Phonological and Logographic Influences on Errors in Written Chinese Words

Chao-Lin Liu[1] Kan-Wen Tien[2] Min-Hua Lai[3] Yi-Hsuan Chuang[4] Shih-Hung Wu[5]
[1-4]National Chengchi University, [5]Chaoyang University of Technology, Taiwan
{[1]chaolin, [2]96753027, [3]95753023, [4]94703036}@nccu.edu.tw, [5]shwu@cyut.edu.tw

Abstract

We analyze a collection of 3208 reported errors of Chinese words. Among these errors, 7.2% involved rarely used character, and 98.4% were assigned common classifications of their causes by human subjects. In particular, 80% of the errors observed in the writings of middle school students were related to the pronunciations and 30% were related to the logographs of the words. We conducted experiments that shed light on using the Web-based statistics to correct the errors, and we designed a software environment for preparing test items whose authors intentionally replace correct characters with wrong ones. Experimental results show that using Web-based statistics can help us correct only about 75% of these errors. In contrast, Web-based statistics are useful for recommending incorrect characters for composing test items for "incorrect character identification" tests about 93% of the time.

1 Introduction

Incorrect writings in Chinese are related to our understanding of the cognitive process of reading Chinese (e.g., Leck et al., 1995), to our understanding of why people produce incorrect characters and our offering corresponding remedies (e.g., Law et al., 2005), and to building an environment for assisting the preparation of test items for assessing students' knowledge of Chinese characters (e.g., Liu and Lin, 2008).

Chinese characters are composed of smaller parts that can carry phonological and/or semantic information. A Chinese word is formed by Chinese characters. For example, 新加坡 (Singapore) is a word that contains three Chinese characters. The left (土) and the right (皮) part of 坡, respectively, carry semantic and phonological information. The semantic information, in turn, is often related to the logographs that form the Chinese characters. Evidences show that production of incorrect characters are related to phonological, logographic, or the semantic aspect of the characters. Although the logographs of Chinese characters can be related to the lexical semantics, not all errors that are related to semantics were caused by the similarity in logographs. Some were due to the context of the words and/or permissible interpretations of different words.

In this study, we investigate issues that are related to the phonological and logographical influences on the occurrences of incorrect characters in Chinese words. In Section 2, we present the details about the sources of the reported errors. We have collected errors from a published book and from a group of middle school students. In Section 3, we analyze the causes of the observed errors. Native speakers of Chinese were asked to label whether the observed errors were related to the phonological or the logographic reasons. In Section 4, we explore the effectiveness of relying on Web-based statistics to correct the errors. We submitted an incorrect word and a correct word separately to Google to find the number of web pages that contained these words. The correct and incorrect words differed in just the incorrect character. We examine whether the number of web pages that contained the words can help us find the correct way of writing. In Section 5, we employ Web-based statistics in the process of assisting teachers to prepare test items for assessing students' knowledge of Chinese characters. Experimental results showed that our method outperformed the one reported in (Liu and Lin, 2008), and captured the incorrect characters better than 93% of the time.

2 Data Sources

We obtained data from three major sources. A list that contains 5401 characters that have been believed to be sufficient for everyday lives was obtained from the Ministry of Education (MOE) of Taiwan, and we call the first list the **Clist**, henceforth. The 5401 characters form the core basis for the BIG-5 code, and an official introduction of these 5401 characters is available at http://www.cns11643.gov.tw/AIDB/encodings.do#encode4.

We have two lists of words, and each word is accompanied by an incorrect way to write the word. The first list is from a book published by MOE (1996). The MOE provided the correct words and specified the incorrect characters which were mistakenly used to replace the correct characters in the correct words. 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 incorrect characters were entered into computers based on students' writings, ignoring those characters that did not actually exist and could not be entered.

We will call the first list of word the **Elist**, and the second the **Jlist** from now on. Elist and Jlist contain, respectively, 1490 and 1718 entries. Each of these entries contains a correct word and the incorrect character. Hence, we can reconstruct the incorrect words

Proceedings of the 7th Workshop on Asian Language Resources, ACL-IJCNLP 2009, pages 84–91,
Suntec, Singapore, 6-7 August 2009. ©2009 ACL and AFNLP

easily. Two or more different ways to incorrectly write the same words were listed in different entries and considered as two entries for simplicity of presentation.

3 Error Analysis of Written Words

Two human subjects, who are native speakers of Chinese and are graduate students in Computer Science, examined Elist and Jlist and categorized the causes of errors. They compared the incorrect characters with the correct characters to determine whether the errors were **pronunciation-related** or logographs-related. Referring to an error as being "semantics-related" is ambiguous. Two characters might not contain the same semantic part, but are still semantically related, e.g., misusing "偷"(tou1) for "投"(tou2) in "投機取巧". In this study, we have not considered this factor. For this reason we refer to the errors that are related to the sharing of logographic parts in characters as **composition-related**.

Among the 1490 and 1718 words in Elist and Jlist, respectively, the two human subjects had consensus over causes of 1441 and 1583 errors. It is interesting to learn that native speakers had a high consensus about the causes for the observed errors, but they did not always agree. To have a common standard in comparison, we studied the errors that the two subjects had agreed categorizations.

The statistics changed when we disregarded errors that involved characters not included in Clist. An error would be ignored if either the correct or the incorrect character did not belong to the Clist. It is possible for students to write such rarely used characters in an incorrect word just by coincidence.

After ignoring the rare characters, there were 1333 and 1645 words in Elist and Jlist, respectively. The subjects had consensus over the causes of errors for 1285 and 1515 errors in Elist and Jlist, respectively.

Table 1 shows the percentages of five categories of errors: ***C*** for the composition-related errors, ***P*** for the pronunciation-related errors, ***C&P*** for the intersection of *C* and *P*, ***NE*** for those errors that belonged to neither *C* nor *P*, and ***D*** for those errors that the subjects disagreed on the error categories. There were, respectively, 505 composition-related and 1314 pronunciation-related errors in Jlist, so we see 505/1645=30.70% and 1314/1645=79.88% in the table. Notice that *C&P* represents the intersection of *C* and *P*, so we have to deduct *C&P* from the sum of *C*, *P*, *NE*, and *D* to find the total probability, namely 1.

It is worthwhile to discuss the implication of the statistics in Table 1. For the Jlist, similarity between pronunciations accounted for nearly 80% of the errors, and the ratio for the errors that are related to compositions and pronunciations is 1:2.6. In contrast, for the Elist, the corresponding ratio is almost 1:1. The Jlist and Elist differed significantly in the ratios of the error types. It was assumed that the dominance of pronunciation-related errors in electronic documents was

Table 1. Error analysis for Elist and Jlist

	C	*P*	*C&P*	*NE*	*D*
Elist	66.09%	67.21%	37.13%	0.23%	3.60%
Jlist	30.70%	**79.88%**	20.91%	2.43%	7.90%

a result of the popularity of entering Chinese with pronunciation-based methods. The ratio 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 than composition-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 being so used to entering Chinese text with pronunciation-based method that the organization of their mental lexicons is also pronunciation related. The ratio for the Elist suggests that editors of the MOE book may have chosen the examples with a special viewpoint in their minds – balancing pronunciation and composition related errors.

4 Reliability of Web-based Statistics

In this section, we examine the effectiveness of using Web-based statistics to differentiate correct and incorrect characters. The abundant text material on the Internet gives people to treat the Web as a corpus (e.g., webascorpus.org). When we send a query to Google, we will be informed of the number of pages (**NOPs**) that possibly contain relevant information. If we put the query terms in quotation marks, we should find the web pages that literally contain the query terms. Hence, it is possible for us to compare the NOPs for two competing phrases for guessing the correct way of writing. At the time of this writing, Google found 107000 and 3220 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 individual words.) Hence, "strong" appears to be a better choice to go with "tea". This is an idea similar to the approach that compute collocations based on word frequencies (cf. Manning and Schütze, 1999). Although the idea may not work very well for small database, the size of the current Web should be considered large enough.

Using the quotation marks for the query terms enforced the influences of the surrounding characters in Chinese words, and provides a better clue for judging correct usage of Chinese characters. For instance, without the context, "每" and "美" might be used incorrectly to replace each other because they have the same pronunciation, i.e., Mei3. It is relatively unlikely for one to replace "每" with "美" when we write "每個" (*every one*), but these two characters can become admissible candidates when we write "美國" (*USA*) and "每國" (*every country*).

4.1 Field Tests

We test this strategy by sending the words in Elist and Jlist to Google to find the NOPs. We can retrieve the NOPs from the documents returned by Google, and compare the NOPs for the correct and the incorrect words to evaluate the strategy. Again, we focused on those in the 5401 words that the human subjects had consensus about their error types. Recall that we have 1285 and 1515 such words in Elist and Jlist, respectively. As the information available on the Web changes all the time, we also have to note that our experiments were conducted during the first half of March 2009. The queries were submitted at reasonable time intervals to avoid Google's treating our programs as malicious attackers.

Table 2 shows the results of our investigation. We considered that we had a correct result when we found that the NOP for the correct word was larger than the NOP for the incorrect word. If the NOPs were equal, we recorded an ambiguous result; and when the NOP for the incorrect word was larger, we recorded an incorrect event. We use '**C**', '**A**', and '**I**' to denote "correct", "ambiguous", and "incorrect" events in Table 2.

The column headings of Table 2 show the setting of the searches with Google and the set of words that were used in the experiments. We asked Google to look for information from web pages that were encoded in traditional Chinese (denoted **Trad**). We could add another restriction on the source of information by asking Google to inspect web pages from machines in Taiwan (denoted **Twn+Trad**). We were not sure how Google determined the languages and locations of the information sources, but chose to trust Google. The headings "**Comp**" and "**Pron**" indicate whether the words whose error types were composition and pronunciation-related, respectively.

Table 2 shows eight distributions, providing experimental results that we observed under different settings. The distribution printed in bold face showed that, when we gathered information from sources that were encoded in traditional Chinese, we found the correct words 73.12% of the time for words whose error types were related to composition in Elist. Under the same experimental setting, we could not judge the correct word 4.58% of the time, and would have chosen an incorrect word 22.30% of the time.

Statistics in Table 2 indicate that web statistics is not a very reliable factor to judge the correct words. The average of the eight numbers in the 'C' rows is only 71.54% and the best sample is 76.59%, suggesting that we did not find the correct words frequently. We would made incorrect judgments 24.75% of the time. The statistics also show that it is almost equally difficult to find correct words for errors that are composition and pronunciation related. In addition, the statistics reveal that choosing more features in the advanced search affected the final results. Using "Trad" offered better results in our experiments than using "Twn+Trad". This observation may arouse a perhaps controversial argument. Although Taiwan is the main area to use traditional Chinese, their web pages might not have used as accurate Chinese as web pages located in other regions.

Table 2. Reliability of Web-based statistics

		Trad		Twn+Trad	
		Comp	Pron	Comp	Pron
Elist	C	**73.12%**	73.80%	69.92%	68.72%
	A	**4.58%**	3.76%	3.83%	3.76%
	I	**22.30%**	22.44%	26.25%	27.52%
Jlist	C	76.59%	74.98%	69.34%	65.87%
	A	2.26%	3.97%	2.47%	5.01%
	I	21.15%	21.05%	28.19%	29.12%

4.2 An Error Analysis for the Field Tests

We have analyzed the reasons for why using Web-based statistics did not always find the correct words. Frequencies might not have been a good factor to determine the correctness of Chinese. However, the myriad amount of data on the Web should have provided a better performance.

The most common reason for errors is that some of the words are really confusing such that the majority of the Web pages actually used the incorrect words. Some of errors were so popular that even one of the Chinese input methods on Windows XP offered wrong words as possible choices, e.g., "雄起起" (the correct one) vs. "雄糾糾". It is interesting to note that people may intentionally use incorrect words in some occasions; for instance, people may choose to write homophones in advertisements.

Another popular reason is that whether a word is correct depends on a larger context. For instance, "小斯" is more popular than "小廝" because the former is a popular nickname. Unless we had provided more contextual information about the queried words, checking only the NOPs of "小斯" and "小廝" led us to choose "小斯", which happened to be an incorrect word when we meant to find the right way to write "小廝". Another difficult pair of words to distinguish is "紀錄" and "記錄".

Yet another reason for having a large NOP of the incorrect words was due to errors in segmenting Chinese character strings. Consider a correct character string "WXYZ", where "WX" and "YZ" are two correct words. 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" to contribute to the count for "海繽" which is an incorrect form of "海濱".

5 Facilitating Test Item Authoring

Incorrect character correction is a very popular type of test in Taiwan. There are simple test items for young children, and there are very challenging test items for the competitions among adults. Finding an attractive incorrect character to replace a correct character to form a test item is a key step in authoring test items.

We have been trying to build a software environment for assisting the authoring of test items for incorrect character correction (Liu and Lin, 2008, Liu et al., 2009). It should be easy to find a lexicon that contains pronunciation information about Chinese characters. In contrast, it might not be easy to find visually similar Chinese characters with computational methods. We expanded the original Cangjie codes (OCC), and employed the expanded Cangjie codes (ECC) to find visually similar characters (Liu and Lin, 2008).

Cangjie encoding (Chu, 2009) is a special system for representing the formation of Chinese characters with a sequence of at most five basic symbols. For instance, "坡" and "波" are represented by "土木竹水" and "水木竹水", respectively. It is evident that the Cangjie codes are useful for finding visually similar characters.

With a lexicon, we can find characters that can be pronounced in a particular way. However, this is not enough for our goal. We observed that there were different symptoms when people used incorrect characters that are related to their pronunciations. They may use characters that could be pronounced exactly the same as the correct characters. They may also use characters that have the same pronunciation and different tones with the correct character. Although relatively infrequently, people may use characters whose pronunciations are similar to but different from the pronunciation of the correct character.

We reported that replacing OCCs with ECCs to find visually similar characters could increase the chances to find similar characters. Instead of saving "土木竹水" for "坡" directly, we divide a Chinese character into subareas systematically, and save the Cangjie codes for each of the subareas. A Chinese character is stored with the information about how it is divided into subareas and the Cangjie sequences for each of its subareas. The internal code for how we divide "坡" is 2, and the ECC for "坡" has two parts: "土" and "木竹水". Yet, it was not clear as to which components of a character should use ECCs (Liu and Lin, 2008; Liu et al., 2009).

5.1 Formalizing the Extended Cangjie Codes

We analyzed the OCCs for all the characters in Clist to determine the list of basic components, with computer programs. We treated a basic Cangjie symbol as if it was a word, and computed the number of occurrences of n-grams based on the OCCs of the characters in Clist. Since the OCC for a character contains at most five symbols, the longest n-grams are 5-grams. Because the reason to use ECCs was to find common components in characters, we saved n-grams that repeated no less than three times in a list. After obtaining this initial list of n-grams, we removed those n-grams that were substrings of longer n-grams in the list.

In addition, the n-grams that appeared no less than three times might not represent an actual part in any Chinese characters. This may happen by chance because we considered only frequencies of n-grams when we generated the initial list at the previous step. For instance, the OCC codes for "曬" (shai4), "晤" (wu4), and "晨" (chen2) are "日一一心", "日一一口", and "日一一女", respectively. Although the substring "日一一" appears three times, it does represent an actual part of Chinese characters. Hence, we manually examined all of the n-grams in the initial list, and removed such n-grams from the list.

In addition to considering the frequencies of n-grams formed by the basic Cangjie codes to determine the list of components, we also took advantage of radicals that are used to categorize Chinese characters in typical printed dictionaries. Radicals that are standalone Chinese words were included in the list of components.

After selecting the list of basic components with the above procedure, we encoded the words in Elist with these basic components. We inherited the 12 ways reported in a previous work (Liu and Lin, 2008) to decompose Chinese characters. There are other methods for decomposing Chinese characters into components. Juang et al. (2005) and their team at the Sinica Academia propose 13 different ways for decomposing characters.

At the same time when we annotated individual characters with their ECCs, we may revise the list of basic components. If a character that actually contained an intuitively "common" part and that part had not been included in the list of basic component, we would add this part into the list to make it a basic component and revised the ECC for all characters accordingly. The judgment of being "common" is subjective, but we still maintained the rule that such common parts must appear in more than three characters. When defining the basic components, not all judgments are completely objectively yet, and this is also the case of defining the original Cangjie codes. We tried to be as systematic as possible, but intuition sometimes stepped in.

We repeated the procedure described in the preceding paragraph five times to make sure that we were satisfied with the ECCs for all of the 5401 characters. The current list contains 794 components, and we can revise the list of basic components in our work whenever necessary.

5.2 Recommending Incorrect Alternatives

With the pronunciation of Chinese characters in a dictionary and with our ECC encodings for words in the Elist, we can create lists of candidate characters for replacing a specific correct character in a given word to create a test item for incorrect character correction.

There are multiple strategies to create the candidate lists. We may propose the candidate characters because their pronunciations have the **s**ame **s**ound and the **s**ame **t**one with those of the correct character (denoted ***SSST***). Characters that have **s**ame **s**ounds and

different **t**ones (***SSDT***), characters that have si**m**ilar **s**ounds and **s**ame **t**ones (***MSST***), and characters that have si**m**ilar **s**ounds and **d**ifferent **t**ones (***MSDT***) can be considered as candidates as well. It is easy to judge whether two Chinese characters have the same tone. In contrast, it is not trivial to define "similar" sound. We adopted the list of similar sounds that was provided by a psycholinguistic researcher (Dr. Chia-Ying Lee) at the Sinica Academia. "坡" (po) and "玻" (bo) and "犯"(fan4) and "患"(huan4) are pairs that have similar sounds. It was observed that these are four possible reasons that people used incorrect characters in writing.

Because a Chinese character might be pronounced in multiple ways, character lists generated based on these strategies may include the same characters. More specifically, the lists *SSST* and *SSDT* may overlap when a character that can be pronounced in multiple ways, and these pronunciations share the same sound and have different tones. The characters "待" and "好" are such examples. "待" can be pronounced as "dai1" or "dai4", and "好" can be pronounced as "hao3" or "hao4". Hence, characters that can be pronounced as "hao3" will be listed in both SSST and SSDT for "好".

In addition, we may propose characters that look similar to the correct character. Two characters may look similar for many reasons (Liu et al., 2009). The most common reason is that they contain the same components, and the other is that they belong to the same **r**adical category and have the same total number of **s**trokes (***RS***), e.g., the pairs "己" and "巳", "記" and "計", and "谿" and "豁". When two characters contain the same component, the shared component might or might not locate at the same position, e.g., "部" and "陪".

In an authoring tool, we could recommend a selected number of candidate characters for replacing the correct character. We tried two different strategies to compare and choose the visually similar characters. The similarity is computed based on the number and the locations of shared Cangjie symbols in the ECCs of the characters. The first strategy (denoted ***SC1***) gave a higher score to the shared component that located at the same location in the two characters being compared. The second strategy (***SC2***) gave the same score to any shared component even if the component did not reside at the same location in the characters. The characters "頸", "勁", and "徑" share the same component "巠". When computing the similarity between these characters with *SC1*, the contribution of "巠" will be the same for any pair. When computing with *SC2*, the contribution of "巠" will be larger for the pair "頸" and "勁" than for the pair "徑" and "勁". In the former case, "巠" appears at the same location in the characters.

When there were more than 20 characters that receive nonzero scores in the *SC1* and *SC2* categories, we chose to select at most 20 characters that had leading scores as the list of recommended characters.

We had to set a bound on the number of candidate characters, i.e., 20, for strategies *SC1* and *SC2*. The number of candidates generated from these two strategies can be large and artificial, depending on our scoring functions for determining similarities between characters. We did not limit the sizes of candidate lists that were generated by other strategies because those lists were created based on more objective methods. The rules for determining "similar" sounds were given by the domain experts, so we considered the rules objective in this research.

For the experiments that we reported in the following subsection, we submitted more than 300 thousand of queries to Google. As we mentioned in Section 4.1, a frequent continual submission of queries to Google will make Google treat our programs as malicious processes. (We are studying the Google API for a more civilized solution.) Without the bound, it is possible to offer a very long list of candidates. On the other hand, it is also possible that our program does not find any visually similar characters for some special characters, and this is considered a possible phenomenon.

5.3 Evaluating the Recommendations

We examined the usefulness of these seven categories of candidates with errors in Elist and Jlist. The first set of evaluation (the inclusion tests) checked whether the lists of recommended characters contained the incorrect character in our records. The second set of evaluation (the ranking tests) was designed for practical application in computer assisted item generation. Only for those words whose actual incorrect characters were included in the recommended list, we replaced the correct characters in the words with the candidate incorrect characters, submitted the incorrect words to Google, and ordered the candidate characters based on their NOPs. We then recorded the ranks of the incorrect characters among all recommended characters.

Since the same character may appear simultaneously in *SC1*, *SC2*, and *RS*, we computed the union of these three sets, and checked whether the incorrect characters were in the union. The inclusion rate is listed under ***Comp***, representing the inclusion rate when we consider only logographic influences. Similarly, we computed the union for *SSST*, *SSDT*, *MSST*, and *MSDT*, checked whether the incorrect characters were in the union, and recorded the inclusion rate under ***Pron***, representing the inclusion rate when we consider only phonological influences. Finally, we computed the union of the lists created by the seven strategies, and recorded the inclusion rate under ***Both***.

The second and the third rows of Table 3 show the results of the inclusion tests when we recommended candidate characters with the methods indicated in the column headings. The data show the percentage of the incorrect characters being included in the lists that

Table 3. Incorrect characters were contained and ranked high in the recommended lists

	SC1	*SC2*	*RS*	*SSST*	*SSDT*	*MSST*	*MSDT*	*Comp*	*Pron*	*Both*
Elist	73.92%	76.08%	4.08%	91.64%	18.39%	3.01%	1.67%	81.97%	99.00%	93.37%
Jlist	67.52%	74.65%	6.14%	92.16%	20.24%	4.19%	3.58%	77.62%	99.32%	97.29%
Elist	3.25	2.91	1.89	2.30	1.85	2.00	1.58			
Jlist	2.82	2.64	2.19	3.72	2.24	2.77	1.16			
Elist	19.27	17.39	11.34	19.13	8.29	19.02	9.15			
Jlist	17.58	16.24	12.52	22.85	9.75	22.11	7.68			

were recommended by the seven strategies. Notice that the percentages were calculated with different denominators. The number of composition-related errors was used for *SC1*, *SC2*, *RS*, and *Comp* (e.g., 505 that we mentioned in Section 3 for Jlist); the number of pronunciation-related errors for *SSST*, *SSDT*, *MSST*, *MSDT*, and *Pron* (e.g., 1314 mentioned in Section 3 for the Jlist); the number of either of these two types of errors for *Both* (e.g., 1475 for Jlist).

The results recorded in Table 3 show that we were able to find the incorrect character quite effectively, achieving better than 93% for both Elist and Jlist. The statistics also show that it is easier to find incorrect characters that were used for pronunciation-related problems. Most of the pronunciation-related problems were misuses of homophones. Unexpected confusions, e.g., those related to pronunciations in Chinese dialects, were the main reason for the failure to capture the pronunciation-related errors. (Namely, few pronunciation-related errors were not considered in the information that the psycholinguist provided.) *SSDT* is a crucial complement to *SSST*.

There is still room to improve our methods to find confusing characters based on their compositions. We inspected the list generated by *SC1* and *SC2*, and found that, although *SC2* outperformed *SC1* on the inclusion rate, *SC1* and *SC2* actually generated complementary lists in many cases, and should be used together. The inclusion rate achieved by the *RS* strategy was surprisingly high. We found that many of the errors that were captured by the *RS* strategy were also captured by the *SSST* strategy.

The fourth and the fifth rows of Table 3 show the effectiveness of relying on Google to rank the candidate characters for recommending an incorrect character. The rows show the average ranks of the included cases. The statistics show that, with the help of Google, we were able to put the incorrect character on top of the recommended list when the incorrect character was included. This allows us to build an environment for assisting human teachers to efficiently prepare test items for incorrect character identification.

Note that we did not provide data for all columns in the fourth and the firth rows. Unlike that we show the inclusion rates in the second and the third rows, the fourth and the fifth rows show how the actual incorrect characters were ranked in the recommended lists. Hence, we need to have a policy to order the characters of different lists to find the ranks of the incorrect characters in the integrated list.

However, integrating the lists is not necessary and can be considered confusing to the teachers. The selection of incorrect characters from different lists is related to the goals of the assessment, and it is better to leave the lists separated for the teachers to choose. The same phenomenon and explanation apply to the sixth and the seventh rows as well.

The sixth and the seventh rows show the average numbers of candidate characters proposed by different methods. Statistics shown between the second and the fifth rows are related to the recall rates (cf. Manning and Schütz, 1999) achieved by our system. For these four rows, we calculated how well the recommended lists contained the reported errors and how the actual incorrect characters ranked in the recommended lists. The sixth and the seventh rows showed the costs for these achievements, measured by the number of recommended characters. The sum of the sixth and the seventh rows, i.e., 103.59 and 108.75, are, respectively, the average numbers of candidate characters that our system recommended as possible errors recorded in Elist and Jlist. (Note that some of these characters were repeated.)

There are two ways to interpret the statistics shown in the sixth and the seventh rows. Comparing the corresponding numbers on the fourth and the sixth rows, e.g., 3.25 and 19.27, show the effectiveness of using the NOPs to rank the candidate characters. The ranks of the actual errors were placed at very high places, considering the number of the originally recommended lists. The other way to use the statistics in the sixth and the seventh rows is to compute the average precision. For instance, we recommended an average 19.13 characters in *SSST* to achieve the 91.64 inclusion rate. The recall rate is very high, but the averaged precision is very low. This, however, is not a very convincing interpretation of the results. Having assumed that there was only one best candidate as in our experiments, it was hard to achieve high precision rates. The recall rates are more important than the precision rates, particularly when we have proved that the actual errors were ranked among the top five alternatives.

When designing a system for assisting the authoring of test items, it is not really necessary to propose all of the characters in the categories. In the reported experiments, choosing the top 5 or top 10 candidates will contain the most of the actual incorrect characters based on the statistics shown in the fourth and the fifth rows. Hence the precision rates can be significantly increased practically. We do not have to merge the candidate characters among different categories

because choosing the categories of incorrect characters depends on the purpose of the assessment. Reducing the length of the candidate list increases the chances of reducing the recall rates. Achieving the best trade off between precision and recall rates relies on a more complete set of experiments that involve human subjects.

Furthermore, in a more realistic situation, there can be more than one "good" incorrect character, not just one and only gold standard as in the reported experiments. It is therefore more reasonable the compute the precision rates based the percentage of "acceptable" incorrect characters. Hence, the precision rates are likely to increase and become less disconcerting.

We reported experimental results in which we asked 20 human subjects to choose an incorrect character for 20 test items (Liu et al., 2009). The best solutions were provided by a book. The recommendations provided by our previous system and chosen by the human subjects achieved comparable qualities.

Notice that the numbers do not directly show the actual number of queries that we had to submit to Google to receive the NOPs for ranking the characters. Because the lists might contain the same characters, the sum of the rows showed just the maximum number of queries that we submitted. Nevertheless, they still served as good estimations, and we actually submitted 103.59×1441(=149273) and 108.75×1583 (=172151) queries to Google for Elist and Jlist in experiments from which we obtained the data shown in the fourth and the fifth rows. These quantities explained why we had to be cautious about how we submitted queries to Google. When we run our program for just a limited number of characters, the problems caused by intensive queries should not be very serious.

5.4 Discussions

Dividing characters into subareas proved to be crucial in our experiments (Liu and Lin, 2008; Liu et al., 2009), but this strategy is not perfect, and could not solve all of the problems. The way we divided Chinese characters into subareas like (Juang et al., 2005; Liu and Lin, 2008) sometimes contributed to the failure of our current implementation to capture all of the errors that were related to the composition of the words. The most eminent reason is that how we divide characters into areas. Liu and Lin (2008) followed the division of Cangjie (Chu, 2009), and Juang et al. (2005) proposed an addition way to split the characters.

The best divisions of characters appear to depend on the purpose of the applications. Recall that each part of the character is represented by a string of Cangjie codes in ECCs. The separation of Cangjie codes in ECCs was instrumental to find the similarity of "苗" and "福" because "田" is a standalone subpart in both "苗" and "福". The Cangjie system has a set of special rules to divide Chinese characters (Chu, 2009; Lee, 2008). Take "副" and "福" for example. The component "畐" is recorded as an standalone part in "副", but is divided into two parts in "福". Hence, "畐" is stored as one string, "一ㄖ田", in "副" and as two strings, "一ㄖ" and "田", in "福". The different ways of saving "畐" in two different words made it harder to find the similarity between "副" and "福". An operation of concatenation is in need, but the problems are that it is not obvious to tell when the concatenation operations are useful and which of the parts should be rejoined. Hence, using the current methods to divide Chinese characters, it is easy to find the similar between "苗" and "福" but difficult to find the similar between "副" and "福". In contrast, if we enforce a rule to save "畐" as one string of Cangjie code, it will turn the situations around. Determining the similarity between "苗" and "福" will be more difficult than finding the similarity between "副" and "福".

Due to this observation, we have come to believe that it is better to save the Chinese characters with more detailed ECCs. By saving all detailed information about a character, our system can offer candidate characters based on users' preferences which can be provided via a good user interface. This flexibility can be very helpful when we are preparing text materials for experiments for psycholinguistics or cognitive sciences (e.g., Leck et al, 1995; Yeh and Li, 2002).

6 Summary

The analysis of the 1718 errors produced by real students show that similarity between pronunciations of competing characters contributed most to the observed errors. Evidences show that the Web statistics are not very reliable for differentiating correct and incorrect characters. In contrast, the Web statistics are good for comparing the attractiveness of incorrect characters for computer assisted item authoring.

Acknowledgments

This research was supported in part by the National Science Council of Taiwan under grant NSC-97-2221-E-004-007-MY2. We thank anonymous reviewers for their invaluable comments.

References

B.-F. Chu. 2009. *Handbook of the Fifth Generation of the Cangjie Input Method*, available at http://www.cbflabs.com/book/ocj5/ocj5/index.html. Last visited on 30 April 2009.

D. Juang, J.-H. Wang, C.-Y. Lai, C.-C. Hsieh, L.-F. Chien, J.-M. Ho. 2005. Resolving the unencoded character problem for Chinese digital libraries, *Proc. of the 5th ACM/IEEE Joint Conf. on Digital Libraries*, 311–319.

S.-P. Law, W. Wong, K. M. Y. Chiu. 2005. Whole-word phonological representations of disyllabic

words in the Chinese lexicon: Data from acquired dyslexia, *Behavioural Neurology*, **16**, 169–177.

K. J. Leck, B. S. Weekes, M. J. Chen. 1995. Visual and phonological pathways to the lexicon: Evidence from Chinese readers, *Memory & Cognition*, **23**(4), 468–476.

H. Lee. 2008. *Cangjie Input Methods in 30 Days*, http://input.foruto.com/cjdict/Search_1.php, Foruto Company, Hong Kong. Last visited on 30 April 2009.

C.-L. Liu, K.-W. Tien, Y.-H. Chuang, C.-B. Huang, J.-Y. Weng. 2009. 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.

C.-L. Liu, J.-H. Lin. 2008. Using structural information for identifying similar Chinese characters, *Proc. of the 46th ACL*, short papers, 93−96.

C. D. Manning, H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press. 1999.

MOE. 1996. *Common Errors in Chinese Writings* (常用國字辨似), Ministry of Education, Taiwan.

S.-L. Yeh, J.-L. 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.