

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

Sandra Williams and Ehud Reiter ( [swilliam@csd.abdn.ac.uk](mailto:swilliam@csd.abdn.ac.uk), [ereiter@csd.abdn.ac.uk](mailto:ereiter@csd.abdn.ac.uk) )

## Abstract

Part of an evaluation of a natural language generation (NLG) system with real users, this study investigated whether readability rules used by the system were effective or not. We found that adults with poor reading skills made fewer substitution errors (miscues) reading aloud system output that had been generated when the rules were activated. So the rules seem to be at least partially successful.

## Aim

To find out if adults with poor reading skills reading texts output by a natural language generation system make fewer reading errors when the system uses readability preference rules.

## Background

- If a text is more readable, readers make fewer reading errors and disfluencies.
  - Oral reading errors increase as text difficulty increases (Blaxhall and Willows 1984, Young and Bowers 1995).
  - Poor readers' ability to recognise phrase boundaries decreases as text difficulty increases, but good readers' ability remains unaffected (Young and Bowers 1995).
- NLG system, GIRL (Sandra Williams' PhD project).
  - System generates a report after a literacy assessment.
  - System plans output using knowledge about discourse relations.
  - System uses readability preference rules, e.g. prefer solutions with:
    - Discourse cue phrases – to aid comprehension (Degand 1999).
    - Short, common cue phrases – more readable (Williams, Reiter and Osman 2003).
    - End-of-sentence punctuation – shorter sentences improve comprehension (Coleman 1962).
    - End-of-phrase punctuation.
    - Order with a statement first then related information.
  - Using the rules, the system would select (a) from the example below – "easy" version.
  - The system has rules base on frequency to generate an alternative text – "hard" version.

E.g. 2 possible outputs for a reason relation:

- your score is excellent, *because* you made no mistakes
- you made no mistakes *and therefore* your score is excellent

- In (a), cue phrase = *because*, positioned before 2<sup>nd</sup> text span, punctuation = comma and the order is statement, then reason.
- In (b), cue phrase = *and therefore*, positioned before 2<sup>nd</sup> text span, punctuation = none and the order is reason, then statement.

## Hypotheses

- 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.

## Participants

- 36 participants.
- 19 good readers, score = L2 in literacy test, 17 poor readers, score < L2.
- Aged 16 years to over 60 years.

## Method

### 1. Administer Literacy Test

**Sample Question**

You will see some adverts on screen for a short time only. What are the adverts about? (Select one)

bikes	<b>Bargain Buys</b>
furniture	Boy's Mountain Bike. Table and 4 chairs.
jobs	Jupiter Model / Black Diak Oak with Borai seats. £250 when condition £300 only. new. V811 accept
bargains	Tel. Wexford 63481 668439
	Three-piece suite, blue cotton covers – good condition. £250 only. Buyer must collect. Venetian blinds, two at 4ft wide by 5ft deep. New condition. £200 or nearest offer. Tel. 01332 460591 Tel. 01332 327178

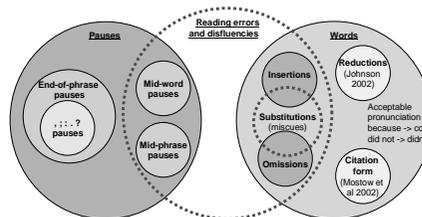
### 2. System Automatically Generates a Feedback Report (version chosen randomly)

<p><b>READING SKILLS REPORT</b></p> <p>Your scores were excellent in two tests, because you made no mistakes. You got full marks for word order and missing words.</p> <p>OR</p> <p>For word order, you selected the correct sentences. You picked them up in an application form.</p> <p>You got high marks for speed reading and form filling. That is, you made only one mistake in form filling. You did very well.</p> <p>Sometimes in form filling, you did not find the right item. You did not click on the title, for example.</p> <p>Your reading was good. Well done.</p> <p>easy version</p>	<p><b>READING SKILLS REPORT</b></p> <p>Your scores were excellent in two tests because you made no mistakes. You got full marks for word order and missing words.</p> <p>OR</p> <p>For word order, you selected the correct sentences. You picked them up in an application form.</p> <p>You got high marks for speed reading and form filling (you made only one mistake in form filling). You did very well.</p> <p>Sometimes in form filling, you did not find the right item. You did not click on the title, for example.</p> <p>Your reading was good. Well done.</p> <p>hard version</p>
--	---

### 3. Digitally Record Reading Aloud

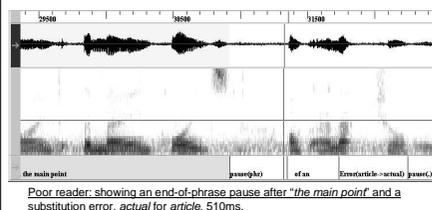
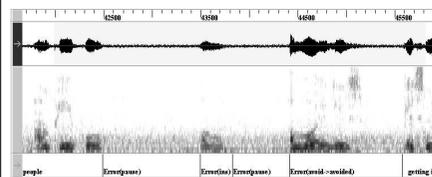
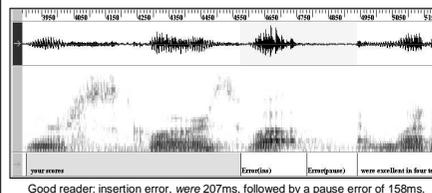


## Classifying reading errors



- Substitutions**
  - Miscues, mispronounced words (Labov et al 1998, Hulstander et al. 2002, van Hasselt 2002).
- Insertions**
  - Before- or mid-word, "struggles" (Hulstander et al. 2002).
  - Miscue (insertion) followed by correct word, i.e. self-corrections (Hulstander et al. 2002, van Hasselt 2002).
- Pauses**
  - Pauses before- or mid-word, i.e. reading disfluencies (Hulstander et al. 2002).
- Omissions**
  - Words or syllables missed out (van Hasselt 2002).

## Analysis



## Results Summary

- Poor readers**
  - More substitution errors reading *hard* texts than easy texts.
    - Means of 0.4 errors on easy texts and 2.3 errors on *hard* texts.
    - Significant variances in distribution ( $p < 0.001$  Levine's test).
    - But not enough data.
  - Insertion errors: no significant differences.
  - Pause errors (disfluencies) were longer reading *hard* texts than easy texts, but no significant differences.
  - Total errors (substitutions plus insertions plus pauses): no significant differences.
- Good readers**
  - Pause error times were longer on *hard* texts (mean 432ms) than on easy texts (mean 101ms), but no significant differences.
  - Substitution errors: no significant differences.
  - Insertion errors: no significant differences.
  - Total errors (substitutions plus insertions plus pauses): no significant differences.
- 167 errors found** (21 substitutions, 49 insertions, 96 pauses, 1 omission).

## 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.