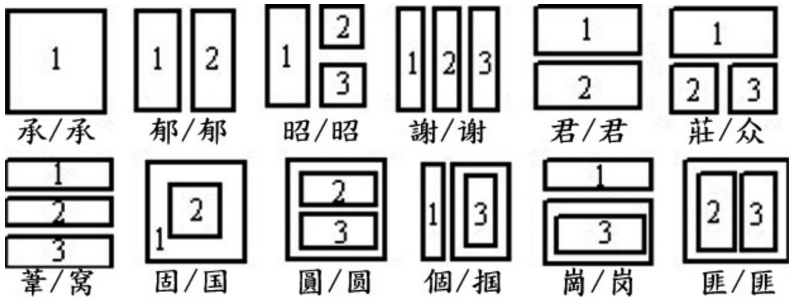

Fig. 2. Layouts of Chinese characters (used in Cangjie).

Again, the Cangjie codes offer the possibility to determine the degrees of similarity between characters efficiently. It is possible to find that the characters c_{23}, c_{24}, and c_{25}share a common component because their Cangjie codes share "戈女" (ge1 nu3). Using the common substrings (shown in boldface) of the Cangjie codes, we may also find the shared component "勾" (gou1, encoded by "心戈") for the characters in group 8 (c_{28}, c_{29}, and c_{30}), the shared component "员" (yuan2, encoded by "口月人") in c_{31} and c_{32}, the shared component "力" (li4, encoded by "大尸") in c_{33} and c_{34}, and the shared component "⻏" (jing1, encoded by "弓一") in c_{34} and c_{35}.

Similar to the problem of using the original Cangjie codes for traditional Chinese, we would encounter ambiguity problems when comparing the similarities between simplified Chinese characters. The shared component "员" in c_{32} and c_{33} is encoded by "口月人" (kao3 yue4 ren2) and "口人", respectively. The prefix "弓一" (gong1 yi1) in c_{34} and c_{35} can represent "⻏", "鱼" (yu2; e.g., in c_{26}), and "马" (ma3; e.g., in c_{27}). Characters whose Cangjie codes include "弓一" may contain any of these three components, but c_{26}, c_{27},and c_{34} do not really look alike.

Given the observations reported in the previous subsection and in this present one, we augmented the original Cangjie codes by using the complete Cangjie codes and annotated each Chinese character with a layout identification that encodes the overall contours of the characters.

3.3 Augmenting the Cangjie Codes

Figure 2 shows the 12 possible layouts that are considered for the Cangjie codes for both traditional and simplified Chinese characters. Most of the layouts contain two or three small regions (called subareas henceforth), and the rectangles show individual subareas within a character. The subareas are assigned IDs, but to maintain readability of the figures, not all of the IDs for subareas are shown in Figure 2. From left to right and from top to bottom, each layout is assigned an identification number from 1 to 12. An example pair of characters, separated by a slash, is provided below each layout. A traditional Chinese character is on the left, and a simplified one is on the right. For example, the layout ID of "固/国" is 8. "固" (gu4) is a traditional Chinese character, and has two parts, that is, "囗" (wei2) and "古" (gu3). "国" (guo2) is a simplified Chinese character and has two parts, that is, "囗" and "玉" (yu4).

When Chinese characters are transformed from the traditional to simplified forms, the layout of the same characters may or may not be changed, and a more comprehensive discussion about the significant change in the structures of the characters is available in Lee [2010b]. Hence, we may and may not use the traditional and simplified forms of the same character as a pair in Figure 2. Except for layouts 6, 7, 8, and 10, the pairs of characters shown under the layouts are the same characters in both traditional and simplified forms. For instance, "谢" (xie4) and "謝" (xie4) are

examples of layout 4, and "谢" is the simplified form of "謝". In contrast, "固" and "国" are two different characters, but both are examples of layout 8. The traditional form of "国" is "國", which belongs to layout 9.

Researchers have come up with other ways to decompose individual Chinese characters. A team at the Shanghai Jiao-Tong University (SJTU) report an early attempt, and they consider five major ways to decompose Chinese characters [p. 1071, SJTUD 1988]. In this study, the SJTU team report detailed analysis of the compositions of Chinese characters. Based on their analysis, "口" (kou3) and "木" (mu4) are the most frequent components in Chinese characters [p. 1027, SJTUD 1988]. Juang et al. [2005] employ four relationships for components of Chinese characters, and Sun et al. [2002] six relationships. The Chinese Document Laboratory at the Academia Sinica in Taiwan considers 13 possible ways to decompose Chinese characters [CDL 2010]. Lee [2010b] proposes more than 30 possible layouts. In Unicode standard 4.0.1, 12 operators are considered to build Chinese characters from a set of building blocks [UNICODE 2010].

The layout of a character affects how people perceive the visual similarity between characters [Yeh and Li 2002]. For instance, c_{16} ("勁", jing4) in Table II is more similar to c_{17} ("頸", jing3) than to c_{18} ("經", jing1), because the shared component of c_{16} and c_{17} is on the same side of the words. Overall, c_{16}, c_{17}, and c_{18} are more similar to each other than to "瘜" (jing1), although they share "巠" (jing1). We follow the style by which Chinese characters are decomposed in the Cangjie system, and rely on the expertise in Cangjie codes reported in Lee [2010a] to divide a Chinese character into subareas, which we showed in Figure 2.

Table IV shows the extended codes for some of the characters listed in Table I. The ID column provides links between the characters listed in both Table II and Table IV. The CC column repeats the Chinese characters. The LID column shows the identifications for the layouts of the characters. The columns with headings P1, P2, and P3 show parts of the extended Cangjie codes, where P*i* shows the i^{th} part of the Cangjie codes, as indicated in Figure 2.

LIDs are useful for comparing the degree of similarity between characters. Consider the case that we want to determine whether "勁" (jing4) is more similar to "頸" (jing3) than it is similar to "經" (jing1). Their extended Cangjie codes indicate that "頸" is a better answer for two reasons. First, both "勁" and "頸" are examples of layout 2; and, second, the shared components reside in the same subarea, that is, P1, in "勁" and "頸".

We decide the extended Cangjie codes for the individual parts with the help of computer programs and subjective judgments. Starting from the original Cangjie codes, we can compute the most frequent substrings among the original Cangjie codes for all characters. This process is similar to the one which can be used to compute the frequencies of *n*-grams in corpora [cf., Jurafsky and Martin 2009]. Computing the most frequently appearing substrings in the original codes is not a complex task because the longest original Cangjie codes contain just five symbols.

Often, the frequent substrings are simplified codes for common components in Chinese characters, for example, "言" (yan2) and "巠" (jing1). The complete codes for "言" and "巠" should be "卜一一口" (bu3 yi1 yi1 kou3) and "一女女女一" (yi1 nu2 nu2 nu2 yi1), but they are simplified to "卜口" and "一一", respectively, in the original Cangjie codes. When the Cangjie codes are simplified, "巠" has the same code as "正" (zheng4) and "二" (er4), as we have illustrated by c8, c9, and c16 in Table II. The simplified Cangjie codes for "言" are the same as the Cangjie codes of "亠", which is in the upper part of "高" (gao1).

After finding the frequent substrings, we verify whether these frequent substrings are simplified codes for meaningful components, which, in our definition, form parts

Table IV. Examples of Extended Cangjie Codes for Traditional Chinese

ID	CC	LID	P1	P2	P3
5	許	2	卜**一一**口	人十	
6	訐	2	卜**一一**口	一十	
7	計	2	卜**一一**口	十	
10	購	10	月**山**金	廿廿	**土**月
11	溝	10	水	廿廿	**土**月
12	構	10	木	廿廿	**土**月
13	員	5	口	月山金	
14	圓	9	田	口	月**山**金
15	勛	2	口**月山**金	大尸	
16	勁	2	一**女女女**一	大尸	
17	頸	2	一**女女女**一	一月**山**金	
18	經	3	女**戋**火	一女**女女**	一

of one or more Chinese characters. For meaningful components, we replace the simplified codes with their complete codes. For instance, the Cangjie codes for "許" (xu3) and "訐" (jie2) are extended to contain "卜一一口" in Table IV, where we indicate the extended keys that did not belong to the original Cangjie codes in boldface and with a surrounding box. After recovering the dropped codes for "言", our programs will have the information necessary to be able to tell "言" and "讠" apart.

Although we have tried to employ computer programs to help us find the frequent substrings in as many instances as we can, the work to recover the simplified codes remained labor-intensive, and we had to devote particular attention to certain anomalous cases at times. Fortunately, the process to implement the extended Cangjie codes proved to be worthwhile as we will show in the experimental studies.

Using a structure that is similar to Table IV, Table V shows the extended Cangjie codes for some of the simplified Chinese characters that we show in Table III. The "ID" column provides links between the characters listed in both Table III and Table V.

In Table V, we recover the Cangjie codes for "讠" (yan2) and "𠂤" (jing1). Using "戈弓女" (ge1 gong1 su3), rather than "戈女", for "讠" prevents us from confusing "讠" with "戉" (yue4). Similarly, using "弓人一" (gong1 ren2 yi1), rather than "弓一", for "𠂤" avoids the confusion of "马" (ma3) and "鱼" (yu2) with "𠂤", e.g., c26, c27, and c34 in Table III.

Replacing simplified codes with complete codes not only helps us avoid incorrect matches but also helps us find matches that would be missed due to the simplification of the Cangjie codes. If we only use the original Cangjie codes in Table III, it is not easy

Table V. Examples of Extended Cangjie Codes for Simplified Chinese

ID	CC	LID	P1	P2	P3
23	许	2	戈弓女	人十	
24	讦	2	戈弓女	一十	
25	计	2	戈弓女	十	
28	购	10	月人	心	戈
29	沟	10	水	心	戈
30	构	10	木	心	戈
31	员	5	口	月人	
32	圆	9	田	口	月人
33	勋	2	口月人	大尸	
34	劲	2	弓人一	大尸	
35	颈	2	弓人一	一月人	
36	经	3	女女一	弓人	一
37	恸	4	心	一一戈	大尸

to determine that c36 ("经", jing1) in Table III shares the component "⺯"(jing1) with c34 ("劲", jing4) and c35 ("颈", jing3). In contrast, there is a chance to find the similarity with the extended Cangjie codes in Table V, given that all of the three Cangjie codes include "弓人一" (gong1 ren2 yi1).

Although most of the examples provided in Table V indicate that we expanded only the first part of the Cangjie codes for the simplified Chinese, it is possible that the other parts, that is, P2 and P3, may need to be extended too. Sample c37 shows such an example.

3.4 Similarity Measure

The main differences between the original and the extended Cangjie codes are the degrees of detail about the structures of the Chinese characters. By recovering the details that were ignored in the original codes, our programs will be better equipped to find the similarities between characters.

We experiment with three different scoring methods to measure the visual similarity between two characters based on their extended Cangjie codes. Two of these methods were tried in our studies for traditional Chinese characters [Liu et al. 2009b, 2009c]]. The first method, denoted SC1, considers the total number of matched keys in the matched parts. Two parts are considered as matched as long as their contents are the same. They do not have to locate at the region within a character. Let c_i denote the i^{th} character listed in Table V. We have SC1(c_{33}, c_{34}) = 2 because of the matched "大尸"

(da4 shi1). Analogously, we have SC1(c_{37}, c_{34}) = 2 because of the matched "大尸". Notice that, although "一" (yi1) is in the P1 of c_{34} and in the P2 of c_{37}, it is not considered a match because the P1 of c_{34} and the P2 of c_{37} do not match as a whole.

The second method, denoted SC2, includes the score of SC1 and considers the following conditions: (1) add one point if the matched parts locate at the same place in the characters and (2) if the first condition is met, an extra point will be added if the characters belong to the same layout. Hence, we have SC2(c_{33}, c_{34}) = SC1(c_{33}, c_{34}) + 1 + 1 = 4 because (1) the matched "大尸" is the P2 of both characters and (2) c_{33} and c_{34} belong to the same layout. Assuming that c_{34} belongs to layout 5, than SC2(c_{33}, c_{34}) would become 3. In contrast, we have SC2(c_{37}, c_{34}) = 2. No extra points were added for "大尸" in the Cangjie codes for c_{37} and c_{34} because "大尸" is not at the same position in the characters. The extra points consider the spatial influences of the matched parts on the perception of similarity.

While splitting the extended Cangjie codes into parts allows us to tell that c_{33} is more similar to c_{34} than to c_{37}, it also creates a new barrier in computing similarity scores. An example of this problem is that SC2(c_{35}, c_{36}) = 0. This is because that "弓人一" (gong1 ren2 yi1) at P1 in c_{35} can match neither "弓人" at P2 nor "一" at P3 in c_{36}.

To alleviate this problem, we consider SC3 which computes the similarity in three steps. First, we concatenate the parts of a Cangjie code for a character. Then, we compute the longest common subsequence (LCS) (cf., Cormen et al. [2009]) of the concatenated codes of the two characters being compared, and compute a Dice coefficient (cf., Croft et al. [2010]) as the similarity. The Dice coefficients are used in many applications, including defining the strength of the relatedness of two terms (or similarity of two strings) in natural language processing (cf., Manning and Schütze [1999]). Let X and Y denote the concatenated, extended Cangjie codes for two characters, and let Z be the LCS of X and Y. The similarity is defined by the following equation.

$$Dice_{LCS} = \frac{2 \times |Z|}{|X| + |Y|}, \text{ where } |S| \text{ is the length of string } S \tag{1}$$

We compute another Dice coefficient between X and Y. The formula is the similar to Equation (1), except that we set Z to the longest common consecutive subsequence. We call this score $Dice_{LCCS}$. Notice that $Dice_{LCCS} \leq Dice_{LCS}$, $Dice_{LCCS} \leq 1$, and $Dice_{LCS} \leq 1$. Using both $Dice_{LCS}$ and $Dice_{LCCS}$ allows us to compute the visual similarity from two aspects. Finally, the SC3 of two characters is the sum of their SC2, $10 \times Dice_{LCCS}$, and $5 \times Dice_{LCS}$. We multiply the Dice coefficients with constants to make them as influential as the SC2 component in SC3. Since the LCCSs of two strings are generally quite shorter than the LCSs, we multiply $Dice_{LCCS}$with a larger weight. The constants were not scientifically chosen, but were selected heuristically.

Using the extended Cangjie codes and a selected score function, we can select a list of visually similar characters for a given character. Using SC3, we can find that every character in the string "逕涇輕徑痙莖紜氫" (jing4 jing1 qing1 jing4 jing4 jing1 yun2 qing1) is similar to "經" (jing1) in some way. Interestingly, each character in the string "逕涇輕徑痙莖紜氫" belongs to different radicals: "辶氵車彳疒艸糸气" (chuo4 shui3 che1 chi4 chuang2 cao3 mi4 qi4). One may verify that not all of the characters in the string are listed under the same radical as "經", so our approach offers chances to find visually similar characters that belong to different radicals.

3.5 Appropriate Similarity Measures

We discuss the main functions of these measures qualitatively, before we compare their effectiveness experimentally in Section 5.

Consider the problem of finding a set of "visually similar characters" for a given character within a Chinese word. Except resorting to the remembrance of human experts, how can we find similar characters from thousands of characters? A more important question may be what we mean by "similar" characters. Certainly, some characters can look similar individually. However, when putting into the contexts of words, some words can become more attractive than others. For instance, "浩" (hao4), "皓" (hao4), and "治" (zhi4) are almost equally similar to each other in some ways. Each pair differs in only in one of their components. Replacing the radical "氵" (shui3) in "浩" with the radical "白" (bai2) will create "皓", and replacing the component "告" (gao4) in "浩" with the component "台" (tai2) will create "治". However, "皓大" (hao4 da4), and "治大" (zhi4 da4) are not equally attractive for the writing of "浩大" (hao4 da4). Psycholinguistic evidence has shown that humans do not read text letter by letter for alphabetic languages or character by character for languages such as Chinese (e.g., Jackendoff [1995]). The contexts matter in determining the similarities.

As a result, the "best" similarity measure for computer software depend on the goals of the applications. Do we want to build a model for how humans judge visually-similar characters? Do we want to build a model for how human process confusing words in which some characters are visually similar?

In this article, the target application is more closely related to the latter question. In another application that we are building for learning Chinese characters [Liu et al. 2011], we are more concerned with similarity between individual characters. Hence, the similarity measures that we presented in the previous section were just to find "good candidates", and we will have another measure to compare these first-round candidates.

This is the main reason that we did not report experiments in which we carefully tuned the weights for SC3. The main function of SC3 in the current study was to find good first-round candidates.

Changing the current weights for $Dice_{LCS}$ and $Dice_{LCCS}$ also changes the order in the recommended characters. We illustrate the results of using three different sets of weights below. We show the alternative formulas for SC3, and the resulting recommended lists for "韻" (yun4) and "捐" (juan1) below. From left to right, recommended characters are listed in the order of descending scores. We do not show the Hanyu pingyin symbols of all of the listed characters to avoid congesting the page with the symbols for Chinese pronunciations.

Original: SC3 = SC2 + 10 × $Dice_{LCCS}$ + 5 × $Dice_{LCS}$.
韻: 損隕圓賞韶嘖遺磧貴員韹蹟賸績積臢瓚潰價漬槓慣債讚鑽賽
捐: 損涓娟扣狷拐揖絹揹搞胡咀硼拮採抬喟撰操捉抿湖啁高亭葫臺豪

Alternative 1: SC3 = SC2 + 10 × $Dice_{LCCS}$ + 0 × $Dice_{LCS}$
韻: 隕損賞圓韶嘖韹蹟賸績遺積磧臢瓚潰價漬槓慣債鑽貴員
捐: 損涓娟扣狷拐揖絹揹搞撰操硼拮採抬喟隕噯韻捉胡抿咀捌湖瑚糊啁

Alternative 2: SC3 = SC2 + 0 × $Dice_{LCCS}$ + 5 × $Dice_{LCS}$
韻: 隕損韶嘖賞圓遺讚韹蹟賸績躓積磧臢瓚潰價漬槓慣債鑽賽貴
捐: 涓娟損狷揹揖拐絹扣拮採抬喟捂唷探授掙啃揮援招搞搖

We can see that there are differences in the lists when we adopted different sets of weights in SC3. However, most, if not all, visually similar characters are included in the lists. Hence, treating these as the first-round candidates, we were satisfied with

the weights that we selected. A more practical mechanism to rank these candidate characters in the context of words will be introduced in Section 4.4. Due to that ranking mechanism, the resulting performances of different weights will not differ significantly for the current application, as long as we have chosen a satisfactory set of weights.

When we focus on just finding just visually similar characters, there will be no contextual information, that is, the words, available to rank the characters. In such cases, the weights certainly matter. Song et al. [2008] discuss related issues when they build a system for Chinese spelling checker. We [Liu et al. 2011] also face a similar problem when we need software to find characters that find Chinese characters that contain specific components.

4. DATA SOURCES AND PRELIMINARY ANALYSES

We provide information about our lexicons, the sources from which we obtained the reported errors in Chinese text, and our analyses of these reported errors in this section.

4.1 Lexicons

For both traditional and simplified Chinese, we prepare a lexicon that provides information on the pronunciation and a database that contains the extended Cangjie codes for the characters. Our programs rely on these databases to generate lists of characters that are phonologically and visually similar to a given character.

It is not difficult to acquire lexicons that contain information about standard pronunciations for Chinese characters. As we stated in Section 2, the main problem is that it is not easy to predict how people in different areas in China and Taiwan actually pronounce the characters. In the current study we employ the standards for Mandarin Chinese that are recorded in the lexicons and published by the official agency in Taiwan[6]. Experimental results reported in Section 5 will show that the ethnic background and mother tones did not influence the performance of our methods very much (at most 1%).

With the procedure reported in Section 3.3, we built databases of extended Cangjie codes for both the traditional and the simplified Chinese. Our database for the traditional Chinese was designed to contain 5,401 common characters in the BIG5 encoding system (between 0xa440 and 0xc67e), which was originally designed for the traditional Chinese. We will call this list of characters TCdict. We converted the traditional Chinese characters to their simplified counterparts and built the database of Cangjie codes for the simplified Chinese. Because two different traditional Chinese characters may be transformed to a common simplified form, this simplified list contains only 5,170 different characters, and we call this list of characters SCdict.

Count from the very first day of the conception of the main ideas, it took us a long time to develop the current TCdict and SCdict. The original idea was published in Liu and Lin [2008], but we continued to try different ideas since then. With the help of the software, that we explained in Section 3.3, to analyze the frequent substrings of the original Cangjie codes, two graduate students (the third and the fourth authors) were able to come up with a good version of the extended Cangjie code for the 5,401 traditional Chinese characters in a couple of weeks. That initial version was modified once in a while afterward. The modification operations were motivated by results of sporadic tests we ran with some data (Elist and Jlist, to be explained in Section 4.2), so

[6]See http://www.cns11643.gov.tw/AIDB/welcome_en.do.

we used some new data (Wlist and Blist, also to be explained in Section 4.2) to examine the performance of our system.

We employed our experience with the traditional Chinese to build the first and only version of the extended Cangjie codes for the simplified Chinese characters in few weeks. Most of the work was conducted only by the second author. We did not run experiments for the simplified Chinese while we are building the extended codes. Therefore, the experimental results that we report in Section 5.6 were not already based on new data.

4.2 Sources of Incorrect Words and Their Roles in Experiments

We acquired five lists of reported errors in Chinese at different stages of our study. By 2009, we collected two lists of errors for traditional Chinese, and in 2010, we added two lists of errors for traditional Chinese and a list of errors for simplified Chinese.

All of these lists contained information about the observed errors. In order to facilitate our experiments, we saved the reported errors in a simple format. An item of a reported error contains three parts: the correct word, the correct character that will be replaced, and the actual incorrect character. For instance, the correct way to write a type of banana is "芭蕉" (ba1 jiao1) and sometimes people use "笆" (ba1) for "芭" (ba1). In this case, we will maintain a data item "芭蕉,芭,笆" for this error.

At the beginning of our study, we acquired two lists of reported errors for traditional Chinese. The first list was obtained from a book published by the Ministry of Education (MOE) in Taiwan [MOE 1996]. The second list was collected in 2008 from the written essays of students of the seventh and the eighth grades in a middle school in Taipei. The errors were entered into computers based on students' writings, not including those characters that did not actually exist and could not be entered. We call the first list of errors the Elist, and the second the Jlist. Elist and Jlist contain, respectively, 1,490 and 1,718 items of errors.

Two or more different ways to write the same words incorrectly were listed in different items and considered as two items. When the same character of a word can be written incorrectly in multiple ways, for example, writing "應付" (ying4 fu4) as "應附" (ying4 fu4) or "應咐" (ying4 fu4) in Jlist, we considered them different errors. Cases like these make a program difficult to find the best actual incorrect character, as we will see in Sections 5.5 and 5.6.

Repeated or semantically related errors were treated as many times as the errors were committed by writers. Writing "變得更好" (bian4 de2 geng4 hao3) as "變的更好" (bian4 de1 geng4 hao3) and writing "變得更強" (bian4 de2 geng4 qiang2) as "變的更強" (bian4 de1 geng4 qiang2) can be considered repeated errors. Writing "變得更好" as "變的更好" and writing "作得不錯" (zuo4 de2 bu2 cuo4) as "作的不錯" (zuo4 de1 bu2 cuo4) can be considered related errors in lexical semantics. (These errors were observed in Jlist.)

These decisions helped us preserve the original distribution of the reported errors. That is, we took the test data as they were and did not try to manipulate or change the reported incorrect Chinese words. However, this also allowed a larger influence of the repeated errors on the reported experiment results.

In order to conduct further experiments, we collected two more lists of errors for traditional Chinese in 2010. The main reason for obtaining these lists was to use them as extra test data for our Cangjie codes that were improved during 2008 and 2009. Since we had access to both Elist and Jlist while we were improving the extended Cangjie codes for TCList, we thought it would be necessary to have new test data that we had never seen before to examine the effectiveness of the improved codes.

Table VI. Quantities of Reported Errors in Different Lists

Data Source	Original	Reduced	Data Source	Original	Reduced
Elist	1490	1333	Wlist	199	188
Jlist	1718	1645	Blist	487	385
			Ilist	684	621

The new datasets were acquired from independent sources. The first new list was collected from the Internet,[7] and the second new list came from errors discussed in a published book that was compiled by scholars [Tsay and Tsay 2003]. The first and the second lists contain 199 and 487 incorrect words, and we refer to these lists as Wlist and Blist, respectively.

In order to test whether our approach works for capturing errors in simplified Chinese, we searched the Internet for reported errors for simplified Chinese, and obtained two lists of errors. The first list[8] came from the entrance examinations for senior high schools in China, and the second list[9] contained errors that were observed at senior high schools in China. We used 160 and 524 errors from the former and the latter lists, respectively. Both of these lists of errors were produced by students at the senior high school levels, so we combined them into one list and refer to the combined list as Ilist.

We dropped some of the reported errors in our experiments because of the current scope of study. Some of the reported errors involved characters that did not belong to TCdict (for traditional Chinese) or SCdict (for simplified Chinese). Since we have extended the Cangjie codes for characters that were included only in TCdict for traditional Chinese and in SCdict for simplified Chinese, we ignored reported errors that did not occur in either TCdict or SCdict. This reduced the sizes of the lists that we collected. Table VI shows the sizes of the original and the reduced lists, respectively, under and the Original and Reduced columns.

4.3 Preliminary Error Analyses

In order to know the main reasons that caused the production of the observed errors, we asked two native speakers to classify the causes of these errors into three categories based on whether the errors were related to phonological similarity, visual similarity, or neither. Since the annotators did not always agree on their classifications, the final results are presented in five categories: P, V, N, D, and B in Table VII. P and V indicate that the annotators agreed on the types of errors to be related to, respectively, phonological and visual similarity. N indicates that both annotators believed that the errors were not due to phonological or visual similarity. D indicates that the annotators believed that the errors were due to phonological or visual similarity, but they did not have a consensus on the category. B indicates the intersection of P and V, that is, errors that are related to both phonological and visual similarities. Table VII shows the percentages of errors in these categories.

We used the quantities of reported errors in the reduced lists as the denominators to compute the percentages in Table VII. Hence, 79.9% in the "Jlist" row indicates that 1,314 (= 1645×0.799) errors were classified as related to phonological similarity. To get 100% for a row in the table, we need to add P, V, N, and D, and subtract B from the total.

[7]See http://www.eyny.com/archiver/tid-2529010.html; last visited on 30 September 2010.
[8]See http://www.0668edu.com/soft/4/12/95/2008/2008091357140.htm; last visited on 10 June 2010.
[9]See http://gaozhong.kt5u.com/soft/2/38018.html; last visited on 30 September 2010.

Table VII. Error Analysis of the Errors: Phonological Influences Dominated in These Errors

	P	V	N	D	B
Elist (traditional)	67.2%	66.1%	0.2%	3.6%	37.1%
Jlist (traditional)	79.9%	30.7%	2.4%	7.9%	20.9%
Wlist (traditional)	69.1%	54.8%	4.8%	8.0%	36.7%
Blist (traditional)	81.6%	34.8%	1.6%	4.7%	22.6%
Ilist (simplified)	83.1%	48.3%	0%	3.7%	35.1%

In all of these five lists, phonological similarity showed a dominant influence in respect to the visual similarity of the reported errors. Most of the reported errors were related to similar pronunciations, while the percentage of errors that were related to visual similarity depended on the lists of the reported errors. It should not be very surprising that the annotators may disagree sometimes.

The weighted proportion of phonologically related errors is 76.0%. Based on the statistics shown in Table VI and Table VII, this analysis considered 4,172 errors (the total of the errors in the reduced lists). The total number of errors that were related to similar pronunciation is $1333 \times 0.672 + 1645 \times 0.799 + 188 \times 0.691 + 385 \times 0.816 + 621 \times 0.831 = 3170.25$. The result of dividing 3170.25 by 4172 is 76.0%. Similarly, we can compute that the weighted proportion of visually related errors is $(1333 \times 0.661 + 1645 \times 0.307 + 188 \times 0.548 + 385 \times 0.348 + 621 \times 0.483) \div 4172 = 46.1\%$.

It is particularly noticeable that although the errors in Jlist were collected from written documents, the phonological factor still dominated. It is a common belief that the dominance of pronunciation-related errors in electronic documents occurs as a result of the common habit of entering Chinese with pronunciation-based methods. The ratio between P and V, that is, P ÷ V, for the Jlist challenges this popular belief and indicates that even though the errors occurred during a writing process, rather than typing on computers, students still produced more pronunciation-related errors. Distribution over error types is not as related to input method as one may have believed. Nevertheless, the observation might still be a result of students in Taiwan being so used to entering Chinese text with a pronunciation-based method that the organization of their mental lexicons is also pronunciation related. The P ÷ V ratio for the Ilist also supports this phenomenon, suggesting that the dominance of phonological influence may be a common phenomenon in the use of both traditional and simplified Chinese. The ratio for the Elist suggests that editors of the MOE book may have chosen the examples with a special viewpoint in their minds—that of balancing pronunciation and composition related errors. (The Blist is so short that we do not consider it representative in regard to this issue.)

It is worthwhile to note that a large percentage of errors are related to either phonological or visual similarity in Chinese. The sum of the statistics under N and D columns indicates the proportion of errors that were related to neither visual nor phonological similarity. The weighted average of (N + D) for the five lists was just 7%. The lowness of this figure can be explained by the large percentage of phono-semantic compounds (xingsheng words, "形聲字") in Chinese.

4.4 Web-Based Statistics

In this section, we examine the effectiveness of using Web-based statistics to differentiate correct and incorrect characters. The abundance of text material on the Internet allows people to treat the Web as a corpus[10]. When we send a query to Google, we will

[10]See http://webascorpus.org.

Table VIII. Reliability of Web-Based Statistics (Based on Data Collected in April 2010)

	Elist			Jlist			Ilist		
	C	A	I	C	A	I	C	A	I
P	92.4%	0.1%	7.5%	91.3%	0.9%	7.8%	97.1%	0.0%	2.9%
V	92.6%	0.0%	7.4%	91.5%	0.6%	7.9%	98.0%	0.0%	2.0%

be informed of the estimated number of pages[11] (ENOPs) that possibly contain relevant information. If we put the query terms in quotation marks, we should find the Web pages that replicate the query forms in the exact sequence and with the same adjacency as those originally entered. Hence, it is possible for us to compare the ENOPs for two competing phrases for guessing the correct way of writing a word. For instance, at the time of this writing, Google reported 116,000 and 33,000 relevant pages, respectively, for "strong tea" and "powerful tea". (When conducting such advanced searches with Google, the quotation marks are needed to ensure the adjacency of the individual words.) Hence, "strong" appears to be a better choice to go with "tea". This is an idea similar to one of the approaches for computing collocations based on word frequencies [cf., Manning and Schütze 1999]. Although the idea may not work very well when using a small database, the size of the current Web should be large enough.

We ran experiments for only those items that the annotators were in consensus over the causes of the error. Hence, for instance, we had 1285(= 1333 × (1-0.036), cf. Table VI and Table VII), 1515 (= 1645 × (1-0.079)), and 598(= 621 × (1-0.037)) such words for Elist, Jlist, and Ilist, respectively. As the information available on the Web may change over time, we also have to note that the statistics reported in Table VIII were based on experiments conducted during April 2010.

Table VIII shows the results of our investigation. For each reported error, we submitted the correct word and the incorrect word to Google and considered that we had a correct result when we found that the ENOP for the correct word was larger than the ENOP for the incorrect word. If the ENOPs were equal, we recorded an ambiguous result; and when the ENOP for the incorrect word was larger, we recorded an incorrect event. We use C, A, and I to denote correct, ambiguous, and incorrect events, respectively, in the table. We record a correct result for the "strong tea vs. powerful tea" test, for instance.

The Web-based statistics did not work very well for the Elist and Jlist, but seemed to work well enough for Ilist. The most common reason for the errors is that certain words are confusing to the extent that the majority of the Web pages showed the incorrect words. Some of the errors are so common that even one of the Chinese input methods on Windows XP offered wrong words as possible choices, for example, "雄赳赳" (xiong2 jiu1 jiu1; the correct one) vs. "雄糾糾" (xiong2 jiu1 jiu1). It is also interesting to note that people may intentionally use incorrect words on some occasions; for instance, people may choose to write homophones in advertisements.

Another possible reason for the mistakes is that whether a word is correct depends on a larger context. For instance, "小斯" (xiao3 si1) is more popular than "小廝" (xiao3 si1) because the former is a popular nickname. Unless we provided more contextual information about the queried words, checking only the ENOPs of "小斯" and "小廝" would lead us to choose "小廝", which would be an incorrect word if we meant to find the right way to write "小斯". Other difficult pairs of words to distinguish are"紀錄" (ji4 lu4) vs. "記錄" (ji4 lu4) and "須要" (xu1 yao4) vs. "需要" (xu1 yao4).

[11]According to Croft et al. [2010], the ENOPs may not reflect the actual number of pages on the Internet, they may result from statistical estimations.

Yet another reason for having a large ENOP for the incorrect words was due to errors in segmenting Chinese character strings (cf., Ma and Chen [2003]). Consider a correct character string "WXYZ". It is possible that "XY" happens to be an incorrect way to write a correct word. This is the case for having the counts for "花海繽紛" (hua1 hai3 bin1 fen1) to contribute to the count for "海繽", which is an incorrect form of "海濱" (hai3 bin1).

A reason why the Web-based statistics worked for Ilist is that all of the correct words in Ilist contained four characters. None of the factors that we list above for explaining the errors that we observed from Elist and Ilist may reasonably apply in the case of four-character strings. Hence, Web-based statistics worked almost perfectly for Ilist.

We compared the statistics reported in Table VIII and our reports in Liu et al. [2009c] and found that the effectiveness for Elist and Jlist improved greatly. To understand this phenomenon, we examined the records of our experimental results. Let x and y be the ENOPs of the correct and incorrect words, respectively. When we reported cases in which our systems failed to identify the correct character, that is, the I cases, in 2009, many of the ratios of y against x were within approximately 5%. Namely, we had $(y/x) \leq 1.05$in many cases, indicating that the margins were very small in 2009. The differences between the ENOPs may have changed between 2009 and 2010, so making us to achieve the better performance reported in Table VIII.

The improvement shows that the reliability of the Web-based statistics may be improving as the correct usage of words increase. This also shows that our approach might not work very well if the majority of Web-page authors do not use the characters in standard ways. To avoid the problem of using only the raw values of ENOPs to judge the correctness of words, we will introduce an alternative method in the next section.

5. EXPERIMENTAL EVALUATION

We evaluate the effectiveness of using the phonologically and visually similar characters to capture errors in Chinese words in this section.

In Section 5.1, we provide details about the procedures for the experiments, and, in Section 5.2, we explain the definitions of inclusion rates that we used to evaluate the basic performance of our systems. In Section 5.3, we offer a comparison of the statistics about our experiments that were conducted in 2009 and 2010, in order to gather information about the reliability of Web-based statistics, which is crucial for the success and applicability of our systems in the long run. In Section 5.4, we report the inclusion rates of our systems on the five sources of test data. The experiments also show the robustness of our methods and data for processing test data that were not reported in previous conference articles. In Sections 5.5 through 5.7, we deepen and widen our investigation of the effectiveness of our systems by applying an alternative method to rank the candidate characters and by recommending a limited number of candidates characters based on two ranking mechanisms.

5.1 Experimental Procedure

We designed and employed the ICCEval procedure for the evaluation task. We needed two types of data for the experiments. The information about the pronunciation and structures of the Chinese characters (Section 4.1) helped us generate lists of similar characters. We also needed reported errors (Section 4.2) so that we could evaluate whether the similar characters catch the reported errors.

At step 1, we created a list of characters based on the selection criterion, given the correct word and the correct character to be replaced. We may choose to evaluate the

Procedure ICCEval
Input:
A test item that includes the following information:
(Section 4.2)
cwd: the correct word,
ccr: the correct character,
aic: the actual incorrect character;
crit: the selection criterion;
num: number of requested characters;
rnk: the criterion to rank the incorrect words;
Output: a list of ranked candidates for **ccr**
Steps:
1. Select a candidate list, L, of characters, from TCdict or SCdict, for **ccr** with the specified criterion, **crit**. When using SC1, SC2, or SC3 to select visually similar characters, at most **num** characters will be selected.
2. Check whether **aic** belongs to L (the inclusion test). if yes, continue; else, return nil.
3. For each x in L, replace **ccr** in **cwd** with x, submit the resulting x and incorrect word to Google, and record the ENOPs for x and the incorrect word.
4. Rank the characters in L with the criterion specified by **rnk**.
5. Return the ranked list.

effectiveness of phonologically or visually similar characters. For a given correct character, ICCEval can generate characters that are in the SS, SD, MS, and MD categories for phonologically similar characters (Section 2). For visually similar characters, ICCEval can select characters based on different score functions, i.e., SC1, SC2, and SC3 (Section 3.4). In addition, ICCEval can generate a list of characters that belong to the same radical and have the same number of strokes with the correct character. In the experimental results, we refer to this type of similar characters as RS.

At step 2, we checked whether the selected list of characters indeed contained the actual incorrect character.

At step 3, for a correct word that people should write, we replaced the correct character with a character from the candidate list that was generated at step 1, submitted the incorrect word to Google AJAX Search API (or directly to the Google interface[12], and extracted the ENOP of pages that contained the incorrect words. In an ordinary interaction with Google, an ENOP can be retrieved from the search results, and it typically follows the string "Results 1-10 of about" in the browser window. Using the Google AJAX Search API, we have only to parse the returned results using a simple method.

Larger ENOPs for incorrect words suggest that these words are incorrect words that people frequently used on their Web pages. Hence, we could rank the similar characters based on their ENOPs at step 4 and return the list.

Since the reported errors contained information about the actual incorrect ways to write the correct words (Section 4.2), we could check whether the actual incorrect characters were among the similar characters that our programs generated at step 2 (inclusion tests). We could also check whether the actual incorrect characters were ranked higher in the ranked lists (ranking tests).

Take the word "和藹可亲" (he2 ai3 ke3 qin1) as an example. In the collected data, it was reported that people wrote this word as "和靄可亲" (he2 ai3 ke3 qin1), that is, the second character was incorrect. Hence the test item (correct word, correct character, actual incorrect character) is ("和藹可亲", "藹", "靄"). Hoping to capture the error, ICCEval generated a list of possible substitutions for "藹" at step 1. Depending on the

[12]See http://www.google.com.

categories of sources of errors, ICCEval generated a list of characters. When aiming to test the effectiveness of visually similar characters, we could ask ICCEval to apply SC3 to selected a list of alternatives for "藹" from SCdict, and the results may include "靄" (ai3), "谒" (ye4), "葛" (ge3), and other candidates. At step 2, we found that the actual incorrect character was included in the candidate list. At step 3, we created and submitted the query strings "和靄可亲", "和谒可亲", and "和葛可亲" to obtain the ENOPs for the candidates. If the ENOPs were, respectively, 571,000, 445,000, and 8,580, these candidates would be returned in the order of "靄", "谒", and "葛". As a result, the returned list contained the actual incorrect character "靄", and placed "靄" at the top of the ranked list.

Notice that we considered the contexts in which the incorrect characters appeared to rank the candidate characters. We did not rank the incorrect characters with the unigrams such as "靄", "谒", and "葛" alone. Instead, the candidates were ranked with the ENOPs of "和靄可亲", "和谒可亲", and "和葛可亲".

In addition, although this running example shows that we ranked the characters directly with the ENOPs, we also tried to rank the list of alternatives with the pointwise mutual information [PMI; cf., Jurafsky and Martin 2009]:

$$PMI(C, X) = \frac{\Pr(C \wedge X)}{\Pr(C) \times \Pr(X)}, \tag{2}$$

where X is the candidate character to replace the correct character and C is the correct word excluding the correct character to be replaced. To compute the score for the replacement of "藹" with "靄" in "和藹可亲", X = "靄", C="和□可亲", and $(C \wedge X)$ is "和靄可亲". (□ denotes a character to be replaced.) We chose to try the frequency-based method and PMI-based method because both are used to compute the strength of collocation in natural language processing [Manning and Schütze 1999].

It would demand a considerable amount of computation effort to find $\Pr(C)$ in general, if this is a required task. Fortunately, we do not have to consider the effect of $\Pr(C)$ because it is a common denominator for all possible incorrect characters for a given incorrect word. Let X_1 and X_2 be two competing incorrect characters for the correct character. We can ignore $\Pr(C)$ because of the following relationship.

$$PMI(C, X_1) \geq PMI(C, X_2) \Leftrightarrow \frac{\Pr(C \wedge X_1)}{\Pr(X_1)} \geq \frac{\Pr(C \wedge X_2)}{\Pr(X_2)} \tag{3}$$

Hence, X_1 prevails if $score(C, X_1)$ is larger, where $score(C, X)$ for any$X$is listed in Equation (4).

$$score(C, X) = \frac{\Pr(C \wedge X)}{\Pr(X)} \tag{4}$$

In our work, we approximate the probabilities used in Equation (4) by the corresponding frequencies. Namely, we replace $\Pr(C \wedge X)$ with the Web-based counts for $(C \wedge X)$, for example, "和靄可亲"; and substitute $\Pr(X)$ with the Web-based counts for X, for example, "靄". The counts were obtained with exactly the same mechanism that we used to obtain the ENOPs that we explained at the beginning of this section.

5.2 Performance Measures

Recall that we used only those errors for which the annotators had reached consensus on the causes of the errors. The errors that involved characters that were not in TCdict and not SCdict were not considered in the current study either.

Given the errors in the lists in Table VI, we ran ICCEval**,** and measured the performance in two ways. First, we would like to have the candidate list (step 1 in ICCEval)

include the actual incorrect character in the inclusion test. Second, we would prefer that the actual incorrect character be placed at the top of the ranked list (step 4 in ICCEval).

Assume that there were n items of errors in a given list and that the candidate lists for these n errors contained m of the n actual incorrect character. Then, we compute the inclusion rates in the following manner.

$$\text{inclusion rate} = \frac{m}{n} \tag{5}$$

In order to compare whether it is easier to capture either the phonologically similar or the visually similar errors, we separate the test instances according to the annotators' consensuses. Hence, we will provide separate inclusion rates for two types of error. When reporting the inclusion rates for phonologically similar errors, we use the number of phonologically-similar errors in place of n, and when reporting the inclusion rates for visually-similar errors, we use the number of visually-similar errors in place of n.

The inclusion rates are similar to the recall rates that are commonly-used in the field of information retrieval (e.g., Croft et al. [2010]). Yet, they are different. The recall rates measure how well a search engine retrieves the correct answers for a query. In our experiments, each test has only one answer, that is, the actual incorrect character. Hence, we measure the probability whereby we would capture the actual incorrect character across the reported errors.

In addition, we prefer to put the actual incorrect character higher in the ranked list. Hence, we analyzed the accumulative inclusion rates of the top k characters in the candidate list. Assume that there were n items of errors in a given list. Also assume that ICCEval placed the actual incorrect characters at the j^{th} position t_j times in the n experiments, where the first position is the best position. The ratio ($t_j \div$n) is the probability that the actual incorrect character was ranked at the j^{th} position in an individual test. Then, we compute the accumulative inclusion rate (AIR) in the following way.

$$\mathrm{R}_k = \frac{\sum_{j=1}^{k} t_j}{n} \tag{6}$$

If ICCEval returns the candidate list completely, then it will achieve the inclusion rate. If ICCEval returns only the top k candidates, it will achieve the accumulative inclusion rates, which are upper-bounded by the inclusion rate.

5.3 Temporal Comparison

In April 2010, we repeated the experiments that we conducted in March 2009. The main purpose was to inspect whether we would achieve performance of similar quality with Web-based statistics in 2009 and 2010. The experiments had two goals: (1) to have the candidate lists (step 1 in ICCEval) include the incorrect characters in the reported errors, and (2) to put the actual incorrect characters at top of the ranked results (step 4). We used Elist and Jlist in the experiments.

Because we kept the lists of candidate characters that we generated at step 1 in ICCEval in 2009, we were able to resubmit the incorrect words at step 3 and collected the ENOPs to rank the incorrect characters in 2010. In 2009, we submitted our queries directly through the Google interface[13] and we repeated the same procedure in the experiments in 2010 again. (Submitting a large amount of queries directly to the standard Google interface resulted in Google considering our programs to be a malicious

[13]See http://www.google.com.

Table IX. Inclusion Rates in 2009 [Liu et al. 2009b]

	SS	SD	MS	MD	SC1	SC2	RS	Phone	Visual	All
Elist	91.6	18.4	3.0	1.9	76.1	73.9	4.1	99.0	82.0	93.4
Jlist	92.2	20.2	4.2	3.7	74.5	67.5	6.1	99.3	77.6	97.3

attacking agent, and our computers might be blocked. Hence, we have switched to the Google AJAX search API for the other new experiments.)

Since we reused the candidate lists, the inclusion rates did not change over time, so we show the inclusion rates in Table IX. We show the inclusion rates of the candidate lists that were generated with different criteria, that is, SS or SC1 for Elist and Jlist. We did not run SC3 for this experiment because we did not have this score function in 2009. Using SS to recommend candidate characters, we captured 91.6% of the errors that were related to pronunciation in Elist. Since in Table VII, 67.2% of the errors in Elist was related to phonological similarity, using the SS list alone captured (91.6 $\times$ 67.2)%, that is, 61.6%, of all of the errors. Using SC1 to recommend candidate characters, we captured 74.5% of the errors that were caused by visual similarity, and that is (74.5 $\times$ 66.1)%, that is, 49.2%, of all of the errors in Elist. The Phone column shows the inclusion rates when we used the union of the recommend lists created with SS, SD, MS, and MD criteria to guess the actual incorrect characters that were caused by phonological similarity. The Visual column shows the inclusion rates when we used the union of the candidate lists created with SC1, SC2, and RS criteria to guess the actual incorrect characters that were caused by visual similarity. The All column shows the inclusion rates when we used the union of the candidate lists created with SS, SD, MS, MD, SC1, SC2, and RS criteria to guess all of the actual incorrect characters.

Apparently, it was much easier to capture errors that were related to the phonological similarity. All together, we were able to capture more than 95.5% of the nearly 2,978 observed errors. (cf., Table VI and Table IX; 95.5 = (93.4 $\times$ 1333 + 97.3 $\times$ 1645) $\div$ (1333 + 1645)) Recall that SS stands for "same sound and same tone." SS is the most effective criterion to use to select a candidate list that can capture the observed errors. Even so, candidate lists selected with other criteria were necessary to achieve 99% inclusion rates for the phonologically related errors.

In contrast, using the extended Cangjie code that we constructed manually in 2009 did not perform comparatively well for visually-related errors, although achieving an inclusion rate of 80.4% in 2009 was very encouraging. It is worth mentioning that the RS category was able to capture 4.8% of the visually similar errors. (cf., Table VI, Table VII, and Table IX; 80.4 = (82.0 $\times$ 1333 $\times$ 0.661 + 77.6 $\times$ 1645 $\times$ 0.307) $\div$ (1333 $\times$ 0.661 + 1645 $\times$ 0.307). We used only those errors that were caused by visual similarity in this calculation.)

Table X shows the accumulative inclusion rates up to the 10^{th}-ranked candidates for the errors in Elist. If we subtract the results for 2009 from the corresponding numbers for 2010 in the table, we will see that there is no significant difference between the data shown in the upper and the lower part of Table X. We also reran the tests for Jlist, and did not observe any significant differences in statistics for the ranked lists in 2009 and 2010, either. Web-based statistics therefore appeared to be very stable, for the task of ranking the candidate incorrect characters.

5.4 Inclusion Tests

We have changed several aspects of our system since we conducted the experiments reported in the previous subsection. We have built a new version of the extended Cangjie codes for the characters in TCdict and the first version of the extended Cangjie codes for the characters in the SCdict, using the procedure that we discussed in Section 3.3.

Table X. AIRs for Ranking the Candidates for Elist Based on the Frequencies Collected in 2009 and 2010

		R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_{10}
March 2009	SS	55.5	74.6	82.4	85.8	87.6	89.3	90.1	90.7	90.7	91.1
	SD	10.0	13.8	15.4	16.3	16.6	17.1	17.4	17.6	17.9	18.3
	MS	2.1	2.3	2.5	2.7	2.9	3.0	3.0	3.0	3.0	3.0
	MD	1.1	1.4	1.4	1.4	1.4	1.4	1.7	1.7	1.7	1.7
	SC1	33.6	49.4	55.9	61.4	65.8	67.7	68.8	71.2	72.0	73.2
	SC2	30.6	43.5	50.7	56.8	60.6	64.1	65.5	67.1	68.7	70.1
	RS	2.7	3.5	3.7	4.0	4.1	4.1	4.1	4.1	4.1	4.1
April 2010	SS	55.1	73.4	81.7	86.0	88.4	90.0	90.4	90.9	91.1	91.2
	SD	10.0	13.2	15.1	16.1	16.6	16.9	17.3	17.6	17.8	17.9
	MS	2.1	2.5	2.8	2.8	2.9	2.9	3.0	3.0	3.0	3.0
	MD	1.0	1.4	1.4	1.4	1.6	1.6	1.6	1.6	1.7	1.7
	SC1	31.1	48.0	57.1	62.8	66.6	68.4	69.9	70.8	72.7	73.7
	SC2	29.4	43.7	52.1	58.5	62.7	64.5	65.9	67.0	68.6	69.6
	RS	2.8	3.5	3.8	3.8	4.0	4.0	4.0	4.1	4.1	4.1

We have also added a new score function, that is, SC3, that we did not have in 2009. Furthermore, we do not submit our queries to the ordinary Google interface anymore, as we explained in Section 5.3.

To make the results of the experiments with traditional Chinese more convincing, we used two new lists, Wlist and Blist that we did not know when we improved the Cangjie codes in TCdict. These two lists serve as unforeseen test instances for our programs and databases.

We ran ICCEval with Elist, Jlist, Wlist, Blist, and Ilist. The experiments were conducted for all categories of phonological and visual similarity. When using SS, SD, MS, MD, and RS as the selection criteria, we did not limit the number of candidate characters. Any characters that conform to the selection criteria were selected in the candidate list, L in ICCEval. When using SC1, SC2, and SC3 as the selection criteria, we limited the number of candidates to be no more than 30. We inspected samples of the candidate lists generated with SC1, SC2, and SC3, and found that the number of visually similar characters for a given character rarely exceeded 30. Hence this limit was chosen heuristically.

We considered only words that the native speakers were in consensus over the causes of errors. There is a limit on the maximum number of queries that one can submit to the Google AJAX API. As a consequence, we could not complete our experiments in a short time, and the ENOPs were obtained during March and April 2010. Table XI shows the inclusion rates that the candidate lists, generated with different crit at step 1, contained the incorrect characters in the reported errors. The columns have the same meaning as they have in Table IX. The rows show the statistics that we observed while using the lists, listed in Table VI, as the test data. For instance, we achieved an inclusion rate of 90.3% for the visually similar errors when we applied SC3 to generate the candidate lists for errors in the Wlist.

`ICCEval` and our databases worked well for traditional Chinese. Although we have slightly expanded the definitions of similar sound since 2009, the effectiveness of SS, SD, MS, and MD remain the same for Elist and Jlist in Table IX and Table XI. The statistics for RS did not change because we were using the same list in 2009 and in 2010. Statistics about SC1 and SC2 are slightly better in Table XI for both Elist and Jlist, but the improvements are not significant. Using the new score function, that is,

Table XI. Inclusion Rates for the Different Experiments

	SS	SD	MS	MD	SC1	SC2	SC3	RS	Phone	Visual	All
Elist	91.6	18.4	3.0	1.9	77.7	76.3	87.3	4.1	99.0	89.8	96.2
Jlist	92.2	20.2	4.2	3.7	77.0	71.3	89.3	6.1	99.3	91.9	99.3
Wlist	94.6	23.1	0.8	0.8	80.6	78.6	90.3	1.1	99.2	90.3	96.9
Blist	82.2	24.2	3.2	1.9	77.6	75.4	90.3	3.7	95.2	91.8	94.7
Ilist	82.6	29.3	2.1	1.6	78.3	71.0	87.7	1.3	97.3	90.0	96.5

SC3, and the new Cangjie codes significantly improved the inclusion rates for visually related errors. We were able to include 88.3%[14] of the visually similar errors with SC3. We were able to include approximately 10% more of the actual incorrect characters in experiments, when we used SC3 rather than SC1 or SC2 to generate the candidate lists.

Even though we had not seen the errors in Wlist and Blist[15] previously, our program had shown a robust performance. ICCEval achieved a comparable performance when running with Wlist than running with Elist and Jlist. When working with Blist, ICCEval did not perform as well with phonologically similar errors, but showed a similar performance for visually similar errors. However, the change was not very significant, and the results reflected the preference of the experts who wrote the book from which we obtained Blist. The effectiveness of the SS lists dropped, while the effectiveness of the SD lists increased. We inspected the errors in Blist closely, and found many challenging instances—those that native speakers found the incorrect words are becoming more frequently used in practice. For this reason, we considered that ICCEval achieved reasonably well.

When running with Ilist, ICCEval achieved a performance similar to the one which it had achieved with Blist. Like the results observed in the other experiments, it is easier to find phonologically similar incorrect characters than visually similar ones. Using SS and SC3 as the selection criterion at step 1 in ICCEval were the most effective criteria for phonologically and visually similar characters, respectively. The contribution of SD was quite significant for Ilist.

When we used the unions of the phonologically similar characters to compute the inclusion rates, we captured 98.6% of the phonologically similar errors for the five lists. The unions of the visually similar characters were also very effective, though capturing only about 90.5% of the visually similar errors. When we used the union of all of the candidate lists, we captured 97.4% of all the errors. These are the weighted averages of the inclusion rates which we calculated with a procedure similar to the one provided in the tenth footnote.

It is certainly desirable for applications to have the potential to capture all of the reported errors. However the inclusion rates were achieved at different costs. For each reported error and the actual incorrect character of the error, ICCEval generated a candidate list at step 1. In an experiment that used a list of y errors, we would have y candidate lists. Hence, for a particular experiment, we can calculate the average length of such $y$candidate lists. Table XII shows the average lengths of the corresponding experiments that are reported in Table XI.

[14]The rates are averages computed considering the numbers of errors in the error lists, listed in Table VI and Table VII. For instance, 88.3 = (87.3 × 1333 × 0.661 + 89.3 × 1645 × 0.307 + 90.3× 188 × 0.548 + 90.3 × 385 × 0.348 + 87.7 × 621 × 0.483) ÷ (1333 × 0.661 + 1645 × 0.307 + 188 × 0.548 + 385 × 0.348 + 621 × 0.483).

[15]We used only 20 of the reported errors in Blist in Liu et al. [2009a].

Table XII. Average Lengths of the Candidate Lists

	SS	SD	MS	MD	SC1	SC2	SC3	RS	Phone	Visual	All
Elist	11.3	18.6	9.7	21.8	23.3	26.7	25.4	9.2	56.6	48.8	102.1
Jlist	12.4	22.0	11.6	25.4	21.9	24.3	25.4	7.7	64.3	46.0	107.4
Wlist	11.8	17.4	10.7	22.0	22.6	26.1	25.5	9.2	56.4	48.8	102.0
Blist	14.2	22.2	10.4	22.5	22.2	25.6	25.7	8.1	62.5	47.4	106.9
Ilist	12.6	19.1	9.1	19.5	24.3	27.1	25.5	9.4	55.5	47.8	100.2

Clearly, longer candidate lists would increase the chances to achieve higher inclusion rates. Hence, it would be more preferable if a shorter candidate list can achieve the same inclusion rate as that of a longer candidate list. From this perspective, the statistics in Table XII show that SS is very effective for capturing phonologically-similar errors, as we were able to capture better than 89.8% of the phonologically-similar errors by an average of 12.2 characters. (12.2 is the weighted average of 11.3, 12.4, 11.8, 14.2, and 12.6.) Taking the union of SS, SD, MS, and MD lists to obtain the "Phone" list, the average lengths increased from 12.2 to 60.0, but the inclusion rates increased from 89.8% only to 98.5%.

Using SC3 as the selection criterion for visually similar errors offered a significant improvement in both effectiveness and efficiency. The weighted average lengths of the candidate lists that were selected with SC1, SC2, and SC3 were 22.8, 25.7, and 25.4, respectively. The weighted average inclusion rates for SC1, SC2, and SC3 were 77.6%, 73.6%, and 88.6%, respectively. Using SC3 allows us to achieve higher inclusion rates with shorter candidate lists. We took the union of the SC1, SC2, SC3, and RS lists to form the Visual list, increased the average lengths of the candidate lists to 47.4, but increased the inclusion rates only marginally to 90.9%.

To achieve the inclusion rates in the All column in Table XI, we would have to allow ICCEval to recommend 104.3 characters. Although a list of 104 characters will be too long to be practically useful, we have to keep in mind that we had reduced the search space from more than 5,100 characters to approximately 100 characters—which is 98% for the compression rate. This point is particularly important to bear in mind as we seek to applying our findings to help teachers select "attractive incorrect characters" when authoring test items of the ICC tests.

5.5 Ranking Tests with Elist and Jlist

To make the candidate lists applicable, we wish to place the actual incorrect character high in the ranked list. This will help the efficiency in supporting computer-assisted test-item writing. Having shorter lists that contain relatively more confusing characters may facilitate the data preparation for psycholinguistic studies as well.

Table XIII shows the results when we recommended only the leading ten candidates for the errors in Jlist. The table is divided into two parts. The upper part (with row heading "Frequency") shows the results when we used the raw values of the ENOPs to rank the candidate characters, and the lower part (with row heading "PMI") shows the results when we used Equation (4) in Section 5.1 to rank the candidate characters. The column "R_i" shows the accumulative inclusion rates (AIRs) that we defined in Equation (6) in Section 5.2. The sub-row headings show the selection criteria that were used in the experiments. For instance, using SS as the criterion and ranking with the raw values of ENOPs, 55.1% of phonologically related errors were included if we offered only one candidate, 70.6% of the phone-related errors were included if we offered two candidates, etc. If we recommended only the top five candidates in SS (ranked with ENOPs), we captured the phonologically similar errors 84.3% of the time. For errors that were related to visual similarity, recommending the top five candidates

Table XIII. AIRs for Ranking the Candidates for Jlist Based on ENOPs and PMIs

		R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_{10}	R_{num}
Frequency	SS	55.1	70.6	77.6	81.7	84.3	86.2	87.6	88.2	89.2	89.7	92.2
	SD	10.6	13.8	15.9	16.9	17.7	18.0	18.1	18.5	18.7	18.8	20.2
	MS	3.0	3.3	3.5	3.7	3.7	3.7	3.7	3.8	3.8	3.9	4.2
	MD	2.2	2.7	2.9	3.0	3.1	3.3	3.3	3.3	3.3	3.3	3.7
	Phone	42.9	57.5	64.6	70.1	73.7	77.7	81.1	83.0	84.6	86.1	99.3
	SS+SD	43.7	56.3	64.9	71.7	75.5	78.7	81.1	83.2	84.4	85.5	95.1
	SC1	40.2	53.5	59.8	64.2	65.9	67.7	68.3	69.1	70.5	72.1	77.0
	SC2	34.7	48.5	52.1	55.4	57.6	60.0	62.0	63.0	64.4	65.5	71.3
	SC3	42.6	56.6	64.4	69.7	73.9	77.0	78.8	80.4	81.6	83.6	89.3
	RS	5.3	5.9	5.9	5.9	5.9	5.9	5.9	5.9	6.1	6.1	6.1
	Visual	35.3	50.4	57.9	61.5	66.1	69.2	72.4	74.0	76.0	77.6	91.9
PMI	SS	47.0	62.2	70.7	75.8	79.0	82.8	84.1	85.5	86.9	87.9	92.2
	SD	9.4	12.0	14.1	15.3	15.8	16.6	16.8	17.7	18.2	18.4	20.2
	MS	2.7	3.3	3.3	3.6	3.6	3.6	3.7	3.7	3.7	3.7	4.2
	MD	2.1	2.5	2.7	2.8	2.8	3.0	3.0	3.0	3.1	3.1	3.7
	Phone	37.0	51.0	59.7	64.5	69.0	72.7	75.6	77.7	79.3	80.5	99.3
	SC1	38.2	49.9	56.4	59.4	63.0	66.1	67.5	69.1	70.5	71.7	77.0
	SC2	33.7	45.0	50.1	54.9	57.0	59.0	61.8	63.4	64.4	64.6	71.3
	SC3	39.6	54.3	63.2	69.5	73.9	75.4	78.2	79.4	80.6	82.4	89.3
	RS	3.8	5.9	5.9	5.9	5.9	5.9	5.9	5.9	6.1	6.1	6.1
	Visual	34.7	47.4	56.9	61.9	66.9	69.8	72.8	76.4	77.0	78.4	91.9

in SC3 (ranked with PMIs) would capture the actual incorrect characters 73.9% of the time. As we explained in Section 5.2, the AIRs must be smaller or equal to the inclusion rates of the individual experiments. In Table XIII, we copy the inclusion rates of the Jlist row in Table XI into the R_{num} column.

The statistics listed in Table XIII show the effectiveness of our ranking mechanisms – both ENOPs and also PMIs. The difference (R_{num}-R_i) is a good indicator of the degree of sacrifice required in the situation when we recommend only the top i candidate rather than the complete candidate lists. When we shortened the candidate lists to contain less than 10 characters, we did not sacrifice the inclusion rates significantly. When we recommended 10 characters, the differences (R_{num}-R_{10}) were not large, especially when we considered that we would have to put forward much longer lists of candidate characters, for example, Table XII, to achieve R_{num}. One exception to this observation is that providing the complete candidate lists that were selected with the SS criterion may be worthwhile. According to Table XII, suggesting an average of 12.4 characters achieved R_{num}.

Recall that using the union of the candidate lists, such as Phone and Visual in Table XII, helped us to achieve higher inclusion rates. Although higher inclusion rates are desirable, the detailed statistics in the Phone and Visual sub-rows in Table XIII shed light on the drawbacks of the union lists. If we present only the top k candidates to those who need the similar characters of a given character, the union of the lists might not provide better performance profiles than that of the individual lists, separately.

It is not very difficult to understand the potentially inferior performances of the union lists. Assume that the rank of the actual incorrect character is j in a list, say LL. This implies that there already are at least $(j-1)$ characters that are mistakenly considered as better candidates by the score functions. After we put the lists together

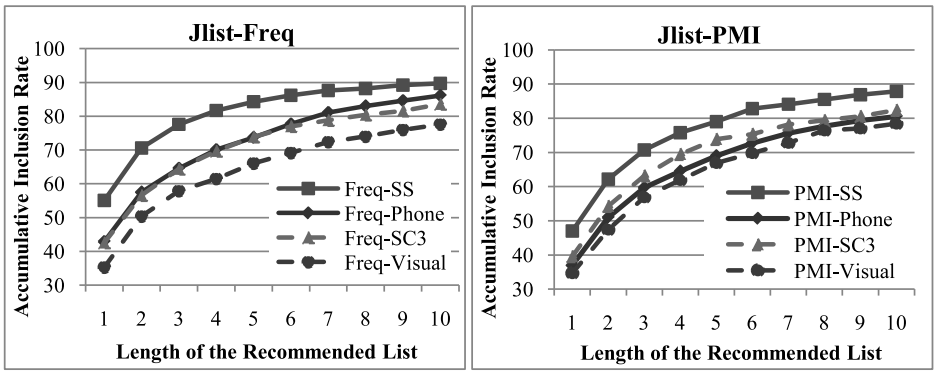

Fig. 3. AIRs of the union lists might not be as good as individual lists.

and rank the joined lists, these $(j-1)$ characters still win against the actual incorrect character. In addition, other characters that were not in LL might be ranked higher than the actual incorrect character in the union. As a consequence, the rank of the actual incorrect character might not improve in the union lists.

Consider, in one particular test, two ranked lists $SS = \{A, B\}$ and $SD = \{C, D\}$, where A, B, C, and D are four different characters. Hence, we must have score $(A) \geq$ score(B) and score$(C) \geq$ score (D). Assume that B is the actual incorrect character, that score $(C) \geq$ score(B), and that score$(B) \geq$ score(D). The union of SS and SD will be $\{A, C, B, D\}$. The rank of the actual incorrect character will drop from 2 in SS to 3 in the union. This is a situation in which the joined list might not outperform the best individual list.

However, it remains possible that the joined lists perform better than the individual ones. This could happen when the actual incorrect characters were included in only one of the individual lists and when the ranks of the actual incorrect characters remain the same in the joined lists. If, in the previous example, score(B) is larger than score(C), than the joined lists will perform as well as SS. Moreover, if, in another test, we have $SS = \{E, F\}$, $SD = \{G, H\}$, where E, F, G, and H are different characters, and where G is the actual incorrect character, than the joined list will perform better than SS.

Given the reasons provided in the previous two paragraphs, we cannot tell whether or not the joined lists will perform better than the best performing individual lists.

Figure 3 shows four pairs of examples for the experiments with Jlist. It happened that the joined lists did not perform as well as the best-performing individual lists in the joined lists. The charts were drawn based on the statistics listed in Table XIII. For instance, the curve "Freq-SS" in the chart with the title "Jlist-Freq" was based on the data in the sub-row "SS" in the "Frequency" part in Table XIII. The performance profiles of SS lists dominated those of Phone lists, and the performance profiles of SC3 lists dominated those of Visual lists in these cases. When we ranked the candidate characters with PMIs, the results were similar, and are shown in the chart titled "Jlist-PMI."

It is interesting to explore whether we may improve the performance of the Phone list by not considering the characters that were in the MS and MD lists. We conducted such an experiment, and in the middle of Table XIII, the row SS + SD shows the AIRs of the union list of words that were formed by using the candidate characters originally in the SS and SD lists. We can compare the performances of the Phone list and the SS + SD list. Overall, the SS + SD list provides better, but not significantly better, performance.

Table XIV. AIRs for Ranking the Candidates for Elist Based on ENOPs and PMIs

		R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_{10}
Frequency	SS	55.2	72.7	81.9	85.2	87.5	88.9	89.7	90.0	90.5	90.7
	SD	11.3	14.4	15.6	16.2	16.8	17.2	17.6	17.7	17.8	17.9
	MS	2.0	2.2	2.6	2.7	2.9	2.9	2.9	2.9	3.0	3.0
	MD	1.3	1.6	1.6	1.6	1.6	1.6	1.8	1.8	1.8	1.8
	Phone	39.7	57.1	68.0	72.8	77.1	79.9	82.5	84.6	86.0	87.1
	SS+SD	46.7	62.8	73.5	78.9	83.7	86.2	88.6	89.9	90.6	92.0
	SC1	32.4	46.8	55.7	62.0	65.1	67.7	69.4	71.0	72.7	73.1
	SC2	27.7	41.6	49.2	55.3	59.2	62.6	64.9	66.8	67.9	69.5
	SC3	33.6	49.3	57.6	63.2	68.4	71.4	74.4	77.7	79.2	81.5
	RS	2.8	3.6	3.6	3.7	3.8	4.0	4.0	4.1	4.1	4.1
	Visual	27.7	41.7	49.1	55.1	59.8	63.7	66.7	69.3	71.7	74.1
PMI	SS	51.8	73.9	82.2	85.6	88.1	89.0	89.6	89.7	90.2	90.5
	SD	10.5	14.5	16.1	16.9	17.3	17.3	17.4	17.6	17.8	17.8
	MS	1.7	2.1	2.6	2.6	2.6	2.7	2.8	2.8	2.9	2.9
	MD	1.1	1.6	1.7	1.7	1.7	1.8	1.8	1.8	1.9	1.9
	Phone	40.5	61.6	72.5	77.7	81.6	84.5	86.6	88.2	89.6	91.0
	SC1	35.5	50.5	59.1	63.7	67.7	69.8	71.0	72.2	73.1	74.4
	SC2	32.5	47.1	53.8	58.6	62.5	66.1	67.9	69.6	70.4	71.3
	SC3	36.4	53.2	64.1	69.4	74.9	77.5	79.7	80.7	82.0	82.9
	RS	2.6	3.7	3.8	4.0	4.1	4.1	4.1	4.1	4.1	4.1
	Visual	30.5	47.1	56.5	62.6	67.3	70.7	72.9	75.2	76.8	78.7

In addition to ranking the candidate characters directly with their ENOPs, we also ranked the characters with their PMIs, shown in Equation (2) and Equation (4) in Section 5.1, and repeated the experiments with Elist and Jlist. The lower part of Table XIII shows the observed statistics. Qualitatively, the statistics in the upper and the lower parts of Table XIII do not show different trends: SS and SC3 remain to be the most effective methods to find phonologically and visually similar errors. Overall, finding phonologically similar errors is easier than finding visually similar errors. Providing candidate lists that had only 10 characters achieved reasonable performances.

The most noticeable difference between the upper and the lower part of Table XIII is in the R_1 column. It appears that, if we would recommend only one candidate character, using the raw values of ENOPs to rank will offer better inclusion rates than using PMI. Although this observation appears to be appropriate for the statistics in Table XIII that we collected from the experiments that used Jlist, this trend did not survive in our experiments with Elist.

Table XIV shows exactly the same sets of statistics as those shown in Table XIII. The only difference is that we used Elist, rather than Jlist, to repeat all of the experiments that we used to obtain Table XIII. Statistics in Table XIV indicate the same trends as those suggested by the most of statistics in Table XIII, so we do not repeat the same statements.

A major contribution of the statistics in Table XIV appears in Figure 4, which shows that using PMIs or ENOPs did not guarantee that there would be differences in the performances. We drew the left chart based on the data in Table XIII and the right chart based on data in Table XIV. The curves in the left chart show that using PMIs offered inferior performance than using the raw values of ENOPs, and the curves in the right chart show the opposite trend.

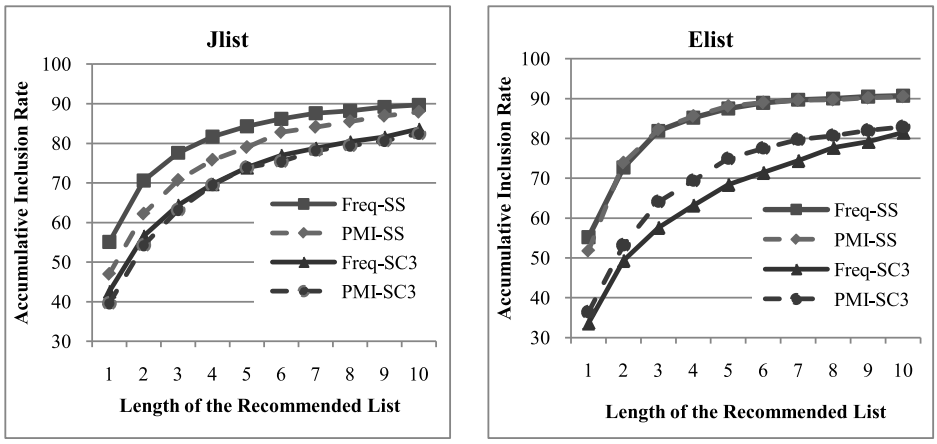

Fig. 4. Using PMI does not necessarily outperform using ENOPs.

Table XV. Ranking the Candidates Based on ENOPs and PMIs for Ilist

	Frequencies					PMIs				
	R_1	R_2	R_3	R_4	R_5	R_1	R_2	R_3	R_4	R_5
SS	70.3	77.7	80.6	81.0	81.6	66.7	76.0	80.2	81.0	81.6
SD	25.6	28.3	28.9	28.9	29.3	24.8	28.1	28.9	29.1	29.3
MS	1.4	1.7	1.7	1.7	1.7	1.6	2.1	2.1	2.1	2.1
MD	1.6	1.6	1.6	1.6	1.6	1.2	1.6	1.6	1.6	1.6
Phone	76.7	89.1	93.6	94.8	95.5	70.5	86.2	92.1	94.4	95.7
SC1	64.7	72.0	76.0	78.0	78.3	61.7	72.3	75.0	76.7	77.7
SC2	58.0	64.7	68.3	70.7	71.0	53.7	66.3	69.3	70.3	70.3
SC3	71.3	80.0	85.0	86.7	87.0	67.0	80.0	84.7	86.0	86.3
RS	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3	1.3
Visual	71.0	82.3	86.3	89.3	89.3	65.0	81.7	86.7	88.3	89.0

Although PMIs are frequently used to compute the co-occurrences of two events, including the collocation of words, examining Formula (4) discussed in Section 5.1 reveals an intuitive interpretation of the PMIs in our applications. The formula measures the percentages of observing the incorrect words (i.e., $C \wedge X$) given that the candidate characters appeared (i.e., X in the formula; recall that this is the character that would replace the correct character). Such percentages can be diluted if X happens to represent a high frequent character.

Similar to an experiment that we conducted for the Jlist, we also created the union list SS + SD for the experiments with Elist, and the results are shown in the middle of Table XIV. This time, the SS + SD list outperformed the Phone list by a margin by about 5%. However, the resulting performance profile of SS + SD is still inferior to that of SS. Interestingly, when we consider the top 10 candidate characters, the SS + SD outperformed not only the Phone but also the SS list marginally.

5.6 Ranking Tests with Ilist

Table XV shows the accumulative inclusion rates (AIRs) for the experiments for Ilist, which provides reported errors for simplified Chinese. The inclusion rates for the experiments for Ilist were presented in Table XI.

Table XVI. AIRs for Ranking Candidates for Wlist Based on Frequencies and PMIs

		R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_{10}
Frequency	SS	73.1	84.6	90.8	92.3	93.1	93.8	93.8	93.8	93.8	93.8
	SD	18.5	21.5	22.3	23.1	23.1	23.1	23.1	23.1	23.1	23.1
	MS	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
	MD	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
	Phone	68.5	82.3	88.5	91.5	93.1	93.1	95.4	95.4	96.2	96.2
	SC1	57.3	70.9	74.8	77.7	78.6	78.6	78.6	79.6	79.6	79.6
	SC2	54.4	69.9	72.8	73.8	74.8	76.7	76.7	77.7	77.7	78.6
	SC3	60.2	76.7	82.5	84.5	85.4	86.4	89.3	89.3	89.3	89.3
	RS	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	Visual	52.4	67.0	77.7	81.6	85.4	86.4	86.4	86.4	89.3	89.3
PMI	SS	63.8	83.1	90.0	91.5	92.3	92.3	92.3	92.3	92.3	92.3
	SD	18.5	21.5	22.3	22.3	23.1	23.1	23.1	23.1	23.1	23.1
	MS	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
	MD	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
	Phone	60.8	82.3	90.8	92.3	94.6	96.2	96.2	96.2	96.2	96.9
	SC1	51.5	64.1	71.8	73.8	77.7	78.6	79.6	79.6	79.6	79.6
	SC2	49.5	65.0	71.8	72.8	75.7	75.7	77.7	77.7	77.7	77.7
	SC3	55.3	73.8	80.6	85.4	88.3	89.3	89.3	89.3	89.3	89.3
	RS	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	Visual	52.4	67.0	71.8	77.7	84.5	86.4	86.4	87.4	88.3	88.3

We use a format that is similar to the format for Tables XIII and XIV for Table XV, but we list only the AIRs for the top-five candidates. When we used the top-five candidate characters, the AIRs were almost as good as the inclusion rates, which we listed in Table XI. Statistics in Table XV indicate that our system performed well for the simplified Chinese. We can find results similar to those that we presented for the experimental results for Elist and Jlist. The candidate lists selected with the SS and SC3 criteria were the most effective in capturing phonologically and visually related errors. SD continued to serve as an instrumental complement for SS.

In contrast to what we observed in the experiments for Elist and Jlist, the Phone lists, which are the unions of SS, SD, MS, and MD lists, performed better than the best performing individual lists, that is, SS lists. The top-five candidate characters in the Phone lists captured 95.5% of the phonologically related errors on average. Analogously, the Visual list, which is the union of SC1, SC2, SC3, and RS, captured 89.3% of the visually related errors.

5.7 Further Tests with Wlist and Blist

We explained, in Section 4.2, that we used Wlist and Blist as the new datasets to test how our system will perform with unforeseen data.

Table XVI and Table XVII provided in the Appendix show that the experimental results for Wlist and Blist were not different from the results for Elist and Jlist, which we discussed in Section 5.5. The inclusion rates were good, as we discussed in Section 5.4. Using the top 10 candidate characters enabled us to catch most of the errors that we were able to capture with the complete lists. Using PMI and ENOPs to rank the candidate characters achieved performance profiles of similar quality. SS lists and SC3 lists performed best if we had to use only one of the lists to capture the phonologically and the visually related errors, respectively. In addition, SD lists complemented the SS lists to find those phonologically related errors.

Table XVII. AIRs for Ranking the Candidates for Blist Based on Frequencies and PMIs

		R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_{10}
Frequency	SS	69.1	78.3	80.3	80.9	81.5	81.8	81.8	81.8	81.8	81.8
	SD	20.1	22.0	22.9	23.9	23.9	23.9	23.9	23.9	23.9	23.9
	MS	3.2	3.2	3.2	3.2	3.2	3.2	3.2	3.2	3.2	3.2
	MD	1.6	1.9	1.9	1.9	1.9	1.9	1.9	1.9	1.9	1.9
	Phone	71.3	82.5	86.3	89.2	90.4	92.4	93.0	93.9	93.9	93.9
	SC1	64.2	71.6	76.1	76.9	77.6	77.6	77.6	77.6	77.6	77.6
	SC2	59.7	67.2	70.1	72.4	73.9	74.6	74.6	74.6	74.6	74.6
	SC3	75.4	85.1	87.3	88.1	88.8	88.8	89.6	89.6	90.3	90.3
	RS	3.0	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7
	Visual	71.6	79.9	83.6	86.6	88.1	89.6	89.6	89.6	89.6	89.6
PMI	SS	64.3	75.8	77.4	80.3	80.9	81.2	81.5	81.5	81.5	81.8
	SD	19.7	22.3	22.6	23.9	23.9	23.9	23.9	23.9	23.9	23.9
	MS	3.2	3.2	3.2	3.2	3.2	3.2	3.2	3.2	3.2	3.2
	MD	1.6	1.6	1.6	1.6	1.6	1.9	1.9	1.9	1.9	1.9
	Phone	67.5	83.4	86.9	89.5	90.8	92.4	93.0	93.0	93.0	93.0
	SC1	63.4	73.1	76.9	77.6	77.6	77.6	77.6	77.6	77.6	77.6
	SC2	61.9	70.9	73.9	73.9	74.6	74.6	74.6	74.6	74.6	74.6
	SC3	75.4	83.6	87.3	89.6	89.6	89.6	89.6	89.6	89.6	89.6
	RS	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7	3.7
	Visual	74.6	82.8	87.3	89.6	90.3	91.0	91.0	91.0	91.0	91.0

6. APPLICATIONS

With the capability to capture the actual errors that occurred while people typed and wrote Chinese, we can apply our techniques to computer assisted language learning and to the related fields that we mentioned in Section 1.

The most obvious application is to help teachers prepare test items for "Incorrect Character Correction" tests (ICC tests). In such tests, students have to find and correct an incorrect Chinese character in a given sentence, for example, "小明昨天參加了校外施行" (Ming took part in the field trip yesterday. Xiao3 ming2 zuo2 tian1 can1 jia1 le1 xiao4 wai4 shi1 xing1). In this Chinese string, "施" (shi1) is incorrect and should be changed to "旅" (lu3) to make the statement correct. This is a very common type of test in the assessment of Chinese language proficiency.

To prepare such test items, there may be some characters which teachers wish to check if their students recognize in their correct form or not, that is, "旅" in "旅行" (lu3 xing2) in the previous example. When preparing the test items, teachers will have to figure out what might be the most appropriate incorrect character to use to stand in for the correct character. Depending on the level of difficulty and the purpose of the test, they may prefer incorrect characters that are visually similar or are phonologically similar to the correct character. For phonologically similar characters, teachers may prefer to select incorrect characters that were recommended with the SS, SD, MS, and MD criteria. From this viewpoint, we should present candidate characters by their categories. The union lists are not the best choice.

Figure 5 shows a snapshot of the user interface of our prototype[16] that aims to help teachers prepare test items for ICC tests. In this example, a teacher requested errors that were visually similar ("形體相似" in the figure; xing2 ti3 xiang1 si4) and errors

[16]See http://140.119.164.139/biansz/bianszindex_v2.php. This is our own service and it is always open, except when we experience power outage problems.

Fig. 5. The interface of our prototype for assisting teachers to prepare test items in the "Incorrect Character Correction" tests.

that had the same sound and same tone ("同音同調" in the figure; tong2 yin1 tong2 diao4), and our system returned only the top three candidates. This is a conservative design, given that we are able to capture a large percentage of the actual incorrect characters of previously observed errors with the top 10 candidates (Sections 5.5 to 5.7).

This authoring tool can be evaluated by how often the recommended characters are adopted by teachers in their test items. This style of evaluation is the same as what we have accomplished in Section 5.4 through Section 5.7. The correct words in the lists, in Table VI, serve as the target test items, and the actual incorrect characters are the teachers' choices. From this viewpoint, we have conducted an evaluation with more than 4,100 individual tests. The observed inclusion rates, which were presented in Table XI, show that our system was able to offer candidate characters that included the teachers' choices. The accumulate inclusion rates, which were presented from Table XIII to Table XVII, further indicate that, by providing no more than 10 candidate characters in different categories of similar characters, our system maintained its efficacy for assisting the compilation of test items for ICC tests.

In addition to assisting the preparation of test items for Chinese tests, we can employ the lists of similar characters to automatic detection of errors in Chinese text (e.g., Zhang et al. [2000]). The statistics discussed in Section 4.3 show that a large portion of errors in Chinese texts are related to characters that have the same or similar pronunciations, and a previous work applied phonetic information for this error detection task based on related arguments [Huang et al. 2008]. Using both visually and phonetically similar characters along with statistical methods, we significantly improve the performances of Huang et al.'s system and two other systems that were reported in the literature [Wu et al. 2010].

We plan to offer a free and open Web-based service to the research community. The service will allow users to enter queries to search Chinese characters that meet certain conditions. With a minor change of the interface shown in Figure 5, we can offer psycholinguistic researchers the neighbor words (e.g., Lo and Hue [2008], Tsai et al. [2006]) of a given Chinese word for their studies. In fact, we are applying our system to support the design of educational games for cognition-based learning of Chinese characters (cf., Lee et al. [2010]). Moreover, our work can be used to find the suggested queries when users of search engines enter incorrect words [Croft et al. 2010, p. 197]. Although we are not experts in the recognition of Chinese characters either in printed

(i.e., OCR) or in written form, we can help researchers to find the confusion sets for Chinese characters [Fan et al. 1995] more efficiently.

7. DISCUSSIONS

The experimental results presented in Section 5 showed that it is relatively easy to capture errors that are related to phonological similarity. It is relatively harder to catch errors that are related to visual similarity.

Using the information about the pronunciations of characters that are available in Chinese lexicons is very effective for reproducing phonologically-related errors. Selecting candidate characters with the SS and SD criteria was most fruitful. The main reason that our program did not catch the errors that were related to phonological similarity was that our lists of confusing phonemes (cf. Table I) did not contain the types of errors that actually occurred. This can happen if the types of errors are those which are not considered significant in psycholinguistic studies, but which can occur once in a while in reality.

Using the extended Cangjie codes proved to be the main reason why we could capture a larger portion of the errors that are related to visual similarity, when we compare the performances of our systems that were implemented in 2007 and 2008. The decision to divide characters into subareas further improves our ability to find similar characters. However, the steps demand subjective decisions, in which we observed how characters were divided [Lee 2010a], and these decisions will influence how well we find the incorrect characters. We discussed an example of the problem, that is, the "弓人一" (gong1 ren2 yi1) problem, in Section 3.4. Another example is the question of how our programs may find the similarity between "副" (fu4) and "福" (fu2). According to Lee [2010a], the LIDs for "副" and "福" are 4 and 3, respectively (cf., Section 3.3). Hence, the shared component "畐" (fu2) of these two characters will be saved in two different ways: "一口田" (yi1 kou2 tian1) at P1 for "副"; and "一口" and "田" at P1 and P2, respectively, for "福".

To alleviate the problem, we concatenated the substrings of Cangjie codes into one string and computed the Dice coefficient (Equation (1) in Section 3.4) of the concatenated Cangjie codes for two characters. This strategy proved to be very important. Using SC3 to select visually similar characters outperformed SC2 and SC1 in all of our experiments.

Although we have achieved good experimental results, using Cangjie codes as the basis of defining the visual similarity between characters does not produce perfect results. The original Cangjie codes may not reflect the complexity, for example, the number of strokes, of a component in a character. A complex component can be represented with a simple Cangjie symbol, for example, the Cangjie code for "弗" (fu2) is "中中弓" (zhong1 zhong1 gong1). In contrast, a seemingly simple component can be represented with a longer sequence of Cangjie symbols, for example, the complete Cangjie code for "予" (yu3) is "弓戈弓弓" (gong1 ge1 gong1 gong1). This phenomenon may mislead our score functions, that is, SC1, SC2, and SC3, which rely on the lengths of the matched Cangjie codes to determine the degree of similarity between characters.

A possible solution to this problem is to use our own Cangjie codes for the basic elements, but this strategy has its problems. For instance, we replaced the Cangjie code for "言" (yan2) with "卜一一口" (bu3 yi1 yi1 kou3) in c_5, c_6, and c_7 in Table IV. However, such an operation is extremely subjective and labor intensive. Although we changed the Cangjie codes for a limited number of elements, we cannot guarantee that we have done enough for all possible errors that are related to visual similarity. Moreover, we were not sure whether we were just trying to maximize the performance of our systems in the case of some particular lists of errors. This was the main reason

that we collected Wlist and Blist for further experiments, after we had been using Elist and Jlist for an extended period of time. Fortunately, the experimental results for Wlist and Blist remained satisfactory.

Another problem that came up when we built the database of the extended Cangjie codes is the degree of detail to which we should recover the Cangjie codes. Consider this list of characters: "舞" (wu3), "列" (lie4), "例" (li4), "夥" (huo3), and "麥" (mai4). It is probably not easy for everyone to notice that they all share "夕" (xi4) somewhere inside them. To what degree do users pay attention to relatively small elements? Should we consider this factor when we design the score functions to measure the degree of similarity between two characters? This is a hard question for us. The best design may depend on the actual applications, for example, the needs of psycholinguistic experiments [Leck et al. 1995; Yeh and Li 2002].

So far, we have not touched upon the issue that the Cangjie codes do not provide a good mechanism for comparing the similarities between characters that consist of very few strokes. Examples are c_1(田, tian2), c_2(由, you2), c_3(甲, jia3), and c_4(申, shen1) in Table II. Another group of similar characters are "土" (tu3), "士" (shi4), "工" (gong1), "干" (gan1), and "千" (qian1). Differences among these characters are at the stroke level, so we cannot rely on the Cangjie codes to find their similarities. For such characters, the Wubihua encoding method [Wubihua 2010] should be applied. The Wubihua encoding method assigns identification numbers to a selected set of strokes, for example, "1" for horizontal strokes and "2" for vertical strokes. Because there is exactly one canonical way to write a Chinese character, that is, the standard order of the strokes that form the character, one can convert each of the strokes into their Wubihua digits, and use this sequence of digits to encode a Chinese character. The Wubihua codes for "土", "士", "工", "干", and "千" are, respectively, "121", "121", "121", "112" and "312"; and the Wubihua codes for "田", "由", "甲", and "申" are, respectively, "25121", "25121", "25112", and "25112". Demanding an exact match between strings as the selection criterion, we can find that "土", "士", and "工" are more similar to each other than to "干" and "千". By appropriately integrating the extended Cangjie codes and Wubihua codes, we will be able to extend our ability to find visually similar characters to a larger scope of characters.

For the study of incorrect Chinese characters, we have intentionally put aside an important class of errors at this moment. For written characters, people may write incorrect characters that look like correct characters, for example, writing "試" (shi4) as "試".[17] These so-called pseudo-characters obey the formation principles of Chinese characters, but, in fact, do not belong to the language. These incorrect characters were not considered in the current study because we could not normally enter them into our files as they were not contained in the font files. Nevertheless, studying this type of errors may uncover possible ways that people memorize Chinese characters, and opens another door to the mental lexicons of Chinese learners.

Song and his colleagues propose methods for automatic proofreading for simplified Chinese in Song et al. [2008]. They consider seven operators for building Chinese characters from their components and propose a set of rules for computing the similarity between Chinese characters. They then employ the similar characters with statistical information about language models to detect possible incorrect words. This line of work is very similar to the work presented in this article. It will be very interesting to compare the performances of Song et al.'s and our systems with some common test sets.

SJTUD [1988] provides not just a systematic way to decompose simplified Chinese characters, and it also lists the decompositions of 11,254 individual characters. It

[17]Reported in the *United Daily News* (http://www.udn.com.tw) on 19 May 2010.

will be very interesting to compare the effectiveness for computing visually similar characters with the extended Cangjie codes and the decompositions in SJTUD [1988].

8. CONCLUSIONS

We found methods to reproduce the errors found in the writing of Chinese script. The methods utilized information about the pronunciation of Chinese characters and the heuristics rules that were derived from observations in psycholinguistic studies to judge the degree of similarity between pronunciations. The methods also employed the extended Cangjie codes and score functions to determine the degree of visual similarity between characters.

We evaluated our approach from three aspects. In Section 5.3, we compared the Web-based statistics to show the reliability of Web-based statistics. In Section 5.4, we showed that our approach could capture the incorrect character for a diverse scope of test data at satisfactory rates. In Sections 5.5 through 5.7, we applied and compared two different methods to rank the candidate characters in an attempt to capture the incorrect character with shorter lists of candidate characters. The experiments were carried out with data that we presented in previous conference articles and some new data that covered both traditional and simplified Chinese.

In these experiments, it was found that 76% of these errors were related to phonological similarity and that 46% were related to visual similarity between characters. We showed that the Web-based statistics were reasonably stable when we compared the popularity of word usages by comparing the numbers of Web pages that contained the target words in both 2009 and 2010. Experimental results show that we were able to capture 97% of the 4,100 errors, when we recommended 104 candidate characters. When we recommended only 10 candidate characters, we still caught more than 80% of the 4100 errors. The reported techniques are useful for applications that are related to Chinese, and, in particular, we showed a real-world application that can help teachers to author test items for "incorrect character correction" tests for Chinese.

ACKNOWLEDGMENTS

We thank anonymous reviewers of this journal version and the previous conference articles for their invaluable comments, which strongly influenced this publication. Experiments were added and improved to respond to reviewers' comments, though we did not mark each of such experiments to indicate the reviewers' credits. We would also like to thank Professor Song Rou for his sharing his article with us. We would also like to thank Miss Moira Breen for her indispensable support for our English.

REFERENCES

CANGJIE. 2010. An introduction to the Cangjie input method. http://en.wikipedia.org/wiki/Cangjie_input_method.

CDL. 2010. Chinese document laboratory, Academia Sinica. http://cdp.sinica.edu.tw/cdphanzi/. (In Chinese)

CHEN, M. Y. 2000. *Tone Sandhi: Patterns Across Chinese Dialects* (Cambridge Studies in Linguistics 92). Cambridge University Press.

CHU, B.-F. 2010. *Handbook of the Fifth Generation of the Cangjie Input Method*. http://www.cbflabs.com/book/5cjbook/. (In Chinese)

CORMEN, T. H., LEISERSON, C. E., RIVEST, R. L., AND STEIN, C. 2009. *Introduction to Algorithms* 3rd Ed. MIT Press.

CROFT, W. B., METZLER, D., AND STROHMAN, T. 2010. *Search Engines: Information Retrieval in Practice*. Pearson.

DICT. 2010. An official source of information about traditional Chinese characters. http://www.cns11643.gov.tw/AIDB/welcome.do.

FAN, K.-C., LIN, C.-K., AND CHOU, K.-S. 1995. Confusion set recognition of online Chinese characters by artificial intelligence technique. *Patt. Recog. 28*, 3, 303–313.

FELDMAN, L. B. AND SIOK, W. W. T. 1999. Semantic radicals contribute to the visual identification of Chinese characters. *J. Mem. Lang. 40*, 4, 559–576.

FROMKIN, V., RODMAN, R., AND HYAMS, N. 2002. *An Introduction to Language* 7th Ed. Thomson.

HANDICT. 2010. A source for traditional and simplified Chinese characters. http://www.zdic.net/appendix/f19.htm.

HUANG, C.-M., WU, M.-C., AND CHANG C.-C. 2008. Error detection and correction based on Chinese phonemic alphabet in Chinese text. *Int. J. Uncertainty, Fuzziness Knowl.-Base. Syst. 16*, suppl. 1, 89–105.

JACKENDOFF, R. 1995. *Patterns in the Mind: Language and Human Nature*. Basic Books.

JUANG, D., WANG, J.-H., LAI, C.-Y., HSIEH, C.-C., CHIEN, L.-F., AND HO, J.-M. 2005. Resolving the unencoded character problem for Chinese digital libraries. In *Proceedings of the 5th ACM/IEEE Joint Conference on Digital Libraries (JCDL'05)*. 311–319.

JURAFSKY, D. AND MARTIN, J. H. 2009. *Speech and Language Processing* 2nd Ed. Pearson.

KUO, W.-J., YEN, T.-C., LEE, J.-R., CHEN, L.-F., LEE, P.-L., CHEN, S.-S., HO, L.-T., HUNG, D. L., TZENG, O. J.-L., AND HSIEH, J.-C. 2004. Orthographic and phonological processing of Chinese characters: An fMRI study. *NeuroImage 21*, 4, 1721–1731.

LECK, K.-J., WEEKES, B. S., AND CHEN, M.-J. 1995. Visual and phonological pathways to the lexicon: Evidence from Chinese readers. *Mem. Cogn. 23*, 4, 468–476.

LEE, C.-Y. 2009. The cognitive and neural basis for learning to reading Chinese. *J. Basic Educ. 18*, 2, 63–85.

LEE, C.-Y., HUANG, H.-W., KUO, W.-J., TSAI, J.-L., AND TZENG, O. J.-L. 2010. Cognitive and neural basis of the consistency and lexicality effects in reading Chinese. *J. Neurolinguist. 23*, 1, 10–27.

LEE, C.-Y., TSAI, J.-L., HUANG, H.-W., HUNG, D. L., AND TZENG, O. J.-L. 2006. The temporal signatures of semantic and phonological activations for Chinese sublexical processing: An event-related potential study. *Brain Res. 1121*, 1, 150–159.

LEE, H. 2010a. Cangjie Input Methods in 30 Days 2. Foruto. http://input.foruto.com/cccls/cjzd.html. (In Chinese)

LEE, MU. 2010b. A quantitative study of the formation of Chinese characters. http://chinese.exponode.com/0_1.htm. (In Chinese)

LIU, C.-L., LAI, M.-H., CHUANG, Y.-H., AND LEE, C.-Y. 2010. Visually and phonologically similar characters in incorrect simplified Chinese words. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*. 739–747.

LIU, C.-L., LEE, C.-Y., TSAI, J.-L., AND LEE, C.-L. 2011. Forthcoming. A cognition-based interactive game platform for learning Chinese characters. In *Proceedings of the 26th ACM Symposium on Applied Computing (SAC'11)*.

LIU, C.-L. AND LIN, J.-H. 2008. Using structural information for identifying similar Chinese characters. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL'08)*. 93–96.

LIU, C.-L., TIEN, K.-W., CHUANG, Y.-H., HUANG, C.-B., AND WENG, J.-Y. 2009a. Two applications of lexical information to computer-assisted item authoring for elementary Chinese. In *Proceedings of the 22nd International Conference on Industrial Engineering & Other Applications of Applied Intelligent Systems (IEA/AIE'09)*. 470–480.

LIU, C.-L., TIEN, K.-W., LAI, M.-H., CHUANG, Y.-H., AND WU, S.-H. 2009b. Capturing errors in written Chinese words. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL'09)*. 25–28.

LIU, C.-L., TIEN, K.-W., LAI, M.-H., CHUANG, Y.-H., AND WU, S.-H. 2009c. Phonological and logographic influences on errors in written Chinese words. In *Proceedings of the 7th Workshop on Asian Language Resources (ALR'09)*. 84–91.

LIU, C.-L., JAEGER, S., AND NAKAGAWA, M. 2004. Online recognition of Chinese characters: The state-of-the-art. *IEEE Trans. Patt. Anal. Mach. Intel. 26*, 2, 198–213.

LO, M. AND HUE, C.-W. 2008. C-CAT: A computer software used to analyze and select Chinese characters and character components for psychological research. *Behav. Res. Meth. 40*, 4, 1098–1105.

MA, W.-Y. AND CHEN, K.-J. 2003. Introduction to CKIP Chinese word segmentation system for the first international Chinese word segmentation bakeoff. In *Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing (SIGHAN'03)*. 168–171.

MANNING, C. D. AND SCHÜTZE, H. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press.

MANNING, C. D., RAGHAVAN, P., AND SCHÜTZE, H. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

MOE 1996. Common errors in Chinese writings (常用國字辨似). Ministry of Education, Taiwan. http://140.111.34.54/files/site_content/M0001/biansz/c9.htm.

SISON, R. AND SHIMURA, M. 1998. Student modeling and machine learning. *Int. J. Artif. Intell. Educ. 9*, 1, 128–158.

SJTUD. 1988. Chinese character code workgroup of Shanghai Jiao Tong University. In *A Dictionary of Chinese Character Information*. Beijing Science Press (in Chinese).

SONG, R., LIN, M., AND GE, S.-L. 2008. Similarity calculation of Chinese character glyph and its application in computer aided proofreading system. *J. Chin. Comput. Syst. 29*, 10, 1964–1968. In Chinese.

SUN, X., CHEN, H., YANG, L., AND TANG, Y. Y. 2002. Mathematical representation of a Chinese character and its applications. *Int. J. Patt. Recog. Artif. Intell. 16*, 6, 735–747.

TSAI, J.-L., LEE, C.-Y., LIN, Y.-C., TZENG, O. J.-L., AND HUNG, D. L. 2006. Neighborhood size effects of Chinese words in lexical decision and reading. *Lang. Linguist. 7*, 3, 659–675.

TSAY, Y.-C. AND TSAY, C.-C. 2003. *Diagnoses of Incorrect Chinese Usage* (In Chinese). Firefly Publisher.

UNICODE. 2010. Unicode Standard 4.0.1, Chapter 11. http://www.unicode.org/versions/Unicode4.0.0/ch11.pdf.

VIRVOU, M., MARAS, D., AND TSIRIGA, V. 2000. Student modelling in an intelligent tutoring for the passive voice of English language. *Educ. Technol. Soc. 3*, 4, 139–150.

WU, S.-H., CHEN, Y.-Z., YANG, P.-C., KU, T., AND LIU, C.-L. 2010. Reducing the false alarm rate of Chinese character error detection and correction. In *Proceedings of the 1st Joint Conference on Chinese Language Processing (SIGHAN'10)*. 54–61.

WUBIHUA. 2010. An input method used in mainland China. http://en.wikipedia.org/wiki/Wubihua_method.

YEH, S.-L. AND LI, J.-L. 2002. Role of structure and component in judgments of visual similarity of Chinese characters. *J. Experi. Psych. Hum. Percept. Perform. 28*, 4, 933–947.

ZHANG, L., ZHOU, M., HUANG, C., AND PAN, H. 2000. Automatic detecting/correcting errors in Chinese text by an approximate word-matching algorithm. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'00)*. 248–254.

ZIEGLER, J. C., TAN, L. H., PERRY, C., AND MONTANT, M. 2000. Phonology matters: The phonological frequency effect in written Chinese. *Psychol. Sci. 11*, 3, 234–238.

Received September 2010; revised December 2010; accepted January 2011