Experiments with discourse-level choices and readability

†Sandra Williams, †Ehud Reiter and ‡Liesl Osman
†Department of Computing Science, ‡Department of Medicine and Therapeutics
University of Aberdeen
{swilliam,ereiter}@csd.abdn.ac.uk med078@abdn.ac.uk

Abstract

This paper reports on pilot experiments that are being used, together with corpus analysis, in the development of a Natural Language Generation (NLG) system, GIRL (Generator for Individual Reading Levels). GIRL generates reports for individuals after a literacy assessment.

We tested GIRL's output on adult learner readers and good readers. Our aim was to find out if choices the system makes at the discourse-level have an impact on readability. Our preliminary results indicate that such choices do indeed appear to be important for learner readers. These will be investigated further in future larger-scale experiments. Ultimately we intend to use the results to develop a mechanism that makes discourse-level choices that are appropriate for individuals' reading skills.

1. Introduction

The Generator for Individual Reading Levels (GIRL) project is developing a Natural Language Generation (NLG) system that generates feedback reports for adults after a web-based literacy assessment (Williams 2002).

The literacy assessment was designed by NFER-Nelson for the Target Skills application (2002) and it is aimed at adults with poor basic literacy. It produces a multi-level appraisal of a candidate's literacy skills. It tests eight skills: letter recognition, sentence completion, word ordering, form filling, punctuation and capitals, spelling, skimming and scanning and listening. The entire assessment consists of ninety questions, but the more difficult tests are only given to stronger candidates who have scored well on earlier tests. All questions are multiple choice. Figure 1 shows a screenshot of a typical question in the sentence completion test.

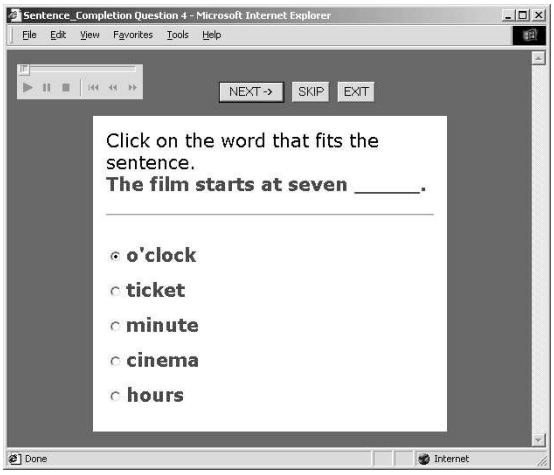

Figure 1. A screen shot of a literacy test question.

In our implementation, each question is assembled on-the-fly by a web server program which retrieves question data (question text, graphics, audio file and multiple-choice answers) from a database. As each question page is downloaded, an audio file of spoken instructions plays automatically. A candidate can play the instructions again by clicking on the audio player graphic.

The inputs to the NLG system, GIRL, are the answers a candidate gives to questions in the literacy assessment. GIRL currently generates a

feedback report after every test in the assessment. An example is shown in Figure 2.

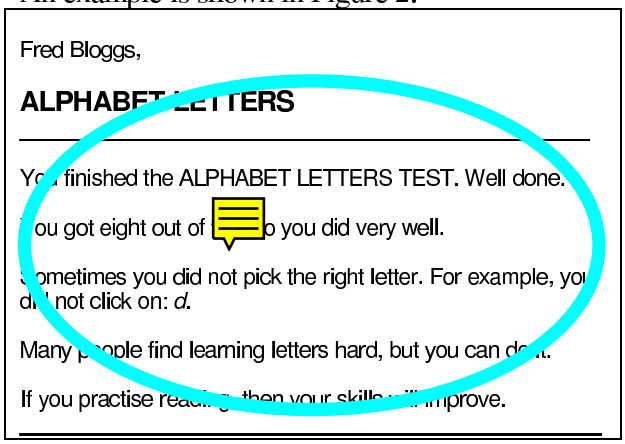

Figure 2. A type A "easy" report generated by GIRL.

GIRL is being developed with the goal of tailoring output texts to the individual reading skills of users (readers). In working towards this goal, we hope to find out more about generating documents for readers at different reading levels, how to test a system on real users, and how to implement reading level decision-making mechanisms as part of the generation process, and which decisions produce the most marked impact on the readability of the output texts.

It is very important to base the development of this system on solid empirical evidence rather than on our own intuitions. There has been very little empirical work on what kinds of texts are most appropriate for people with good reading skills and even less on what is appropriate for people with poor literacy. We were therefore motivated to do our own empirical studies. We are attacking the problem on two fronts: corpus analysis (Williams and Reiter 2003) and experiments with real readers, the subject of this paper.

1.1 Readability

Following Kintsch and Vipond (1979), we relate readability directly to readers' performance on the reading task (i.e. reading speed, ability to answer comprehension questions and ability to recall content). In these experiments, we measured reading speed and comprehension. We also analysed errors made in reading aloud, but that is not described here. The measures of readability we use are thus quantitative and are based on the hypotheses that readability increases or decreases with:

- an increase or decrease in the average reading rate for a particular reader;
- an increase or decrease in the number of correct answers given to comprehension questions.

1.2 Related work

We decided to investigate the impact of discourse-level choices as a novel approach to the problem of how to modify reports for different reading levels. Related work can be found in the PSET project (Devlin et al. 2000). PSET investigated how lexical-level choices and syntactic-level choices affect readability for aphasic readers, but it did not consider discourse choices.

As we mentioned above, there is very little empirical work to date on the impact of NLG system choices on readability and this is why it is so important for this project to carry out empirical work. One exception is the SPOT project (Walker et al. 2002). SPOT investigated which types of system outputs readers prefer. Readers were asked to rate the system's output utterances on understandability, well-formedness and appropriateness for the dialogue context. Apart from understandability, these do not relate to readability and we cannot assume that readers always choose the most readable utterances. They could be influenced by many other factors such as style. Also, all the judges were good readers and their preferences may not in any case be appropriate for learner readers.

2. The NLG system

A data-to-text NLG system like GIRL is able to linguistically express its output in a variety of ways that might affect readability. Here we look at features the microplanner can vary when linguistically realising discourse relations.

2.1 Deriving microplanner rules from a corpus analysis

Decisions about realising discourse relations are made in a module called the microplanner. The input is a tree of discourse relations joining pieces of information. It plans how the information from the tree will be ordered, whether it will be marked with discourse cue phrases (e.g. 'but',

'if' and 'for example'), and how it will be packed into sentences and punctuated.

To determine how human writers make these decisions for **good** readers, we carried out a corpus analysis (Williams and Reiter 2003). We used the RST Discourse Treebank Corpus (Carlson et al. 2002). The discourse relations analysed were *concession*, *condition*, *elaboration-additional*, *evaluation*, *example*, *reason* and *restatement*. The features we analysed are listed below.

- **Text span order.** The order of text spans in discourse relations. For instance "*because you got four out of ten, you need to practise*" or "*you need to practise because you got four out of ten* ".
- **Cue phrase existence and selection.** Whether cue phrases are present in a relation, or not, and which ones are used. For instance, cue phrases *if* and *then* are both present in "*if you practise, then you will improve*", but not in "*if you practise, you will improve*".
- **Cue phrase position.** The positions where cue phrases are located. For instance, for *example* is before the text span in *"for example, you did not click on the letter D"*, mid-span in *"you did not click, for example, on the letter D"*, and after it in *"you did not click on the letter D, for example"*. At present, GIRL can only handle positions *before* and *after*.
- **Existence and selection of between-span punctuation**. Sometimes there is punctuation between texts spans, e.g. the comma in *"many people find learning letters hard, but you can do it"* and the full stop in *"you finished the Alphabet Letters test. Well done"*, sometimes there is none e.g. *"many people find learning letters hard but you can do it"*.
- **First text span length.** The length of the first text span in words.

The inspiration for choosing the first four features was Moser and Moore's analysis (1996).

The features are interdependent. For instance, choosing a particular ordering of text spans can constrain the choice of cue phrase. For instance, the first span in a relation can never have cue phrase *but* (Williams and Reiter 2003). The corpus analysis was therefore extremely useful in deriving a set of rules for choosing legal combinations of features for each relation.

We hypothesised that certain values for features were more likely to increase readability. Commas between segments make the discourse structure more explicit. Sentence-breaking punctuation gives shorter sentences and selection of short, common cue phrases can help learner readers. Sentence length and word length are both believed to have a major impact on readability (Flesch 1949).

The corpus analysis results were input to machine learning algorithms to derive decision trees and rules for GIRL's microplanner. But the analysis they are based on was a corpus written for good readers and we need data to adapt them for learner readers, so we carried out the experiments described here.

2.2 Modifications for the experiments

For the experiments, the system was modified to generate much shorter, more restricted reports than those of the original GIRL system (Williams 2002). It was also modified to produce eight reports, one after each section of the literacy test, rather than a single long report after the entire test. These modifications increased our chances of collecting some reading data from each student, even if the student did not complete the entire literacy assessment (see section 4). Each short report consists of a salutation, a heading and exactly five paragraphs. Each paragraph consists of exactly one discourse relation. The system can produce two versions of each report, A and B (see Table 1).

In text types A and B, discourse relations were generated by varying only one discourse feature per paragraph. Table 1 shows which features were varied. Based on our corpus analysis results and on psycholinguistic evidence, we hypothesised that type A reports would be more readable ("easier") than type B reports ("harder").

Figure 2, shows an example of a type A report and Figure 3, a type B report. The text spans in the first paragraph are in statement:evaluation order in A and evaluation:statement in B. The second paragraph includes the cue phrase *so* in

A, and *therefore* in B. The third paragraph has *for example* before the second span in A and after it in B. The fourth paragraph has a comma present between text spans in A and no comma in B. Finally, the fifth paragraph has cue phrase *then* present in A, but not in B.

Fred Bloggs,

MISSING WORDS

Well done. You finished the MISSING WORDS TEST.

You got four out of fifteen, therefore you need to practise.

Sometimes you did not click on the right word. You did not pick: *recycle*, for example.

Many people find learning words hard but you can do it.

If you practise reading, your skills will improve.

Figure 3. A type B "harder" report generated by GIRL

Varying options in the manner shown in Table 1 ignores any crossover effects since there are actually 32 text types which would be possible from a 2x2x2x2x2 matrix of features. Also, any cumulative effects from having more than one discourse relation per paragraph are ignored in this design. We chose a simplified experimental design, since this was a pilot experiment. We wanted to test a large number of options and were aiming only to get indications of which options would have the greatest effects on readability. Future, more detailed experiments should look at crossover and cumulative effects.

3. Experiments

3.1 Participants

There were twenty-seven participants over the entire series of pilot experiments: twenty-one adults on basic skills literacy programmes (learner readers), four Ph.D. students and on other good readers. Because the design of the experiment was evolving, the conditions changed and we only use results from the final version. That is, nine learners and five good readers for reading speed and eleven learners for comprehension.

People who register for literacy courses are poor readers for a variety of reasons such as: missed school, learning difficulties, dyslexia, poor eyesight, poor hearing, short-term memory problems, or a combination of these. Personal data was recorded for each participant including age range, gender, first language, eyesight problems, hearing problems and any known reading problems (e.g. dyslexia). This data could be used to sub-classify readers, but the number of participants was too small to do this.

3.2 Method

Each participant underwent a web-based literacy assessment as described in the Introduction. After completion of each test in the assessment, GIRL generated feedback on how well the participant had done. The report is one of the two types described above and chosen by the system at random. In total, each participant was presented with between three and five reports of type A and three to five of type B.

Each participant was recorded reading his/her reports aloud, rather than recording silent reading times. This is because we discovered in an earlier pilot that following the more usual procedure of asking participants to read silently and then click a button led to erroneous reading times for learners (Williams 2002). We could not be certain whether they had actually 'read' the reports or not. Recordings provide evidence that reading has, in fact, occurred.

The recordings were made digitally and were annotated by hand by the first author using CSLU's SpeechViewer software (Hosom et al. 1998). The speech waveforms were annotated with beginnings and ends of words and pauses

Paragraph	Discourse relation	Feature Varied	Report type A "easier"	Report type B "harder"
1	evaluation	text span order	statement:evaluation	evaluation:statement
2	reason/result	cue phrase choice	"so"	"therefore"
3	example	cue phrase position	before segment	after segment
4	concession	comma between spans	comma	no comma
5	condition	existence of cue phrase	"if" and "then"	"if" only

Table 1. The discourse features varied in each paragraph in reports type A and B.

and with reading errors. Using the resulting annotation files, timings for each word, pause and paragraph could be calculated accurately (to within, say, 10ms). Only paragraph time and some pause times are used here

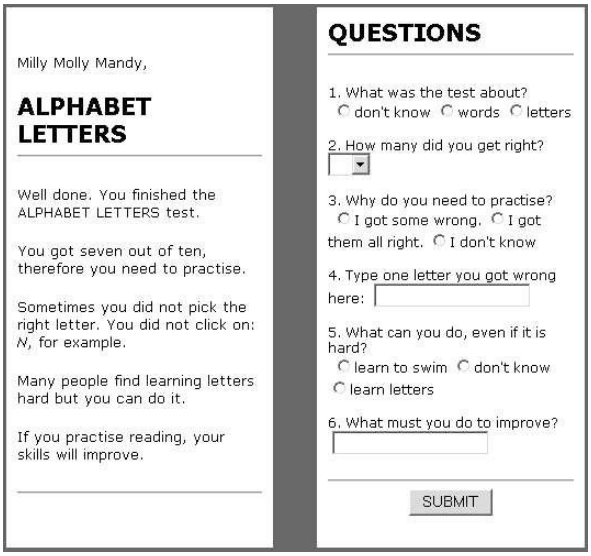

Figure 4. Comprehension questions are presented alongside a second view of the first report

After a participant had seen a screen showing his/her first report, had read aloud from that report and had been recorded, comprehension questions were presented (see Figure 4). The questions are displayed alongside a second view of the report and thus involved only comprehension, not recall. The experimenter read the questions aloud to learners, if necessary. Questions asked the meaning of information items in the report and of certain discourse relations. For example, question three is a 'why' question to determine if the reader has understood the relation in the second paragraph. Comprehension questions were administered only once, because an earlier pilot demonstrated that the meanings of each report are similar enough to prime readers.

On completion of the entire literacy assessment, an overall literacy level was calculated. This is subdivided into overall reading, overall writing and overall listening. Overall reading and overall writing are further subdivided into word focus, sentence focus and text focus scores.

If there was time after the experiment, informal chats with each participant provided useful information about readers' attitudes to the assessment and the reports. Participants offered ideas and suggestions for improvements. Basic skills tutors who were present during the pilots offered valuable suggestions.

4. Results

Not all learners managed to complete the literacy assessment because of time limits. Some learners and all good readers completed the test within an hour. Other learners took much longer, with one taking four hours! Two learners did not wish to be recorded reading aloud (although the majority of people were willing, sometimes even eager, to be recorded). Also, some recordings turned out to be too noisy. So we do not have a complete set of recordings for every person. For fourteen people, a maximum of 154 full text recordings were possible (720 paragraphs). We have good recordings of 297 paragraphs. All eleven learners completed the comprehension test.

4.1 Reading speed

Reading speeds were calculated in milliseconds per word (ms/word). Individuals, and particularly learner readers, vary a great deal in their reading aloud rates, so we calculated adjusted reading times for each person. The adjusted time is an individual's *raw time per word* for a single paragraph less that same person's *average time per word* over all of his/her recordings. In other words, adjusted times are a person's deviations from his/her average time. If the adjusted time is zero, then it is the same as that person's average time. A negative adjusted time means the person read faster and a positive adjusted time means they read slower. We were thus able to compare reading times for both versions of a paragraph, for all readers and calculate which version was read faster.

4.1.1 Paragraph 1: order of text spans

	statement :evaluation		evaluation :statement	
	#	adj.time	#	adj.time
Learners	26	-111.4	12	-47.1
Good readers	23	-37.3	16	-30.6

Table 2. Mean adjusted times in ms/word on two orderings, where # = number of samples.

Table 2 shows that learner readers read statement:evaluation order on average 64.3ms/word faster than evaluation:statement order. This result agrees with our RST Discourse Treebank corpus analysis (Williams and Reiter 2003)

where we found that the statement:evaluation order is far more common. In fact we found this ordering present in 86% of *evaluation* relations. There is little difference in times for good readers (6.7 ms/word).

4.1.2 Paragraph 2: cue phrase selection

Table 3 shows that learner readers read relations with the cue phrase *so* on average 90.8ms/word faster than those containing *therefore*. Good readers' times showed a small difference of only 6.4ms/word. It could be argued that these differences are due to the fact that *therefore* has more, and longer, syllables than *so*. However, if this were the reason, then the differences would be the same for both types of reader.

	"so"		"therefore"	
	#	adj.time	#	adj.time
Learners	20	-126.3	12	-35.5
Good readers	13	-60.8	11	-54.4

Table 3, Average adjusted times in ms/word for *so* and *therefore*, where # = number of samples

4.1.3 Paragraph 3: cue phrase position

	before 2nd span		after 2nd span	
	#	adj. time	#	adj. time
Learners	20	7.5	12	-25.0
Good readers	13	17.0	15	13.8

Table 4, Average adjusted times in ms/word with for *example* before or after the 2nd span, where # = number of samples

Table 4 shows that learner readers were on average 32.5ms/word faster when *for example* was positioned *after* the second span, compared to before it. Again, there is little difference in times for good readers (3.2ms/word). The result for learners was unexpected because we thought people would read faster when they were told in advance that the information they were about to read was going to be an example (i.e. when *for example* is before the second span). If it is after the span, they have to re-evaluate the information they have just read. Also, we found very few examples of the *after* position in our RST Discourse Treebank analysis (Williams and Reiter 2003). Scarcity of the easier-to-read version in the corpus may provide further evidence for Oberlander's theory (Oberlander 1998) that writers do not always 'do the right thing' for readers. This result will be investigated further in future experiments.

4.1.4 Paragraph 4: between-span comma

Table 5 shows that learner readers read this paragraph on average 26.8ms/word faster when a comma was present. This is what we expected. The comma between text spans indicates the discourse structure more explicitly and we would expect it to help learner readers. Once again there is little difference in times for good readers (5.6ms/word). Since the cue phrase is present before the second span, the comma may be redundant for good readers. Future experiments will investigate this.

	comma		no comma	
	#	adj.time	#	adj.time
Learners	20	-59.6	12	-32.8
Good readers	7	-64.3	9	-69.9

Table 5, Average adjusted times in ms/word for a between-span comma or no comma, where # = number of samples

4.1.5 Paragraph 5: presence of second cue

	then		no *then*	
	#	adj. time	#	adj. time
Learners	18	-8.3	12	-37.3
Good readers	7	-57.8	9	-51.0

Table 6, Average adjusted times in ms/word with and without cue phrase *then* where # = number of samples

Table 6 shows that learner readers read relations with no second cue phrase 29.0ms/word faster. Good readers showed little difference in times (6.8ms/word). This is not what we expected. We expected the second cue phrase to help learners because it makes the condition relation more explicit when both *if* and *then* are present. This result ties in with our corpus analysis (Williams and Reiter 2003) where few cases with both cues present were found. Writers do not often use both cue phrases and learner readers seem to find an extra phrase adds difficulty rather than helping.

4.1.6 Sentence length

The figures for reading times vs. sentence length show that all readers are slower on sentences above 23 words in length, and some learners are slower above 18 words. We require more data to verify this.

4.2 Comprehension

We found that learner readers can have problems with answering comprehension questions, even when the questions are administered verbally. Some learner readers are unfamiliar with reasoning about textual meanings. Some find it very hard to create answers using different words from

those that are present in the text they have just read. We therefore implemented a version of the system that generated more explanation for each paragraph (discourse relation), see Figure 5.

Fred Bloggs,

MISSING WORDS

Well done. You finished the MISSING WORDS TEST.

You got four out of fifteen. You made eleven mistakes. That means you need to practise.

Here is one you got wrong. You did not pick: *recycle*.

Many people find learning words hard. Perhaps you find it hard? You can do it.

The more you practise reading, the more your skills will improve.

Figure 5. 'Explanation' text

Figure 6 shows the comprehension scores for eleven learner readers. Scores, shown on the x-axis, are number of questions answered correctly out of six. Learner readers' scores are shown as gray bars and their mean scores are black bars.

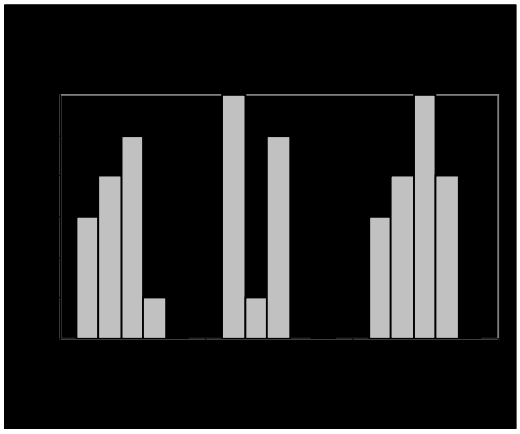

Figure 6. Grey bars = Learner readers' scores, Black bars = [illegible] learner readers' scores

Learners' highest mean comprehension score was on the *explanation* text type, but there is little difference between this and other mean scores. We simplified comprehension questions at the same time as introducing the explanation text, so we need to do further experiments to determine which has the most impact.

5. Conclusions

We previously analysed a corpus to determine how writers linguistically realise a number of discourse relations (Williams and Reiter 2003). Since the features analysed for each relation are interdependent (see section 2.2). Interdependencies can be conceptualised as a matrix like that shown in Table 7, where each cell (shown blank) actually contains rules (e.g. length vs. punctuation rules might be 1-10 words -> comma and >10 words -> full stop).

	length	order	punctuation	cue choice	cue position
length					
order					
punctuation					
cue choice					
cue position					

Table 7. Interdependencies of features for one discourse relation

The corpus analysis results were input to machine learning algorithms to derive decision trees and sets of rules for GIRL's microplanner, so that given input text spans of fixed lengths linked in a discourse relation tree, it can determine ordering, between-span punctuation, cue choice and cue position. Since the corpus analysis was based on a corpus written for good readers, we required data from experiments like this to adapt the rules for learner readers. To find out in detail how to adapt each cell of the matrix for each relation, we need more extensive experiments than these. Nevertheless, our pilot experiments are a good start. They enabled us to develop and refine our experimental method. Our preliminary reading speed results show:

- **Text span order.** Learners were slightly faster reading statement:evaluation order. Good readers' speeds showed only small differences.
- **Cue phrase choice.** Learners were faster reading relations containing *so* than those containing *therefore*. Good readers were also slightly faster reading *so*. This result was the only statistically significant one.
- **Cue phrase position** Learner readers were slightly faster when *for example* was positioned **after** the second segment. The position made very little difference to good readers.

- **Presence of punctuation.** Learner readers were slightly faster when there was a comma between discourse segments. This made very little difference to good readers.
- **Cue phrase existence.** Learner readers were slightly faster when *then* is not present. This made very little difference to good readers.

Sentence length and comprehension results require further investigation. The reading speed results indicate that discourse realisation choices make a greater impact on the reading speeds of learner readers than on those of good readers. This is an important first step in acquiring empirical evidence from real readers who have poor literacy skills. Discourse-level choices do indeed make a difference for these readers. This information is very valuable for the development of the GIRL NLG system.

We require more extensive, larger-scale experiment to derive rules appropriate for adapting our existing corpus-analysis-based models to individuals' reading skills. We need to know the impact of each feature on all the others. For instance (a) and (b), below, are almost equally likely according to our previous corpus analysis of the *condition* relation:

a) If you need help, ask your tutor.
b) Ask your tutor if you need help.

These experiments have shown that learners find *a* easier than *b*, since it has a comma. However, the order of text spans in *b* could be easier. We do not yet know how each feature affects the others and which has the most impact. Nor do we yet know if readability changes for a particular feature as the discourse relation changes. The choices seem deceptively simple, but their impact on people with poor literacy can be considerable.

Acknowledgements

This research is supported by an EPSRC studentship. We are grateful to: the tutors and students who took part in experiments at: Moray College, South Shields College, Southampton City College, Banff and Buchan College and the University of Aberdeen; CTAD Ltd. for their help and advice and for permission to use the Target Skills Literacy Assessment; Somayajulu Sripada and the EWNLG reviewers for comments and advice.

References

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2002. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In *Current Directions in Discourse and Dialogue*, Jan van Kuppevelt and Ronnie Smith (eds.), Kluwer Academic Publishers.

Siobhan Devlin, Yvonne Canning, John Tait, John Carroll, Guido Minnen and Darren Pearce. 2000. An AAC aid for aphasic people with reading difficulties. In *Proceeding of the 9th Biennial Conference of the International Society for Augmentative and Alternative Communication* (ISADC 2000). August 2000, Washington D.C.

Rudolph Flesch. 1949. *The Art of Readable Writing. Harper*. USA.

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

Walter Kintsch and Douglas Vipond. 1979. Reading Comprehension and Readability in Educational Practice and Psychological Theory. In L.G. Nilsson (ed.) *Perspectives on Memory Research.* Lawrence Erlbaum.

Megan Moser and Johanna Moore 1996 On the correlation of cues with discourse structure: results from a corpus study. Unpublished manuscript.

Jon Oberlander. 1998. Do the Right Thing ...but Expect the Unexpected. *Computational Linguistics.* 24, 3. 501-507

Target Skills Literacy Assessment. 2002. Published by the Basic Skills Agency, in association with Cambridge Training and Development Ltd. and NFER-Nelson Ltd.

Marilyn Walker, Owen Rambow and Monica Rogati. 2002. Training a sentence planner for spoken dialogue using boosting. In *Computer Speech and Language*. 16, 409-433.

Sandra Williams. 2002. Natural Language Generation of discourse connectives for different reading levels. In *Proceedings of the 5th Annual CLUK Research Colloquium.*

Sandra Williams and Ehud Reiter. 2003. A corpus analysis of discourse relations for Natural Language Generation. To appear in *Proceedings of Corpus Linguistics 2003*.