Reading errors made by skilled and unskilled readers: evaluating a system that generates reports for people with poor literacy

Sandra Williams and Ehud Reiter
Department of Computing Science
University of Aberdeen
Aberdeen AB24 3UE
Scotland, UK

swilliam@csd.abdn.ac.uk, ereiter@csd.abdn.ac.uk

Abstract

We describe part of an evaluation of a natural language generation system that generates literacy assessment reports for adults with poor basic skills. Research was focused on how to generate more readable documents. To evaluate readability of the system's output, we previously measured comprehension and reading speed. Here we describe a post-hoc investigation where we measured reading errors and disfluencies. The aim was to find out if modifications the system makes for readability resulted in less errors and disfluencies when the output was read aloud. We found that poor readers make less substitution errors on reports generated using readability preference rules.

Introduction

This paper describes an analysis of reading errors and disfluencies (pause errors) in audio recordings made during an evaluation of the readability of automatically generated texts (Williams, PhD thesis, in preparation). The system that generates the texts is a natural language generation (NLG) system called GIRL (Generator for individual Reading Levels) (Williams et al. 2003). GIRL generates reports about how well an adult has done in an assessment of his or her basic literacy skills. The intended audience for the reports is adults with poor literacy. About one fifth of the adult population of most developed countries has poor literacy (Binkley et al. 1997) and the focus of research was on generating more readable documents for poor readers. We focussed in particular on discourse issues such as ordering of information, selection of punctuation, selection of discourse cue phrases (small phrases like "that is", "but", and "for example") and positioning of cue phrases.

We previously evaluated the readability of GIRL's reports by measuring comprehension and reading speed (Williams PhD, in preparation). Comprehension was measured using paper-based comprehension questions, giving help with reading and writing where necessary. To measure reading speed, participants were recorded reading their reports aloud. We noticed that readers made many reading errors. Since reading errors are also an indicator of the reading difficulty of a text, we extend our evaluation here by classifying, annotating and measuring the reading errors in the recordings. The aim is to find out whether modifications the system makes for readability result in less reading errors. Our hypotheses can be stated as follows:

- Poor readers will make fewer reading errors when reading an "easy" version of a report generated by the system than a "hard" version.It will make little difference to good readers which version they read.

Thirty-six participants in the study were classified as good readers or poor readers based on their score in a literacy assessment (Basic Skills Agency et al. 2001). They read reports about their performance in the assessment generated by GIRL. Reports received were randomly one of two types (a) "easy" texts generated using readability preference rules, or (b) "hard" text generated using frequency rules. The experimental design was thus a two by two matrix of good and poor readers reading *easy* and *hard* texts.

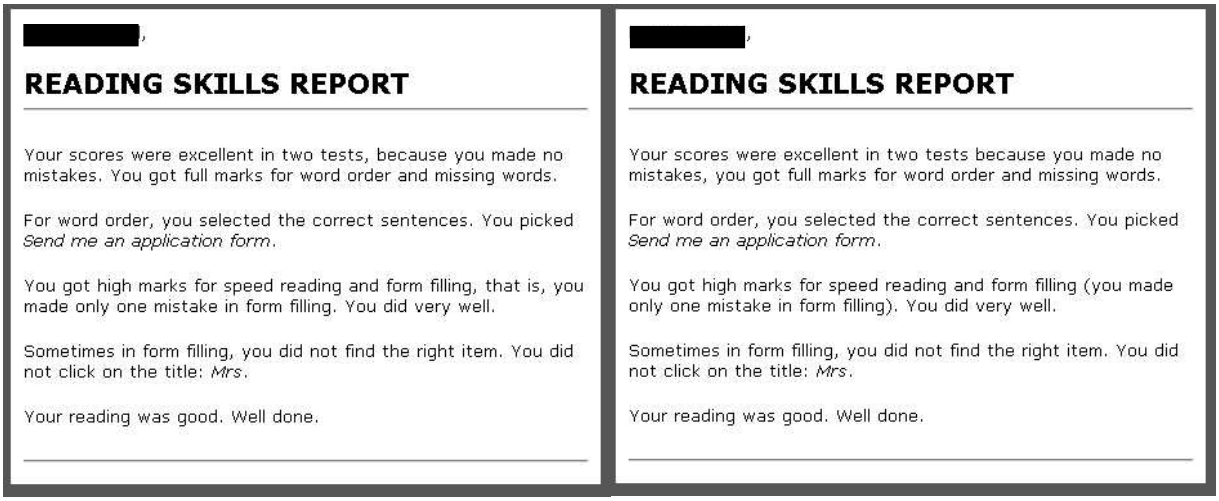
Related work

As text difficulty increases, the numbers of oral reading errors and disfluencies made by children increase (Blaxhall and Willows 1984, Young and Bowers 1995). Additionally, poor readers' ability to recognise phrase boundaries decreases as text difficulty increases, but good readers' ability remains unaffected (Young and Bowers 1995). Although these studies tested children's reading, we could reasonably expect adults to show similar sensitivities to text difficulty. We therefore decided to use measures of oral reading errors to indicate the relative difficulties of texts generated by our system for adults.

Classification of different kinds of reading errors depends on the intended use of the data. Hulslander (2001, Olson Reading Lab.) derived a classification of reading errors that was used to annotate a corpus for training a speech recogniser used in Project LISTEN (Fogarty et al. 2001). In project LISTEN, the recogniser was used to monitor a child reading aloud so that the system could judge when a child was making mistakes and when to interrupt him/her. Their classification scheme identifies a large number of types of errors grouped under the headings of substitutions, insertions, omissions, fluency errors, repetitions and self-corrections. Our usage of reading error data was different to Olson Lab's. We were interested in overall numbers of errors and particularly in errors that caused increases in reading times. These would indicate an increase in reading difficulty of the text being read. Therefore our classification scheme was simpler. We identified insertion errors and pause errors that both increase reading times, omission errors that decrease reading times and substitutions (miscues) where another word or mispronunciation replaces the target word. Our classification was similar to van Hasselt's (2002), but whereas her study measured numbers of errors, ours also measured times of errors.

Materials

The reading materials were reports generated by the GIRL system describing individuals' results in their literacy assessments. Reports received were one of two types (a) "easy" texts generated using readability preference rules, or (b) "hard" text generated using frequency rules. Figure 1 illustrates an example of each type of text.



[redacted],

READING SKILLS REPORT

Your scores were excellent in two tests, because you made no mistakes. You got full marks for word order and missing words.

For word order, you selected the correct sentences. You picked *Send me an application form*.

You got high marks for speed reading and form filling, that is, you made only one mistake in form filling. You did very well.

Sometimes in form filling, you did not find the right item. You did not click on the title: *Mrs*.

Your reading was good. Well done.

[redacted],

READING SKILLS REPORT

Your scores were excellent in two tests because you made no mistakes, you got full marks for word order and missing words.

For word order, you selected the correct sentences. You picked *Send me an application form*.

You got high marks for speed reading and form filling (you made only one mistake in form filling). You did very well.

Sometimes in form filling, you did not find the right item. You did not click on the title: *Mrs*.

Your reading was good. Well done.

Figure 1 Examples of automatically generated texts "easy" text on the left, "hard" text on the right

The readability rules were derived from our own experiments (Williams et al. 2003) and from psycholinguistic data. The system selects the most "readable" of possible alternatives for the discourse choices identified in the introduction. The overall effect is a preference for short, common cue phrases and short sentences, but only when it is "legal", e.g. it is not legal to have a sentence break between the antecedent and consequent of a conditional expression. Empirical data about legal ways to generate discourse choices were derived from a corpus analysis (Williams and Reiter 2003).

Method

Participants were seventeen poor readers and a control group of nineteen good readers; all were native British English speakers from sixteen years of age to over sixty. Some were members of the public who volunteered to take part in psychological experiments; others were enrolled on basic skills courses at a community college. Participants were classified into the two groups based on their performance in a skills-based literacy assessment. Four parts of the literacy assessment were administered, but classification was based on scores in only one part, a timed skimming and scanning test.

After each participant had completed the literacy test, the system generated a report about the participant's skills and he or she was recorded reading it aloud from a computer screen. The recordings were made digitally using a Sony lavalier lapel microphone (model ECM-44B). This is small, lightweight and unobtrusive. It was clipped onto a participant's clothing and placed as close to the throat as possible. The microphone was connected by a long lead to a laptop computer operated by the experimenter.

Analysis of speech recordings

Speech recordings were annotated by hand by the first author using CSLU's SpeechViewer software (Hosom et al. 1998). Each speech file was annotated with the text that was read, with the pauses at the ends of phrases and paragraphs and with any reading errors made. We classified and labelled the errors as:

- insertion errors
- pause errors
- omission errors
- substitutions

Insertion errors are spoken words or parts of words that were not in the text, for instance "sss...." before the word "sometimes". Pauses are extra pauses that were not between-paragraphs or end-of-phrases, often these occurred after insertions, e.g. "sss... [pause] sometimes" or as hesitations before longer words like "selected". Omissions occurred where a word or part of a word in the text had been missed out. These were only labelled when they were obvious. Sometimes if a person was speaking very quickly it was hard to decide whether a short function word, e.g. "of" had actually been voiced, or not, and these were not annotated. Nor did we include common reductions (Johnson 2002) such as "cos" for "because". Like Hulslander (2001) and Mostow et al. (2002) we allowed for some "sloppiness" and did not insist on a match to the citation form such as would be found in a pronunciation dictionary.

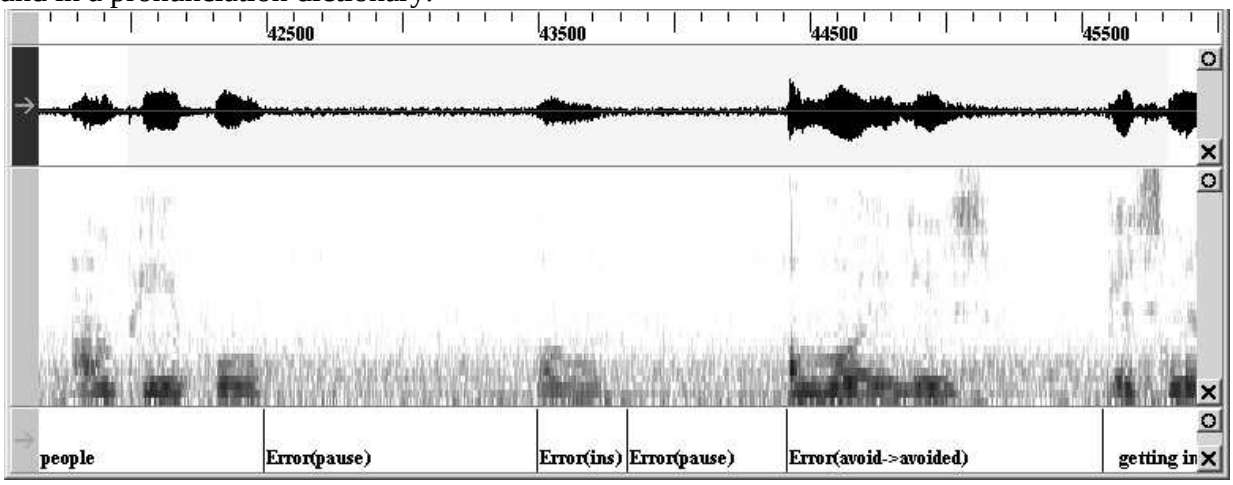


Figure 2 Part of a speech file showing labelled reading errors

Figure 2 shows part of a speech file labelled using SpeechView. At the top of the figure is a time scale in milliseconds. Below that is a section of the time waveform. The next window down is a frequency domain spectrograph. This was used in addition to the time wave as an aid in accurately

marking the beginnings and ends of sections. The tool enables the annotator to play aloud the sections between vertical markers to hear whether the markers have been positioned correctly. The bottom window is the annotation window. A pause error "Error(pause)" has been labelled after "people", followed by an insertion error, marked "Error(ins)", another pause error and a substitution error, marked "Error(avoid->avoided)".

We recorded the numbers of each type of error, the times in milliseconds for each error and the proportions of reading time spent making each type of error.

Results

For thirty-six readers reading texts averaging eighty-five words in length, we found a total of 167 errors.

- 21 substitution errors,
- 49 insertion errors,
- 96 pause errors,
- 1 omission.

Table 1 shows the results for poor readers. The first column on the left-hand side contains the type of error. The second column contains the metric, or the item measured (number of errors, time in milliseconds spent making the errors, and time as a proportion of total reading time). The third column has the type of text generated, as defined above. The fourth column is the number of participants (ten poor readers read the *easy* text and seven read the *hard* text). The fifth, sixth and seventh columns contain the means, standard deviations from the mean and standard errors, respectively.

Type of error	Metric	Condition (type of text)	N	Mean	Std. Deviation	Std. Error Mean
Insertion errors	number	easy	10	2.00	2.11	0.67
		hard	7	2.29	2.29	0.86
	time (ms)	easy	10	714.69	768.59	243.05
		hard	7	1179.57	1556.76	588.40
	proportion of total reading time	easy	10	0.02	0.02	0.01
		hard	7	0.02	0.03	0.01
Pause errors	number	easy	10	4.40	4.50	1.42
		hard	7	5.14	4.22	1.59
	time (ms)	easy	10	3032.20	3877.76	1226.25
		hard	7	2828.29	3041.77	1149.68
	proportion of total reading time	easy	10	0.07	0.08	0.02
		hard	7	0.06	0.06	0.02
Substitution errors	number	easy	10	0.40	0.52	0.16
		hard	7	2.29	2.82	1.06
	time (ms)	easy	10	313.90	431.80	136.55
		hard	7	1736.00	2268.22	857.31
	proportion of total reading time	easy	10	0.01	0.01	.00
		hard	7	0.03	0.04	.02
All errors	number	easy	10	6.80	6.37	2.01
		hard	7	9.71	8.48	3.20
	time (ms)	easy	10	4060.79	4686.747	1482.08
		hard	7	5743.86	6153.242	2325.70
	proportion of total reading time	easy	10	0.09	0.09	0.03
		hard	7	0.12	0.12	0.04

Table 1, reading error results for poor readers

For poor readers, the differences between the measures for errors on *hard* and *easy* text versions are largest for substitution errors. They made more substitution errors on *hard* texts. Differences for other types of error are not so large. Table 2 shows results for Levine's test for equality of variances for poor readers' substitution errors. The figures show that the there are

significant differences in the shapes of the distributions of substitution errors on the two types of text. Neither distribution is normal and, unfortunately, there is insufficient data to show significant differences in substitution errors made reading the two types of text using nonparametric tests.

POOR READERS	**Levine's Test for Equality of Variances**	
SUBSTITUTION ERRORS	**F**	**Sig.**
number	21.86	0.000
time (ms)	17.46	0.001
proportion of reading time	14.50	0.002

Table 2. Differences in distributions for substitution errors made by poor readers

Table 3 shows the results for good readers. It has a layout identical to table 2. For poor readers, the differences between the measures for errors on *hard* and *easy* text versions are largest for pause errors. They made more pause errors on *hard* texts. Differences for other types of error are, once again, not so large.

Type of error	**Metric**	**Condition (text)**	**N**	**Mean**	**Std. Deviation**	**Std. Error Mean**
Insertion errors	number	easy	9	0.33	0.50	0.17
		hard	10	1.00	0.94	0.30
	time (ms)	easy	9	231.33	476.45	158.82
		hard	10	338.40	438.68	138.72
	proportion of total reading time	easy	9	0.01	0.02	0.01
		hard	10	0.01	0.01	0.00
Pause errors	number	easy	9	0.44	0.73	0.24
		hard	10	1.20	1.14	0.36
	time (ms)	easy	9	101.22	165.42	55.14
		hard	10	432.50	513.81	162.48
	proportion of total reading time	easy	9	0.00	0.01	0.00
		hard	10	0.01	0.01	0.00
Substitution errors	number	easy	9	0.00	0.00	0.00
		hard	10	0.10	0.32	0.10
	time (ms)	easy	9	0.00	0.00	0.00
		hard	10	64.20	203.02	64.20
	proportion of total reading time	easy	9	0.00	0.00	0.00
		hard	10	0.00	0.01	0.00
All errors	number	easy	9	0.78	0.83	0.28
		hard	10	2.30	1.64	0.52
	time (ms)	easy	9	332.56	458.10	152.70
		hard	10	835.10	729.64	230.73
	proportion of total reading time	easy	9	0.01	0.02	0.01
		hard	10	0.03	0.02	0.01

Table 3, reading error results for good readers

Conclusions

- Overall, substitution errors (traditional miscues) turned out to give the best evidence that our hypotheses could be correct and that the system is indeed generating more readable texts.
- Poor readers made more substitution errors on *hard* texts, so the NLG system's rules for generating readable texts are working to some extent.
- The text version that was read made little difference to good readers. However, they were slightly more fluent (made fewer pause errors) on *easy* texts, indicating that perhaps the readability rules may help them too.The results indicate that the Natural Language Generation system has gone some way towards generating texts that are easy to read for poor readers. But we feel that further work is necessary to improve performance. **Future Work**

We will continue to investigate how to communicate with people who have poor reading skills in a new project, SkillSum. In this project, the research focus is on:

- How to generate language to motivate people to take up basic skills courses.
- How to generate language that is more readable.

A fair proportion of reading errors were due to clusters of consonants and vowels as Labov et al. found (1998). We will use this idea for improving lexical choice rules.

- Prefer words that are easy to "sound out" and pronounce.
- Prefer words that don't contain consonant and vowel clusters.

References

Basic Skills Agency, Cambridge Training and Development Ltd. and ASE (2001). Target Skills: Initial Assessment, version 1. CD published by Basic Skills Agency, 1-19 New Oxford Street, London.

Marilyn Binkley, Nancy Matheson, and Trevor Williams (1997). Working Paper: Adult Literacy: An International Perspective. National Center for Education Statistics (NCES) Electronic Catalog No. NCES 9733.

Janet Blaxall and Dale M.Willows (1984). Reading ability and text difficulty as influences on second graders' oral reading errors. Journal of Educational Psychology, Volume 76(2), pp. 330-34.

James Fogarty, Laura Dabbish, David Steck, and Jack Mostow (2001). Mining a database of reading mistakes: For what should an automated Reading Tutor listen? Tenth Artificial Intelligence in Education (AI-ED).

Clare van Hasselt (2002). Oral reading achievements, strategies and personal characteristics of New Zealand primary school students reading below normal expectation. National Education Monitoring Project (NEMP) Probe Study, http://nemp.otago.ac.nz

John-Paul Hosom, Mark Fanty, Pieter Vermeulen, Ben Serridge and Tim Carmel (1998). CSLU Toolkit. The Center for Spoken Language Understanding, Oregan Graduate Institute for Science and Technology.

Jacqueline Hulslander (e-mail 2001) General Information on coding oral reading errors. Olson Reading Lab., University of Colorado.

Keith Johnson (2002). Massive reduction in conversational American English. Proceedings of the Workshop on Spontaneous Speech: Data and Analysis, Tokyo.

W. Labov, B. Baker, L. Ross, M. Brown (1998). A Graphic-Phonemic Analysis of the Reading Errors of Inner City Children. www.ling.upenn.edu/~wlabov/home.html

J. Mostow, J.E. Beck, S.V. Winter, S. Wang, and B. Tobin (2002) Predicting oral reading miscues Seventh International Conference on Spoken Language Processing (ICSLP-02).

Sandra Williams, Ehud Reiter and Liesl Osman (2003). Experiments with discourse-level choices and readability. Proceedings of the 9th European Workshop on Natural Language Generation, Budapest, April 2003.

Sandra Williams and Ehud Reiter (2003). A corpus analysis of discourse relations for Natural Language Generation. Proceedings of Corpus Linguistics 2003, Lancaster University, March 2003.

Sandra Williams (in preparation). Natural language generation of discourse relations for different reading levels. Ph.D.

A. Young and P.G. Bowers (1995). Individual Difference and Text Determinants of Reading Fluency and Expressiveness. *J*ournal of Experimental Child Psychology, Volume 60, pp. 428-54.