Capturing 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

A collection of 3208 reported errors of Chinese words were analyzed. Among which, 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 writings of middle school students were related to the pronunciations and 30% were related to the compositions of words. Experimental results show that using intuitive Web-based statistics helped us capture only about 75% of these errors. In a related task, the 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. Evidences show that production of incorrect characters are related to either phonological or the semantic aspect of the characters.

In this study, we investigate several issues that are related to incorrect characters in Chinese words. In Section 2, we present the sources of the reported errors. In Section 3, we analyze the causes of the observed errors. In Section 4, we explore the effectiveness of relying on Web-based statistics to correct the errors. The current results are encouraging but demand further improvements. 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 best candidates for incorrect characters 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. We have two lists of words, and each word is accompanied by an incorrect way to write certain words. The first list is from a book published by MOE (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 words 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 incorrect words 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 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 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 semantic-related. Referring to an error as being "semantic-related" is ambiguous. Two characters might not contain the same semantic part, but are still semantically related. In this study, we have not considered this factor. For this reason we refer to the errors that are related to the sharing of semantic parts in characters as **composition-related**.

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. Hence, we studied the errors that the two subjects had agreed categorizations. Among the 1490 and 1718 words in Elist and Jlist, respectively, the two human subjects had consensus over causes of 1441 and 1583 errors.

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 categories for 1285

Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 25–28,
Suntec, Singapore, 4 August 2009.

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%

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 30.70% (=505/1645) and 79.88% (=1314/1645) 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 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 the errors due to pronunciation and composition.

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". How does this strategy serve for learners of Chinese?

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%

We verified this strategy by sending the words in both the Elist and the 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 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 is 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%, suggest-

ing 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 has proclaimed to be the major region to use traditional Chinese, their web pages might not have used as accurate Chinese as web pages located in other regions.

We have analyzed the reasons for why using Web-based statistics did not 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. Google's rephrasing our submitted queries is an important factor, and, in other cases, incorrect words were more commonly used.

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).

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.

As Liu and Lin (2008) reported, replacing OCC with ECC to find visually similar characters could increase the chances to find similar characters. Yet, it was not clear as to which components of a character should use ECC.

5.1 Formalizing the Extended Cangjie Codes

We analyzed the OCCs for all the words in Clist to determine the list of basic components. We treated a Cangjie basic symbol as if it was a word, and computed the number of occurrences of n-grams based on the OCCs of the words 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 ECC was to find common components in characters, we disregarded n-grams that repeated no more than three times. In addition, the n-grams that appeared more than three times might not represent an actual component in Chinese characters. Hence, we also removed such n-grams from the list of our basic components. This process naturally made our list include radicals that are used to categorize Chinese characters in typical printed dictionaries. The current list contains 794 components, and it is possible to revise the list of basic components in our work whenever necessary.

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

5.2 Recommending Incorrect Alternatives

With a dictionary that provides the pronunciation of Chinese characters and the improved 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 **d**ifferent **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.

In addition, we may propose characters that look similar to the correct character. Two characters may look similar for two reasons. They may contain the same components, or they contain the same **r**adical and have the same total number of **s**trokes (***RS***). When two characters contain the same component, the shared component might or might not locate at the same position within the bounding boxes of characters.

In an authoring tool, we could recommend a limited number of candidate characters for replacing the correct character. We tried two strategies to compare and choose the visually similar 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 strat-

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			

egy (***SC2***) gave the same score to any shared component even if the component did not reside at the same location in the characters. When there were more than 20 characters that receive nonzero scores, we chose to select at most 20 characters that had leading scores as the list of recommended characters.

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 only 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***. 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***. 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. The data show the percentage of the incorrect characters being included in the lists that 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 the 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 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 characters that had exactly the same pronunciations with the correct characters. Unexpected confusions, e.g., those related to pronunciations in Chinese dialects, were the main for the failure to capture the pronunciation-related errors. 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 and should be used together. The inclusion rate achieved by the RS strategy was surprisingly high.

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.

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.

Acknowledgements

This research has been funded in part by the National Science Council of Taiwan under the grant NSC-97-2221-E-004-007-MY2. We thank the anonymous reviewers for invaluable comments, and more responses to the comments are available in (Liu et al. 2009).

References

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.

C.-L. Liu et al. 2009. Phonological and logographic influences on errors in written Chinese words, *Proc. of the 7th Workshop on Asian Language Resources*, 47th ACL.

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

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