

# **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

**Presentation preference:** Paper

## **Abstract**

We describe work in progress evaluating a natural language generation system that generates literacy assessment reports. Research is on generating more readable documents. We previously evaluated comprehension and reading speed. Here we investigate reading errors. Do modifications the system makes result in less errors? We present preliminary results. As expected, poor readers make more errors than good readers. The full paper will report on whether readers make significantly fewer errors on reports modified for readability.

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

## Introduction

This paper describes work in progress on the analysis of reading 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 and the focus of our research with GIRL was on generating more readable documents. This is necessary because about one fifth of the adult population of most developed countries has poor literacy (Binkley et al. 1997). 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.

Thirty-nine 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 that were generated by GIRL. Reports received were randomly either modified for readability, or unmodified. The experimental design was thus a two by two matrix of good and poor readers reading modified and unmodified texts.

## Related work

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 this project, 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 increases in reading times caused by the errors, both indicating 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 mispronunciations that maintain roughly the same reading rate.

## Materials

The reading materials were the reports generated by GIRL describing individuals’ results in their literacy assessments. Reports received were one of two types (a) modified by readability rules and (b) unmodified. 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 twenty-one good readers and seventeen poor readers; all were native British English speakers over sixteen years of age. Some were members of the public who volunteered to take part in psychological experiments, others were adults enrolled on basic skills courses at a community college. Participants were classified 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

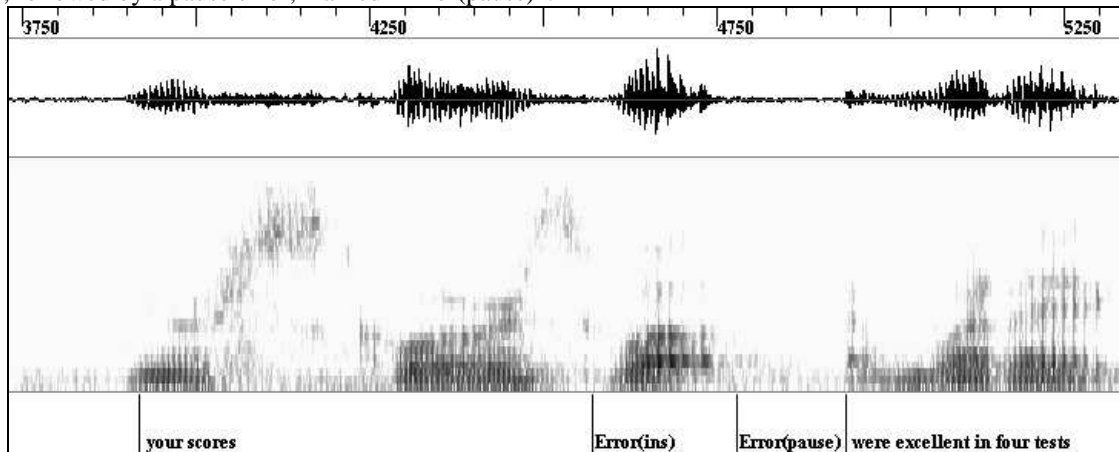
Speech recordings were annotated by hand by the first author using CSLU’s SpeechViewer software (Hosom et al. 1998). Each speech file was annotated with the text that was read, with the pauses at the ends of phrases and paragraphs and with any reading errors made. We classified and labelled the errors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

- insertion errors
- pauses

- omissions
- mispronunciations

Insertion errors are spoken words or parts of words that were not in the text, for instance “ing” added after the word “avoid”. Pauses are extra pauses that were not between-paragraph pauses or end-of-phrase pauses. These pauses often occurred as hesitations before longer words like “selected”. Omissions occurred where a word or part of a word in the text had been missed out, e.g. “ed” missed off the word “selected”. These were only labelled when they were obvious. Sometimes if a person was speaking very quickly it was hard to decide whether a short word had actually been voiced, or not, so these were not annotated. Mispronunciations were labelled when a substitution had been made that did not appear to affect overall reading time, e.g. “times” was mispronounced as “things” in “sometimes”.

The figure 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 where the reader has made two errors “your scores were [pause] were excellent in four tests”. 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. An insertion error “Error(ins)” has been labelled after “your scores”, followed by a pause error, marked “Error(pause)”.



### Preliminary results

Preliminary results for nineteen people’s recordings (eight good readers and eleven poor readers) have been analysed. These indicate that, as expected, poor readers make more errors than good readers ( $p=0.001$  in an independent samples t-test). The eight good readers made a total of 16 errors (7.6 seconds of pauses and insertions) while the eleven poor readers made a total of 102 errors (57.5 seconds of pauses and insertions). Only two omission errors were made, both by poor readers. Of the good readers, four received unmodified texts and four received modified texts, of the eleven poor readers, six received unmodified texts and five received modified texts. Both good readers and poor readers made more errors on the unmodified texts than on the modified texts. Good readers made a total of five errors on the modified texts, a mean of 0.4 seconds of errors per person, and eleven errors on the unmodified texts, a mean of 1.5 seconds of errors per person. Poor readers made a total of forty-six errors on the modified texts, a mean of 4.4 seconds of errors per person and fifty-six errors on the unmodified texts, a mean of 6.2 seconds of errors per person.

Standard deviations are large for data on modified texts vs. unmodified texts that have been analysed so far and statistics are not significant. We will analyse the remaining data to find out whether applying the readability rules does in fact make the modified texts significantly more readable for readers with poor literacy. We will present full results in the final paper.

### 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.
- 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).
- John-Paul Hosom, Mark Fanty, Pieter Vermeulen, Ben Serridge and Tim Carmel (1998). CSLU Toolkit. The Center for Spoken Language Understanding, Oregon Graduate Institute for Science and Technology.
- Jacqueline Hulslander (e-mail 2001) General Information on coding oral reading errors. Olson Reading Lab., University of Colorado.
- 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.