What You See is What You Worry About: Errors – Real and Imagined

Janet C Read

ChiCi

UCLan

Preston

+44(0)1772 893285

jcread@uclan.ac.uk

ABSTRACT

This paper describes a text task in which children wrote their own stories in their regular school books, copied these stories onto digital paper using digital pens, had their handwritten stories recognized by the computer software, and then, looked at the text presented back to them and highlighted errors. There was considerable variability in the ability of the children to spot errors. Some children marked text as being wrong when in fact it was right. Most children spotted almost all the errors but were less likely to notice errors where incorrect but reasonable words had been presented back to them than where the words given by the recognizer were nonsense. In 13 instances, children had one of their own errors corrected by the interface but this was not noticed. The study highlighted several difficulties with the classification and reporting of errors in handwriting recognition based interfaces especially in the counting and classification of errors. Two new metrics for classifying and counting errors in this context are therefore proposed in this paper.

Categories and Subject Descriptors

[H.5.2 User Interfaces \(D.2.2, H.1.2, I.3.6\)](#) Input devices and strategy, Evaluation.

General Terms

Measurement, Performance, Design, Experimentation, Human Factors, Standardization,

Keywords

Children, Handwriting recognition, Text input, Errors, Tolerance, Digital ink

1. INTRODUCTION

Children are now widespread users of computer technology [1] but in many cases they simply use adult technologies that do not match their needs very well. The mismatches of needs to children can be caused by a poor understanding of the context of use or a poor understanding of children's abilities and difficulties. In addition, methods for evaluating technologies are inherited from the adult work domain and are not always well suited to interactions where children are involved. In designing technologies for children, context is especially

important: technologies for the school classroom need to be designed to fit easily into that space; in many instances they need to be almost invisible.

Digital ink is one such 'almost invisible' technology. It requires only a digital notebook and a digital pen. For children, the pen and notebook are easy to use [2], and the possibilities for digital text in the support of writing are significant [3]. Digital ink can exist in two forms, it can remain as an ink rendering (in which case it looks like handwriting) or can be converted using a recognition process into ASCII text. There are several reasons to convert digital ink to text but the most compelling reason is to allow the subsequent manipulation and editing of the text in a traditional word processor.

Studies have shown that children's writing can be converted to ASCII text but that there will always be mistakes: some mistakes come from the children, some from the recognizer [4].

Recognition based systems will always be error prone and it is customary to evaluate these systems based on the number of errors that appear in the final copy [5], [6]. As dealing with user generated text in this sort of work is problematic, most studies of handwriting recognition effectiveness use controlled situations with users copying pre-prepared text [7]. These controlled situations do not really mimic the real world well, especially where the users are children. One problem, when children use digital ink technologies for text input, is that metrics used to evaluate the technologies do not take into account the specific behaviors of children, in particular, their spelling behaviors.

In considering the usefulness of digital ink for writing, not only do the numbers of errors in the system have to be considered, but also the user experience. In a study by [8], it was identified that children invariably failed to see errors and that correction methods were complicated. Thus, the visibility of errors in recognized text, and the ability of a child to find errors is important if digital ink is to have any value in the classroom.

2. THE RESEARCH STUDY

The study described here had two primary purposes; the first was to investigate the viability of carrying out recognition experiments with composed (user generated), rather than copied, text. In this investigation, consideration had to be made as to the usefulness of the usual metrics for the evaluation of text input with digital ink. The second purpose of the study was to discover the extent to which children noticed and found errors in text that had been recognized.

2.1 Procedure

The study took place over three days. On the first day, all the attending children in a class of UK school children, aged 7 and 8, wrote a piece of narrative in their exercise books about a

fable (the tale of Rosy Rhodipus) they had been studying. This activity was carried out in normal class time and the children wrote their stories independently with minimal help from the teacher. Thus, many of the children's stories had grammatical errors and spelling errors.

On the second day, the children came in fours to copy their stories from their textbooks to A5 digital paper books using Logitech IO2 pens. During this task, the children were watched by the research team but there were no interventions. It was noted (but not considered important) that several of the children made small changes to their original stories but in the main, the work was copied without much variation.

Sixteen stories were selected (using a random selection algorithm) to be further investigated. The digital writing from these stories was uploaded to a laptop PC and converted to text using MyScript recognition software. Work recognized in this way was not corrected and was then printed using a size 16 sans serif font ready to be presented back to the children.

On day three, two researchers went back to the classroom and brought the children out of class in twos to see what the computer had done to their work. They were given highlighter pens and told to mark the errors on the printed paper. Children were also asked for their views as to how the computer could have done better; they were then thanked for their participation.

2.2 Data Analysis

From each child, there were two paper outputs, the first (DW) was the digital writing (which was in the notebook), the second was the recognized text (RT) on paper. Reference was not made back to the original story as this was not an investigation of the children's ability to copy!

The lead researcher scrutinized the digital writing (DW) and, for each child, created a third object of interest, the written text (WT) this was a typed presentation of what the researcher assumed the child had written – it included spelling errors and grammar errors. Once the WT was created, the DW was not referenced. The WT was initially looked at on its own to find out how many words the children had written (WW), and how many words they had spelt wrong (SW).

It has previously been noted by the author that the creation of this written text (WT) is error prone. Errors are caused by a) the transcriber (the researcher in this case) guessing a character wrongly, e.g. thinking an 'a' is an 'e', b) the transcriber missing a word or character during transcription, and c) the transcriber 'correcting' an incorrect spelling without realizing. A study of the effects of these errors is out of scope in this paper but in general, an error rate of around 5% should be considered.

In the second phase of analysis, the recognized text (RT) was analyzed and a count was made of the words in it (WR). This was problematic in two ways. Firstly, the markings made by the child during handwriting cause emdashes to appear between words, in which case these words, for example 'Photon-Rosy' and 'idiom-t' in G6 were hard to describe. The procedure adopted was that if one, or both, of the words on the side of the emdash were correct, the two parts would be counted as two words, otherwise the word would count as one. The other area for confusion was where single letters appeared, for example 'E pupfish' in B8 (the child had intended Egyptian!) in these cases, for consistency, a letter on its own was considered a separate word.

Stage three involved the researcher comparing the recognized

text (RT) with the written text (WT) and highlighting the errors (RE) (where the word written by the child had turned out wrong) and the corrections (RC). (where the child had written a word wrong but the recognizer had fixed it). An intelligent pairwise comparison of words was used at this stage rather than the use of an automated algorithm.

Finally, it was then possible to look at the highlighting that the children had done and specify the number of recognition errors found by the children (REF), the number of errors found by the children that were not really errors at all (NE), the number of corrections noted by the children (RCF) and consequently, the recognition errors (REM) and the recognition corrections (RCM) missed by the children.

2.3 Results

Table 1 summarises the key results for the research study: The codes, B1 – G6 refer to the individual children, with G and B indicating girls and boys. Dividing the number of recognition errors (RE) by the number of words in the recognized text (WR) it was possible to calculate the recognition error rate (as a word error rate - WER) for each text sample; shown in Figure 1.

Table 1 - Results from the Study

	WW	SW	WR	RC	RE	REF	RE M	NE
B1	30	4	24	0	14	7	7	1
B2	46	3	43	1	17	17	0	3
B3	21	0	21	0	5	4	1	0
B4	54	3	53	2	6	5	1	1
B5	40	4	41	0	10	10	0	3
B6	41	1	44	0	14	13	1	0
B7	51	4	43	3	13	12	1	4
B8	37	2	38	1	8	6	2	0
B9	72	2	64	0	36	33	3	0
B10	43	8	41	2	16	15	1	4
G1	107	4	109	0	15	14	1	0
G2	74	4	73	1	9	8	1	2
G3	49	4	49	0	26	23	3	16
G4	29	2	28	0	5	5	0	0
G5	98	2	97	1	15	13	2	0
G6	104	3	105	2	10	10	0	0
AVG	56.0		54.6		13.7	12.19	1.5	2.1
SD	27.2		27.8		8.0	7.52	1.8	4.0

The percentages of errors found and missed by the children were also calculated. Most of the children, as shown in Figure 2, found most of the errors. There was a weak positive correlation between the number of words written by the child and the percentage of errors found. There was also a weak negative correlation between the number of words written and the word error rate. These two observations suggest that children who write more write better for the recognizer and are also better at finding errors. This is probably because they both form their letters well and have good reading skills.

Figure 1 - The word error rates across the sample

As shown in the table, half the children highlighted words as errors (NE) which in fact were correct, and three quarters of the children missed errors. It was also interesting to note that none of the children noticed when the recognizer fixed their errors, thus RCF was zero across the sample and therefore RCM was equal to RC. These phenomena were further investigated and are discussed in the next section

Figure 2 -Percentage of Errors found

3. ERRORS REAL AND IMAGINED

The discussion here is in three parts. Firstly, the errors that the children missed are discussed, then, those that children thought were errors but in fact were not errors are considered and finally a lost error type is considered, these are those made by the child at the point of text creation.

3.1 Errors that Children miss

Altogether, the children missed 24 errors. Six of these were the single letters, alluded to in the data analysis section of this paper, that, using the definition of a word, would always have been errors but which it is highly possible the children would have ignored when marking incorrect words. Three of the missed errors were where the recognizer had badly recognized a proper word but where the proper word intended ('Rhodipus') was recognized to a word that was quite close to the intended word. One of the missed words was the presentation 'avery' where the child had initially written 'a very'. The remaining 14 words were standard words that were missed; around a half of these made sense in the context of the sentence where they were situated, for instance, in one case the text read '... was kind...' where '...was rich...' was intended. The others read as nonsense.

3.2 Errors that children invent

The errors that children 'invented' were mainly next to, or between, other (real) errors. There were 35 errors invented in total, 14 of these were in-between two real errors, 15 were alongside real errors and only 5 were isolated words. It is highly likely that the child, on reading the text and finding it didn't make sense, also marked adjacent words as incorrect due to their contribution to the 'no-sense' making.

3.3 Errors that vanished

As can be seen in table 1, thirteen errors that the children made during writing were in fact 'fixed' by the recognition process. In general this fixing was related to the child's spelling of the original word. Some examples of the effects of children's spellings are shown in table 2.

Table 2 - Children's spellings and what happened

Child wrote	Recognised	Child noticed	Coded as
Jells (jealous)	Jets	Y	RE
Marrdie (married)	Mantle	Y	RE
Happly (happily)	happily	N	RC
Mine (mean)	Shrine	Y	RE
Agerse (ages)	Agene	Y	RE
Cidnapped	kidnapped	N	RC
Capgered (captured)	Lapford	Y	RE

It cannot be assumed to suggest that the children's errors that were subsequently fixed, had they remained in the later text, would have been marked as errors by the children, it is certainly the case that these 'vanishing' errors were not noticed by the children. These vanishing errors represented around 25% of the children's spelling errors. In the other spelling errors, the children tended to notice the mistake as the recognizer gave 'strange' suggestions.

3.4 Justice for the Recognizer

This study has shown that in a recognition-based system there are.

Errors created by the children that then end up as recognition errors.

Errors created by the recognizer that were not errors when they were created by the children.

Errors created by the children that were fixed by the recognizer.

Errors that are noticed and

Errors that are not noticed.

When calculating recognition performance, the usual WER metric takes no account of the accuracy of the initial written (WT) text. In addition, there is no compensation for the good work done by the recognizer in correcting (in this case the children's) errors. A modified metric is proposed that measures (albeit in a semi-subjective way) recognition performance. Note that this could be greater than 100% if a child spelt every word wrong but every word were corrected, the RP would be 200%.

$$RP = \frac{WR - ((RE - SW) - RC)}{WR}$$

Where RP = recognition performance, WR = number of words written, RE = number of recognition errors remaining, SW = number of words spelt wrong in the original (WT) text, and RC is the recognition corrections.

A comparison of RP with the traditional Word Error Rate (which takes no account of the children's errors) is shown in Figure 3. In this representation, the Word Error Rate was converted into a Word Correct Rate in order to make a comparison and was expressed as a decimal.

Figure 3 - Traditional rates and recognition performance

It can be seen that in most cases, the recognition performance is better than the Word Correct Rate – this is a result of the adjustment in the recognition performance metric for the ‘smartness’ of the recognizer..

3.5 Reality for the Child

A metric has also been derived to demonstrate how well the children did at error finding. Given that every error found by the child would, if the writing was to be later used, incur a correction, it seemed reasonable to not just reward correctly found errors but also to penalize errors found that were not really errors at all. The following metric is proposed to cover user performance:

$$UP = \frac{(REF) + ((WR - RE) - NE))}{WR}$$

Where REF = recognition errors found, WR = words written, RE = recognition errors in the written text, and NE = errors marked as errors that were not errors. The user performance of each child, as compared to the rougher metric of percentage of errors found, is shown in Figure 4.

4. CONCLUSIONS

This study has shown that children are good at finding mistakes but also rather good at making them! Mistakes would have been reduced had the recognition software had a more ‘phonically’ friendly dictionary – an example can be seen in table 2 where the child’s spelling of ‘captured’ feels like it should have been better recognized.

Figure 4 - User Performance and % Errors found

In finding errors, confidence shading might help children to realize that more of the recognized text is correct than they first thought (this is where ‘uncertain’ words are presented in a shade of black and the more certain a word is, the darker it is displayed). The two new metrics that have been produced take better account of the errors in this sort of system, although both are to some extent subjective but they are considered to be more realistic and less damning.

Future work will compare user performance in error spotting with effort of correction and will focus on the design of a system that takes account of the errors of the children and the smartness of the system whilst also enabling children to discover errors more easily.

5. ACKNOWLEDGMENTS

Thanks to the children who participated in this study and Emanuela Mazzone for discussions on the metrics..

6. REFERENCES

- [1] Wartela, E.A., J.H. Lee, and A.G. Caplovitz, Children and Interactive Media. 2002, Markle Foundation: Texas, US. p. 33.
- [2] Read, J.C. The Usability of Digital Ink Technologies for Children and Teenagers. in HCI2005. 2005. Edinburgh, UK: Springer.
- [3] Read, J.C., S.J. MacFarlane, and M. Horton. The Usability of Handwriting Recognition for Writing in the Primary Classroom. in HCI 2004. 2004. Leeds, UK: Springer Verlag.
- [4] Read, J.C., A study of the usability of handwriting recognition for text entry by children. Interacting with Computers, 2007. 19(1): p. 57 - 69.
- [5] MacKenzie, I.S. and L. Chang, A performance comparison of two handwriting recognizers. Interacting with Computers, 1999. 11(3): p. 283 - 297.
- [6] Mankoff, J. and G. Abowd, Error Correction Techniques for Handwriting, Speech, and other ambiguous or error prone systems. 1999, GVI Centre: Georgia, MA.
- [7] MacKenzie, I.S. and R.W. Soukoreff. Phrase Sets for Evaluating Text Entry Techniques. in CHI 2003. 2003. Ft. Lauderdale, FL: ACM Press.
- [8] Read, J.C., S.J. MacFarlane, and C. Casey. Oops! Silly me! Errors in a Handwriting Recognition-based Text entry Interface for Children. in NordiChi 2002. 2002. Aarhus, Denmark: ACM Press.