Visually and Phonologically Similar Characters in Incorrect Simplified Chinese Words

Chao-Lin Liu[†] Min-Hua Lai[‡] Yi-Hsuan Chuang[↑] Chia-Ying Lee[↕]
[†‡↑]Department of Computer Science; [†↕]Center for Mind, Brain, and Learning
National Chengchi University
[↕]Institute of Linguistics, Academia Sinica
{[†]chaolin, [‡]g9523, [↑]g9804}@cs.nccu.edu.tw, [↕]chiaying@gate.sinica.edu.tw

Abstract

Visually and phonologically similar characters are major contributing factors for errors in Chinese text. By defining appropriate similarity measures that consider extended Cangjie codes, we can identify visually similar characters within a fraction of a second. Relying on the pronunciation information noted for individual characters in Chinese lexicons, we can compute a list of characters that are phonologically similar to a given character. We collected 621 incorrect Chinese words reported on the Internet, and analyzed the causes of these errors. 83% of these errors were related to phonological similarity, and 48% of them were related to visual similarity between the involved characters. Generating the lists of phonologically and visually similar characters, our programs were able to contain more than 90% of the incorrect characters in the reported errors.

1 Introduction

In this paper, we report the experience of our studying the errors in simplified Chinese words. Chinese words consist of individual characters. Some words contain just one character, but most words comprise two or more characters. For instance, "卖" (mai4)[1] has just one character, and "语言" (yu3 yan2) is formed by two characters. Two most common causes for writing or typing incorrect Chinese words are due to visual and phonological similarity between the correct and the incorrect characters. For instance, one might use "划" (hwa2) in the place of "画"(hwa4) in "刻画形象" (ke1 hwa4 xing2 xiang4) partially because of phonological similarity; one might replace "拙" (zhuo2) in "心劳力拙" (xin1 lao2 li4 zhuo2) with "绌" (chu4) partially due to visual similarity. (We do not claim that the visual or phonological similarity alone can explain the observed errors.)

Similar characters are important for understanding the errors in both traditional and simplified Chinese. Liu et al. (2009a-c) applied techniques for manipulating correctness of Chinese words to computer assisted test-item generation. Research in psycholinguistics has shown that the number of neighbor characters influences the timing of activating the mental lexicon during the process of understanding Chinese text (Kuo et al. 2004; Lee et al. 2006). Having a way to compute and find similar characters will facilitate the process of finding neighbor words, so can be instrumental for related studies in psycholinguistics. Algorithms for optical character recognition for Chinese and for recognizing written Chinese try to guess the input characters based on sets of confusing sets (Fan et al. 1995; Liu et al., 2004). The confusing sets happen to be hand-crafted clusters of visually similar characters.

It is relatively easy to judge whether two characters have similar pronunciations based on their records in a given Chinese lexicon. We will discuss more related issues shortly.

To determine whether two characters are visually similar is not as easy. Image processing techniques may be useful but is not perfectly feasible, given that there are more than fifty thousand Chinese characters (HanDict, 2010) and that many of them are similar to each other in special ways. Liu et al. (2008) extend the Cangjie codes (Cangjie, 2010; Chu, 2010) to encode the layouts and details about traditional

[1] We show simplified Chinese characters followed by their Hanyu pinyin. The digit that follows the symbols for the sound is the tone for the character.

Coling 2008: Poster Volume, pages 739–747,
Beijing, August 2010

Chinese characters for computing visually similar characters. Evidence observed in psycholinguistic studies offers a cognition-based support for the design of Liu et al.'s approach (Yeh and Li, 2002). In addition, the proposed method proves to be effective in capturing incorrect traditional Chinese words (Liu et al., 2009a-c).

In this paper, we work on the errors in simplified Chinese words by extending the Cangjie codes for simplified Chinese. We obtain two lists of incorrect words that were reported on the Internet, analyze the major reasons that contribute to the observed errors, and evaluate how the new Cangjie codes help us spot the incorrect characters. Results of our analysis show that phonological and visual similarities contribute similar portions of errors in simplified and traditional Chinese. Experimental results also show that, we can catch more than 90% of the reported errors.

We go over some issues about phonological similarity in Section 2, elaborate how we extend and apply Cangjie codes for simplified Chinese in Section 3, present details about our experiments and observations in Section 4, and discuss some technical issues in Section 5.

2 Phonologically Similar Characters

The pronunciation of a Chinese character involves a sound, which consists of the nucleus and an optional onset, and a tone. In Mandarin Chinese, there are four tones. (Some researchers include the fifth tone.)

In our work, we consider four categories of phonological similarity between two characters: same sound and same tone (**SS**), same sound and different tone (**SD**), similar sound and same tone (**MS**), and similar sound and different tone (**MD**).

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 similar sound should consider the language experience of an individual. One who live in the southern and one who live in the northern China may 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.

Some Chinese characters are heteronyms. 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.

One challenge in defining similarity between characters is that the pronunciations of a character can depend on its context. The most common example of tone sandhi in Chinese (Chen, 2000) is that the first third-tone character in words formed by two adjacent third-tone characters will be pronounced in the second tone. At present, we ignore the influences of context when determining whether two characters are phonologically similar.

Although we have confined our definition of phonological similarity to the context of the Mandarin Chinese, it is important to note the influence of sublanguages within the Chinese language family will affect the perception of phonological similarity. Sublanguages used in different areas in China, e.g., 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 may perceive phonological similarity in very different ways. The study in this direction is beyond the scope of the current study.

3 Visually Similar Characters

Figure 1 shows four groups of visually similar characters. Characters in group 1 and group 2 differ subtly at the stroke level. Characters in group 3 share the components on their right sides. The shared component of the characters in group 4 appears at different places within the characters.

Radicals are used in Chinese dictionaries to organize characters, so are useful for finding visually similar characters. The characters in group 1 and group 2 belong to the radicals "田" and "讠", respectively. Notice that, although the radical for group 2 is clear, the radical for group 1 is not obvious because "田" is not a standalone component.

However, the shared components might not be the radicals of characters. The shared components in groups 3 and 4 are not the radicals. In

Figure 1. Examples of visually similar characters

many cases, radicals are semantic components of Chinese characters. In groups 3 and 4, 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.

Hence, a mechanism other than just relying on information about characters in typical lexicons is necessary, and we will use the extended Cangjie codes for finding visually similar characters.

3.1 Cangjie Codes for Simplified Chinese

Table 1 shows the Cangjie codes for the 13 characters listed in Figure 1 and 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, e.g. "田" and "中" collocate with "W" and "L", respectively. Information about the complete correspondence is available on the Wikipedia[2].

Using the Cangjie codes saves us from using 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. It is possible to find that the characters share a common component, based on the shared substrings of the Cangjie codes, i.e., "戈女". Using the common substring (shown in black bold) of the Cangjie codes, we may also find the shared component "勾" for characters in group 3 (c_{10}, c_{11}, and c_{12}), the shared component "员" in c_{13} and c_{14}, the shared component "力" in c_{15} and c_{16}, and the shared component "⻥" in c_{16} and c_{17}.

Despite the perceivable advantages, these original Cangjie codes are not good enough. In order to maintain efficiency in inputting Chinese characters, the Cangjie codes have been limited to no more than five keys. Thus, users of the Cangjie input method must familiarize themselves with the principles for simplifying the Cangjie codes. While the simplified codes help the input efficiency, they also introduce difficulties and ambiguities when we compare the Cangjie codes for computing similar characters. The prefix "弓一" in c_{16} and c_{17} can represent "⻥", "鱼" (e.g., c_8), and "马" (e.g., c_9). Characters whose Cangjie codes include "弓一" may contain any of these three components, but they do not really look alike.

ID	CC	Cangjie	ID	CC	Cangjie
1	田	田	10	购	月人**心戈**
2	由	中田	11	沟	水**心戈**
3	甲	田中	12	构	木**心戈**
4	申	中田中	13	员	**口月人**
5	许	**戈女**人十	14	圆	田**口月人**
6	讦	**戈女**一十	15	勋	**口人大尸**
7	计	**戈女**十	16	劲	**弓一大尸**
8	鲳	**弓一**日日	17	颈	**弓一**一月人
9	驹	**弓一**心口	18	经	女一**弓人一**

Table 1. Examples of Cangjie codes

Therefore, we augment the original Cangjie codes by using the complete Cangjie codes and annotate each Chinese character with a layout identification that encodes the overall contours of the characters. This is how Liu and his colleagues (2008) did for the Cangjie codes for traditional Chinese characters, and we employ a similar exploration for the simplified Chinese.

3.2 Augmenting the Cangjie Codes

Figure 2 shows the twelve possible layouts that are considered for the Cangjie codes for simplified Chinese characters. Some of the layouts contain smaller areas, and the rectangles show a subarea within a character. The smaller areas are assigned IDs between one and three. Notice that, to maintain read-ability of the figures, not all IDs for subareas are shown in Figure 2. An example character is provided below each layout. From left to right and from top to bottom, each layout is assigned an identification number from 1 to 12. For example, the layout ID of "国" is 8. "国" has two parts, i.e., "口" and "玉".

Researchers have come up with other ways to

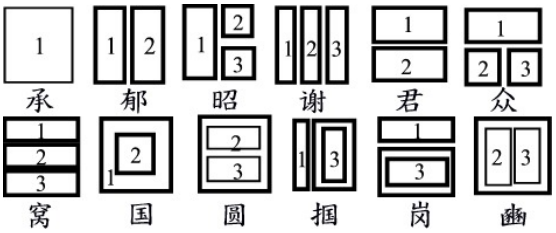

Figure 2. Layouts of Chinese characters

[2] en.wikipedia.org/wiki/Cangjie_input_method#Keyboard_layout ; last visited on 22 April 2010.

decompose individual Chinese characters. The Chinese Document Lab at the Academia Sinica proposed a system with 13 operators for describing the relationships among components in Chinese characters (CDL, 2010). Lee (2010b) propose more than 30 possible layouts.

The layout of a character affects how people perceive visual similarity between characters. For instance, c_{16} in Table 1 is more similar to c_{17} than to c_{18}, although they share "⿰". We rely on the expertise in Cangjie codes reported in (Lee, 2010a) to split the codes into parts.

Table 2 shows the extended codes for some characters listed in Table 1. The "ID" column provides links between the characters listed in both Table 1 and Table 2. The "CC" column shows the Chinese characters. The "LID" column shows the identifications for the layouts of the characters. The columns with headings "P1", "P2", and "P3" show the extended Cangjie codes, where "P*i*" shows the i^{th} part of the Cangjie codes, as indicated in Figure 2.

We decide the extended codes for the parts with the help of computer programs and subjective judgments. Starting from the original Cangjie codes, we can compute the most frequent substrings just like we can compute the frequencies of n-grams in corpora (cf. Jurafsky and Martin, 2009). Computing the most common 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 popular components in Chinese characters, e.g., "⿰" and "⿰". The original codes for "⿰" and "⿰" are "戈弓女" and "弓人一", but they are often simplified to "戈女" and "弓一", respectively. When simplified, "⿰" have the same Cangjie code with "戌", and "⿰" have the same Cangjie code with "马" and "鱼".

After finding the frequent substrings, we verify whether these frequent substrings are simplified codes for meaningful components. For meaningful components, we replace the simplified codes with complete codes. For instance the Cangjie codes for "许" and "讦" are extended to include "弓" in Table 2, where we indicate the extended keys that did not belong to the original Cangjie codes in boldface and with a surrounding box. Most of the non-meaningful frequent substrings have two keys: one is the last key of a part, and the other is the first key of another part. They were by observed by coincidence.

ID	CC	LID	P1	P2	P3
5	许	2	戈**弓**女	人十	
6	讦	2	戈**弓**女	一十	
7	计	2	戈**弓**女	十	
10	购	10	月人	心	戈
11	沟	10	水	心	戈
12	构	10	木	心	戈
13	员	5	口	月人	
14	圆	9	田	口	月人
15	勋	2	口**月**人	大尸	
16	劲	2	弓**人**一	大尸	
17	颈	2	弓**人**一	一月人	
18	经	3	女**女**一	弓人	一
19	恸	4	心	一**一**戈	**大**尸

Table 2. Examples of extended Cangjie codes

Although most of the examples provided in Table 2 indicate that we expand only the first part of the Cangjie codes, it is absolutely possible that the other parts, i.e., P2 and P3, may need to be extended too. c_{19} shows such an example.

Replacing simplified codes with complete codes not only help us avoid incorrect matches but also help us find matches that would be missed due to simplification of Cangjie codes. Using just the original Cangjie codes in Table 1, it is not easy to determine that c_{18} ("经") in Table 1 shares a component ("⿰") with c_{16} and c_{17} ("劲" and "颈"). In contrast, there is a chance to find the similarity with the extended Cangjie codes in Table 2, given that all of the three Cangjie codes include "弓人一".

We can see an application of the LIDs, using "劲", "颈" and "经" as an example. Consider the case that we want to determine which of "颈" and "经" is more similar to "劲". Their extended Cangjie codes will indicate that "颈" is the answer to this question for two reasons. First, "劲" and "颈" belong to the same type of layout; and, second, the shared components reside at the same area in "劲" and "颈".

3.3 Similarity Measures

The main differences between the original and the extended Cangjie codes are the degrees of details 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 similarity between characters.

In the current study, 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 had been tried by Liu and his colleagues' study for traditional Chinese characters (Liu et al., 2009b-c). The first method, denoted **SC1**, considers the total number of matched keys in the matched parts (without considering their part IDs). Let c_i denote the i^{th} character listed in Table 2. We have SC1(c_{15}, c_{16}) = 2 because of the matched "大尸". Analogously, we have SC1(c_{19}, c_{16}) = 2.

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_{15}, c_{16}) =SC1(c_{15}, c_{16})+1+1=4 because (1) the matched "大尸" locate at P2 in both characters and (2) c_{15} and c_{16} belong to the same layout. Assuming that c_{16} belongs to layout 5, than SC2(c_{15}, c_{16}) would become 3. In contrast, we have SC2(c_{19}, c_{16})=2. No extra weights for the matching "大尸" because it locates at different parts in the characters. The extra weight considers 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_{15} is more similar to c_{16} than to c_{19}, it also creates a new barrier in computing similarity scores. An example of this problem is that SC2(c_{17}, c_{18})=0. This is because that "弓人一" at P1 in c_{17} can match neither "弓人" at P2 nor "一" at P3 in c_{18}.

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's coefficient (cf. Croft et al., 2010) as the similarity. 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.

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

We compute another Dice's coefficient between X and Y. The formula is the similar to (1), except that we set Z to the longest common *consecutive* subsequence. We call this score $Dice_{LCCS}$. Notice that $Dice_{LCCS} \le Dice_{LCS}$, $Dice_{LCCS} \le 1$, and $Dice_{LCS} \le 1$. Finally, SC3 of two characters is the sum of their SC2, $10 \times Dice_{LCCS}$, and $5 \times Dice_{LCS}$. We multiply the Dice's coefficients with constants to make them as influential as the SC2 component in SC3. The constants were not scientifically chosen, but were selected heuristically.

4 Error Analysis and Evaluation

We evaluate the effectiveness of using the phonologically and visually similar characters to captures errors in simplified Chinese words with two lists of reported errors that were collected from the Internet.

4.1 Data Sources

We need two types of data for the experiments. The information about the pronunciation and structures of the Chinese characters help us generate lists of similar characters. We also need reported errors so that we can evaluate whether the similar characters catch the reported errors.

A lexicon that provides the pronunciation information about Chinese characters and a database that contains the extended Cangjie codes are necessary for our programs to generate lists of characters that are phonologically and visually similar to a given character.

It is not difficult to acquire lexicons that show 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 actually pronounce the characters. Hence, we can only rely on the standards that are recorded in lexicons.

With the procedure reported in Section 3.2, we built a database of extended Cangjie codes for the simplified Chinese. The database was designed to contain 5401 common characters in the BIG5 encoding, which was originally designed for the traditional Chinese. After converting the traditional Chinese characters to the simplified counterparts, the database contained only 5170

different characters.

We searched the Internet for reported errors that were collected in real-world scenarios, and obtained two lists of errors. The first list[3] came from the entrance examinations for senior high schools in China, and the second list[4] contained errors observed at senior high schools in China. We used 160 and 524 errors from the first and the second lists, respectively, and we refer to the combined list as the **Ilist**. An item of reported error contained two parts: the correct word and the mistaken character, both of which will be used in our experiments.

4.2 Preliminary Data Analysis

Since our programs can compare the similarity only between characters that are included in our lexicon, we have to exclude some reported errors from the Ilist. As a result, we used only 621 errors in this section.

Two native speakers subjectively classified 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 have five interesting categories: "P", "V", "N", "D", and "B" in Table 3. P and V indicate that the annotators agreed on the types of errors to be related to phonological and visual similarity, respectively. N indicates that the 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. B indicates the intersection of P and V.

Table 3 shows the percentages of errors in these categories. To get 100% from the table, we can add up P, V, N, and D, and subtract B from the total. In reality there are errors of type N, and Liu and his colleagues (2009b) reported this type of errors. Errors in this category happened to be missing in the Ilist. Based on our and Liu's observations, the percentages of phonological and visual similarities contribute to the errors in simplified and traditional Chinese words with similar percentages.

	P	V	N	D	B
Ilist	83.1	48.3	0	3.7	35.1

Table 3. Percentages of types of errors

[3] www.0668edu.com/soft/4/12/95/2008/20080913571140.htm ; last visited on 22 April 2010.

[4] gaozhong.kt5u.com/soft/2/38018.html; last visited on 22 April 2010.

4.3 Experimental Procedure

We design and employ the `ICCEval` procedure for the evaluation task.

At step 1, given the correct word and the correct character to be intentionally replaced with incorrect characters, we created a list of characters based on the selection criterion. We may choose to evaluate phonologically or visually similar characters. For a given character, `ICCEval` can generate characters that are in the SS, SD, MS, and MD categories for phonologically similar characters (cf. Section 2). For visually similar characters, `ICCEval` can select characters based on SC1, SC2, and SC3 (cf. Section 3.3). 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, for a correct word that people originally wanted to write, we replaced the correct character with an incorrect character with the characters that were generated at step 1, submitted the incorrect word to Google AJAX Search

```
Procedure ICCEval
Input:
  ccr: the correct character; cwd:
  the correct word; crit: the selec-
  tion criterion; num: number of re-
  quested characters; rnk: the cri-
  terion to rank the incorrect
  words;
Output: a list of ranked candidates
    for ccr
Steps:
1.  Generate a list, L, of charac-
    ters 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.  For each c in L, replace ccr in
    cwd with c, submit the resulting
    incorrect word to Google, and
    record the ENOP.
3.  Rank the list of incorrect words
    generated at step 2, using the
    criterion specified by rnk.
4.  Return the ranked list.
```

API, and extracted the estimated numbers of pages (ENOP) [5] 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`" on the upper part of the browser window. Using the AJAX API, we just have to parse the returned results with 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 ranked the similar characters based on their ENOPs at step 3, and return the list.

Since the reported errors contained information about the incorrect ways to write the correct words, we could check whether the real incorrect characters were among the similar characters that our programs generated at step 1 (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" as an example. In the collected data, it is reported that people wrote this word as "和靄可亲", i.e., the second character was incorrect. Hoping to capture the error, `ICCEval` generated a list of possible substitutions for "藹" at step 1. Depending on the 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 generate a list of alternatives for "藹", possibly including "靄", "谒", "葛", and other candidates. At step 2, we created and submitted 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, 410000, 26100, and 7940, 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" on top of the ranked list.

Notice that we considered the contexts in which the incorrect characters appeared to rank. We did not rank the incorrect characters with just the unigrams. In addition, although this running example shows that we ranked the characters directly with the ENOPs, we also ranked the list of alternatives with pointwise mutual information:

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

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 of replacing "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.) PMI is a common tool for judging collocations in natural language processing. (cf. Jurafsky and Martin, 2009).

It would demand very much computation effort to find Pr(C). Fortunately, we do not have to consider Pr(C) because it is a common denominator for all incorrect characters. Let X_1 and X_2 be two competing candidates for the correct character. We can ignore Pr(C) because of the following relationship.

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

Hence, X_1 prevails if $score(C,X_1)$ is larger.

$$score(C,X) = \frac{\Pr(C \wedge X)}{\Pr(X)} \quad (3)$$

In our work, we approximate the probabilities used in (3) by the corresponding frequencies that we can collect through Google, similar to the methods that we used to collect the ENOPs.

4.4 Experimental Results: Inclusion Tests

We ran `ICCEval` with 621 errors in the Ilist. The experiments were conducted for all categories of phonological and visual similarity. When using SS, SD, MS, MD, and RS as the selection criterion, we did not limit the number of candidate characters. When using SC1, SC2, and SC3 as the criterion, we limited the number candidates to be no more than 30. We consider only words that the native speakers have consensus over the causes of errors. Hence, we dropped those 3.7% of words in Table 3, and had just 598 errors. The ENOPs were obtained during March and April 2010.

Table 4 shows the chances that the lists, gen-

	SS	SD	MS	MD	Phone
Ilist	82.6	29.3	1.7	1.6	97.3
	SC1	SC2	SC3	RS	Visual
Ilist	78.3	71.0	87.7	1.3	90.0

Table 4. Chances of the recommended list contains the incorrect character

[5] According to (Croft et al., 2010), the ENOPs may not reflect the actual number of pages on the Internet.

erated with different `crit` at step 1, contained the incorrect character in the reported errors. In the Ilist, there were 516 and 300[6] errors that were related to phonological and visual similarity, respectively. Using the characters generated with the SS criterion, we captured 426 out of 516 phone-related errors, so we showed 426/516 = 82.6% in the table.

Results in Table 4 show that we captured phone-related errors more effectively than visually-similar errors. With a simple method, we can compute the union of the characters that were generated with the SS, SD, MS, and MD criteria. This integrated list suggested how well we captured the errors that were related to phones, and we show its effectiveness under "Phone". Similarly, we integrated the lists generated by SC1, SC2, SC3, and RS to explore the effectiveness of finding errors that are related to visual similarity, and the result is shown under "Visual".

4.5 Experimental Results: Ranking Tests

To put the generated characters into work, we wish to put the actual incorrect character high in the ranked list. This will help the efficiency in supporting computer assisted test-item writing. Having short lists that contain relatively more confusing characters may facilitate the data preparation for psycholinguistic studies.

At step 3, we ranked the candidate characters by forming incorrect words with other characters in the correct words as the context and submitted the words to Google for ENOPs. The results of ranking, shown in Table 5, indicate that we may just offer the leading five candidates to cover the actual incorrect characters in almost all cases.

The "Total" column shows the total number of errors that were captured by the selection criterion. The column "R*i*" shows the percentage of all errors, due to phonological or visual similarity, that were re-created and ranked i^{th} at step 3 in `ICCEVAL`. The row headings show the selection criteria that were used in the experiments. For instance, using SS as the criterion, 70.3% of actual phone-related errors were rank first, 7.4% of the phone-related errors were ranked second, etc. If we recommended only 5 leading incorrect characters only with SS, we would have captured the actual incorrect characters that were phone related 81.6% (the sum of R1 to R5) of the time. For errors that were related to visual similarity, recommending the top five candidates with SC3 would capture the actual incorrect characters 87.1% of the time. Since we do not show the complete distributions, the sums over the rows are not 100%. In the current experiments, the worst rank was 21.

	Total	R1	R2	R3	R4	R5
SS	426	70.3	7.4	2.9	0.4	0.6
SD	151	25.6	2.7	0.6	0.0	0.4
MS	9	1.4	0.4	0.0	0.0	0.0
MD	8	1.6	0.0	0.0	0.0	0.0
SC1	235	61.3	10.3	4.3	2.0	0.3
SC2	213	53.7	11.0	3.7	2.3	0.3
SC3	263	66.7	12.7	5.7	1.7	0.3
RS	4	1.3	0.0	0.0	0.0	0.0

Table 5. Ranking the candidates

We also used PMI to rank the incorrect words. Due to page limits, we cannot show complete details about the results. The observed distributions in ranks were not very different from those shown in Table 5.

5 Discussion

Compared with Liu et al.'s analysis (2009b-c) for the traditional Chinese, the proportions of errors related to phonological factors are almost the same, both at about 80%. The proportion of errors related to visual factors varied, but the averages in both studies were about 48%. A larger scale of study is needed for how traditional and simplified characters affect the distributions of errors. Results shown in Table 4 suggest that it is relatively easy to capture errors related to visual factors in simplified Chinese. Although we cannot elaborate, we note that Cangjie codes are not good for comparing characters that have few strokes, e.g., c_1 to c_4 in Table 1. In these cases, the coding method for Wubihua input method (Wubihua, 2010) should be applied.

Acknowledgement

This research was supported in part by the research contract NSC-97-2221-E-004-007-MY2 from the National Science Council of Taiwan. We thank the anonymous reviewers for constructive comments. Although we are not able to respond to all the comments

[6]The sum of 516 and 300 is larger than 598 because some of the characters are similar both phonologically and visually.

in this paper, we have done so in an extended version of this paper.

References

Cangjie. 2010. Last visited on 22 April 2010: en.wikipedia.org/wiki/Cangjie_input_method.

CDL. 2010. Chinese document laboratory, Academia Sinica. Last visited on 22 April, 2010; cdp.sinica.edu.tw/cdphanzi/. (in Chinese)

Chen, Matthew. Y. 2000. *Tone Sandhi: Patterns across Chinese Dialects*, (Cambridge Studies in Linguistics 92). Cambridge University Press.

Chu, Bong-Foo. 2010. *Handbook of the Fifth Generation of the Cangjie Input Method*. last visited on 22 April 2010: www.cbflabs.com/book/5cjbook/. (in Chinese)

Cormen, Thomas H., Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. 2009. *Introduction to Algorithms*, third edition. MIT Press.

Croft, W. Bruce, Donald Metzler, and Trevor Strohman, 2010. *Search Engines: Information Retrieval in Practice*, Pearson.

Dict. 2010. Last visited on 22 April 2010, www.cns11643.gov.tw/AIDB/welcome.do

Fan, Kuo-Chin, Chang-Keng Lin, and Kuo-Sen Chou. 1995. Confusion set recognition of on-line Chinese characters by artificial intelligence technique. *Pattern Recognition*, **28**(3):303−313.

HanDict. 2010. Last visit on 22 April 2010, www.zdic.net/appendix/f19.htm.

Jurafsky, Daniel and James H. Martin. 2009. *Speech and Language Processing*, second edition, Pearson.

Kuo, Wen-Jui, Tzu-Chen Yeh, Jun-Ren Lee, Li-Fen Chen, Po-Lei Lee, Shyan-Shiou Chen, Low-Tone Ho, Daisy L. Hung, Ovid J.-L. Tzeng, and Jen-Chuen Hsieh. 2004. Orthographic and phonological processing of Chinese characters: An fMRI study. *NeuroImage*, **21**(4):1721−1731.

Lee, Chia-Ying, Jie-Li Tsai, Hsu-Wen Huang, Daisy L. Hung, Ovid J.-L. Tzeng. 2006. The temporal signatures of semantic and phonological activations for Chinese sublexical processing: An even-related potential study. *Brain Research*, **1121**(1):150-159.

Lee, Hsiang. 2010a. *Cangjie Input Methods in 30 Days 2*. Foruto. Last visited on 22 April 2010: input.foruto.com/cccls/cjzd.html.

Lee, Mu. 2010b. A quantitative study of the formation of Chinese characters. Last visited on 22 April 2010: chinese.exponode.com/0_1.htm. (in Chinese)

Liu, Chao-Lin, and Jen-Hsiang Lin. 2008. Using structural information for identifying similar Chinese characters. *Proc. of the 46th Annual Meeting of the Association for Computational Linguistics*, short papers, 93−96.

Liu, Chao-Lin, Kan-Wen Tien, Yi-Hsuan Chuang, Chih-Bin Huang, and Juei-Yu Weng. 2009a. Two applications of lexical information to computer-assisted item authoring for elementary Chinese. *Proc. of the 22nd Int'l Conf. on Industrial Engineering & Other Applications of Applied Intelligent Systems*, 470−480.

Liu, Chao-Lin, Kan-Wen Tien, Min-Hua Lai, Yi-Hsuan Chuang, and Shih-Hung Wu. 2009b. Capturing errors in written Chinese words. *Proc. of the 47th Annual Meeting of the Association for Computational Linguistics*, short papers, 25−28.

Liu, Chao-Lin, Kan-Wen Tien, Min-Hua Lai, Yi-Hsuan Chuang, and Shih-Hung Wu. 2009c. Phonological and logographic influences on errors in written Chinese words. *Proc. of the 7th Workshop on Asian Language Resources*, the 47th Annual Meeting of the ACL, 84−91.

Liu, Cheng-Lin, Stefan Jaeger, and Masaki Nakagawa. 2004. Online recognition of Chinese characters: The state-of-the-art. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **26**(2):198−213.

Wubihua. 2010. Last visited on 22 April 2010: en.wikipedia.org/wiki/Wubihua_method.

Yeh, Su-Ling, and Jing-Ling Li. 2002. Role of structure and component in judgments of visual similarity of Chinese Characters. *Journal of Experimental Psychology: Human Perception and Performance*, **28**(4):933−947.