Generating readable texts for readers with low basic skills

Sandra Williams and Ehud Reiter

Department of Computing Science
University of Aberdeen
Aberdeen AB24 3UE, U.K.
{ swilliam,ereiter }@csd.abdn.ac.uk

Abstract

Most NLG systems generate texts for readers with good reading ability, but SkillSum adapts its output for readers with poor literacy. Evaluation with low-skilled readers confirms that SkillSum's knowledge-based microplanning choices enhance readability. We also discuss future readability improvements.

1 Introduction

Most existing NLG systems assume that generated texts are read by proficient readers with good literacy levels. However, many people in the UK and elsewhere are not proficient readers; indeed, according to a UK Government survey [Moser, 1999], twenty percent of the UK adult population have problems with reading (and an even greater number have problems with simple maths). Some of these individuals have physical or cognitive disabilities (such as dyslexia), but many have no such problems; their poor basic skills are because of factors such as social deprivation and attending low-quality schools. NLG systems that generate personalised health information, for example Cawsey et al. [2000] and Reiter et al. [2003a], would probably be more effective if they could generate appropriate texts for poor readers as well as good readers. Certainly real world NLG applications should at least consider such readers; otherwise there is a danger that many readers will not understand the texts we generate.

Generating appropriate texts for poor readers is a multifaceted problem. At a content level, texts should be short, explicit, and clearly useful to the reader [Sripada *et al.*, 2003], so that he or she is willing to make the effort required to read it. At a linguistic level, texts should use simple and easy-to-understand words and short sentences with simple syntactic structures [Harley, 2001]. At a presentation level, texts should have an easy-to-understand layout [Bouayad-Agha *et al.*, 2001] and be communicated in clear fonts particularly for dyslexic readers (e.g. K-type fonts, www.k-type.com) and for readers with visual impairment (e.g. tiresias font, www.tiresias.org).

The focus of our research is on the linguistic level, and to date we have looked at choices related to the expression of discourse structure, such as the order in which phrases re-lated by a discourse relation are expressed. Our hope was that rules for linguistic choices at least would be generic and easy to "plug in" to NLG systems intended for poor readers. Future work in the project will look at lexical choice and also at improved content selection and personalisation.

1.1 The SkillSum project

SkillSum is an on-going collaborative project between Cambridge Training and Development Ltd. (CTAD), who build educational resources, and NLG researchers at Aberdeen University. The project is developing a web-based application that assesses adult basic skills in literacy (reading and writing skills) or numeracy (maths skills) and generates feedback reports. Users of SkillSum take a test developed by CTAD that assesses their literacy or numeracy, and then SkillSum generates reports that summarise their performance. SkillSum is being developed in a user-centred manner involving rapid prototyping and frequent evaluations with users.

The ultimate goal of the SkillSum project is to build a system that allows people who are concerned about their literacy or numeracy to assess their skills with minimal support from others, and that encourages people with poor skills to take steps to improve them. Currently most people with poor skills do not in fact enrol in courses to improve their skills, and our hope is that making the assessment process as easy (and private) as possible will encourage more people who need help to seek it out.

The SkillSum project originally used detailed diagnostic literacy and numeracy assessments developed by CTAD. However, in pilots with users, we found that these took too long to complete and it seemed unlikely that people would be able to use them in an unsupported environment. Our current solution uses modified versions of CTAD's shorter literacy and numeracy screeners, i.e. tests that identify in a broader sense whether a user has problems with literacy or numeracy, but without a detailed analysis. These administer twenty-seven questions graded according to the Adult Basic Skills Core Curriculum for England and Wales [Steeds, 2001] and covering a broad range of skills from simpler levels to higher levels in this curriculum. The tests administer the easiest questions first. Ideally, the difficulty of the questions that are administered should change according to

a user's ability to answer correctly, but at present if a user has difficulties with the questions, the test simply ends.

The reports generated by SkillSum are, of course, tailored to individuals, but this tailoring is in terms of content, rather than language. The focus of research to date has been on how to generate appropriate texts for readers with below average literacy and numeracy; i.e. language tailoring for the group as a whole, not for individuals.

1.2 Related work

Using NLG in educational applications is not new, but the type of text generated is different from SkillSum's reports. For example, generating turns in intelligent tutoring system (ITS) dialogues (e.g. [Di Eugenio *et al.*, 2001; [Moore *et al.*, 2004]). Although turns can give feedback, the kind of feedback differs in that it attempts to teach a student about an immediate domain-specific learning problem, rather than to summarise his/her overall skills.

With regard to tailoring texts for different readers, a number of previous researchers have looked at tailoring generated texts according to whether the reader is a domain expert or a novice (for example [Paris, 1988; McKeown *et al.*, 1993; Milosavljevic and Oberlander, 1998]). Less work has been done on tailoring texts according to the reader's literacy. Perhaps the best-known previous work in this area is PSET [Devlin *et al.*, 1999], which focused on syntactic and lexical choices in texts intended for aphasic readers. Unfortunately most of PSET's adaptation rules were not experimentally validated. Siddharthan [2003] similarly proposed and implemented a system for simplifying texts, but did not evaluate how readable his generated texts were for poor readers. Scott and de Souza [1990] suggested some psycholinguistically-motivated rules for expressing discourse relations, but did not evaluate them at all.
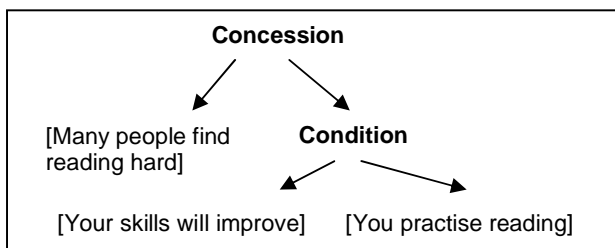
2 Linguistic choices investigated

Figure 1 – Extract from typical content plan

The document (content) planners of our system produce as output a tree, where core messages are related by discourse relations such as *explanation* or *concession*; this basically follows the architecture described by Reiter and Dale [2000]. Discourse relations are essentially rhetorical structure theory (RST) relations [Mann and Thompson, 1987], and messages are represented using a deep-syntactic representation, which is loosely based on RealPro [Lavoie and Rambow, 1997]. An example of an extract from a typical content plan, with messages shown as text glosses instead of deep syntactic structures, is shown in Figure 1.

Our focus to date has been on how discourse relations such as *Concession* and *Condition* in Figure 1 are expressed, in particular:

- *cue phrases*: should a cue phrase (or multiple cue phrases) be used to express a discourse relation? If so, which one(s)? For example, should we generate:
  - *If you practise reading, your skills will improve* (one cue, *If*)
  - *If you practise reading, then your skills will improve* (two cues, *If* and *then*)

- *ordering*: which order should the constituents related by a discourse relation be expressed in? Should the nucleus (core) be first or second? For example, should we generate:
  - *Your skills will improve if you practise reading* (nucleus first)
  - *If you practise reading, your skills will improve* (nucleus second)

- *punctuation* (sentence structure): should constituents be expressed in separate paragraphs, separate sentences, in a single sentence with punctuation separating them, or in a single sentence without punctuation? For example, should we generate (just showing two of these options):
  - *Many people find reading hard, but your skills will improve if you practise reading* (single sentence, comma separation)
  - *Many people find reading hard. But your skills will improve if you practise reading* (two sentences)

These choices are inter-dependent. For example, we cannot say, "*Then your skills will improve, if you practise reading*" (both *if* and *then* cue phrases, nucleus first).

This problem is related to the document structuring task of Power et al. [2002]. Power *et al.*'s approach is essentially algorithmic; whereas our approach is centred on the knowledge required to make the choices, and the algorithm used is less important. Their task is how to map an input RST tree to a set of output trees representing possible alternative document structures and then choose the best; whereas SkillSum's microplanning task is to map an input RST tree to flat, ordered lists of syntactic structures representing possible lists of alternative sentences and then pick the best. Power et al. include document layout in their task, whereas SkillSum makes layout decisions later, during the final realisation stage.

3 Choice rules and the microplanner

We created a microplanner (developed from the one described in [Williams, 2004]) that made the above choices based on hard constraints and optimisation rules; the hard

constraints forbade illegal combinations, and the optimisation rules expressed readability preferences.

3.1 Hard constraints

The hard constraints were intended to forbid combinations of choices that led to ungrammatical texts, such as "Then your skills will improve, if you practise reading". We created these by analyzing the RST Discourse Treebank Corpus (RST-DTC) [Carlson *et al*, 2002]; this is a corpus of Wall Street Journal texts that have been annotated with discourse relations. For each discourse relation of the type that occurs in SkillSum texts, we extracted 200 instances of the relation from the RST-DTC (or as many as possible if the corpus contained fewer than 200 instances), and analysed what combination of the above choices were instantiated in each of these instances. We then created hard constraints that forbade any pair of choices which was not present in any of the RST-DTC instances that we analysed. These constraints were specified on pairs of choices (for example, ordering and punctuation), not a complete choice set (ordering, punctuation, cue phrases), because we did not have enough instances to make rules for complete choice sets.

The RST-DTC corpus was not ideal for this exercise, as it is based on texts (Wall Street Journal articles) that are intended for good readers and written in U.S. English. It would have been preferable to use a corpus of U.K. English texts intended for low-skilled readers. Unfortunately there is no such corpus that includes discourse relation annotations.

3.2 Optimisation rules

The optimisation rules expressed preferences between legal sets of choices. We created two sets of rules: control and enhanced-readability (ER). The control rules were based on the most common choices observed in the RST-DTC; they also penalised cue phrases which were highly ambiguous (could be used for many discourse relations). The ER rules expressed a set of preferences for the above choices which we hypothesized would result in more readable texts for low-skilled readers.

The ER model was based on a literature review of relevant psycholinguistic findings (such as [Millis and Just, 1994; Degand *et al.*, 1999; Harley, 2001]) and also on a series of pilot experiments that we performed with low-skilled readers [Williams *et al.*, 2003]. Essentially, it prefers that

- each discourse relation should be expressed by a cue phrase. Only a single cue phrase should be used (for example, not both *if* and *then* for **condition**);

- lexically common cue phrases are preferred, even if they are ambiguous; for example *but* instead of *however* for **concession**;

- a cue phrase should be placed between the constituents if possible, and the nucleus (core) should come first if possible. For example, "*Your skills will improve if you practise reading*" is preferred over "*If you practise reading, your skills will improve*";

- constituents should preferably be in separate sentences; if they are in the same sentence, they should be separated by a comma.

With regard to the choice of cue phrases, obviously cue phrases do not have identical meanings; even if in a broad sense they express the same discourse relation, they have different connotations and applicability constraints [Knott, 1996]. Hence the choice of cue phrases should be influenced

Figure 2 – Example text produced by SkillSum and generated with enhanced readability (ER) model.

by how well the cue matches the context as well as how common (readable) it is. The issue is finessed in SkillSum by essentially hard-coding which cue phrases can be used to express a particular instance of a relation; this is clearly not a satisfactory long-term solution.

3.3 Microplanner

Our microplanner treats the decision-making problem as a constraint-satisfaction problem (CSP), using the above constraints and optimisation rules. It is in general terms similar to the CSP system that Power [2000] and Power *et al.* [2002] used to make expression choices about rhetorical structures, as discussed above. However, we make different choices. For example, our microplanner decides whether 0, 1, or 2 cue phrases should be used to express a discourse relation, whereas Power *et al.'s* assumes that a single cue phrase is always used. Power *et al.*'s microplanner makes choices about indentation, which our microplanner does not. Power *et al.* 's microplanner also attempts to optimize the document as a whole, whereas ours processes each relation separately, and does not consider interactions between the expressions of different relations.

Our microplanner is implemented in Java, and uses a Java Constraint Library (http://liawww.epfl.ch/JCL) for CSP representation and solving. While we have not explicitly measured the amount of computation time needed by the microplanner, certainly the SkillSum system as a whole produces texts within a few seconds, which is acceptable to users.

3.4 Example

An example of a text produced by SkillSum is shown in Figure 2. This text was generated using the ER model. Some examples of differences in expression resulting from using the ER model over the control model for the *same content* are

- *cue phrase cho*ice: the control model would choose uncommon, unambiguous *thus* to express **evaluation***;* whereas the ER model chooses common but ambiguous *and* (first paragraph in Figure 2);

- *ordering*: the control model would generate, *if you improve your skills, it could help you do a course* whereas the ER model generates *it could help you do a course, if you improve your skills* (sixth paragraph in Figure 2);

- *punctuation and sentence structure*: the control model would choose to express the first three paragraphs in Figure 2 joined by commas and as a single sentence; whereas the same information is expressed in three different sentences (and indeed three different paragraphs) by the ER model.

Some colleagues have asked why the ER model allows sentences to start with *And* (such as *And you did well* in Figure 2). This is because its hard constraints are based on corpus analysis, and sentences beginning with *And* do in fact occur in corpora.

4 Evaluation

4.1 Pilots

We conducted several pilot studies of the readability of SkillSum texts, generally in settings where subjects completed the literacy assessment and then were asked to read the summary texts that described their performance on the assessment. Perhaps the most important decision we made on the basis of these pilots was to evaluate readability by asking subjects to read texts aloud, and timing how long it took them to do so. We initially wanted to measure readability by asking comprehension questions and by timing silent reading, but this proved problematical. The problem with comprehension questions was that subjects responded to them based on their beliefs about their literacy, not on the content of the report. For example, if we showed a subject a text similar to Figure 2 and asked her "*What does the report say you did well on*", she might respond "*I didn't do well on anything*" instead of "*It says I did well on grammar*".

We also tried measuring reading rate using self-timed silent reading (subjects were asked to read text on a screen, and press a button when they finished), which is a common technique in psycholinguistics. Again, this did not work well with poor readers because they tended to skim-read or even simply press the button without reading the text at all.

Therefore we decided to ask subjects to read texts aloud, and measure their reading rate. This also allowed us to measure reading errors, which cannot be measured with silent reading. Oral reading rates and reading errors are commonly used by psychologists [Kintsch and Vipond 1979] and educationalists [ARCS, 2005] to measure reading difficulty.

Generally in pilots, the ER model texts were read faster than control model texts by poor readers but the increase in reading rate was not statistically significant. Good readers' reading rates were not affected. We used these results to calculate the sample size (number of subjects) required for the experiment in section 5.2, which did show a statistically significant increase in reading rate for poor readers.

Other findings of the pilots included:

- Texts needed to be short, no longer than the example shown in Figure 2. We tried giving people more detailed feedback about their assessments, but they did not wish to read longer texts.

- We got better results if we focused on subgroups with similar skills profiles, rather then trying to get subjects with a wide range of reading (dis)abilities. In part this is because low-skilled readers have very different ability profiles; a dyslexic is quite different from someone who never learnt to read because she missed school, for example. Also, from a statistical perspective, a varied subject group meant high standard deviations in reading speed, which made it difficult to obtain statistically significant results.

User preferences elicited in early pilots showed variation across good and poor readers. Some preferred the shorter

sentences and simpler cue phrases of the ER model, while others preferred the control model output, describing its texts as "more flowing".

4.2 Experiment

In this experiment, we focused specifically on people with moderately poor skills but not people with severe learning difficulties.

Goal. To test the readability of texts generated by SkillSum, our hypothesis was that participants would make fewer reading errors and have faster reading rates on a text generated using ER rules than on a text generated using control rules. We also took the opportunity to trial the latest version of SkillSum with real users.

Materials. To measure readability, we showed participants reports generated for someone else (not themselves), in order to de-personalise the experiment; in fact the reports used were the ER version shown in Figure 2 and text with the same content but generated using the control model. Colleagues have suggested that we could have shown a number of different texts tailored for different people, rather than showing everyone the same individual's texts. This is a good idea and we will try it in the future.

Participants. 60 students aged sixteen to twenty-eight years at a UK Further Education college (similar to an American community college), all of whom were enrolled in vocational courses (e.g. Hairdressing, Sport and Travel and Tourism). The participants were selected by staff from the college's basic skills department who knew their approximate levels of literacy and numeracy; they selected people who were known to have problems with basic skills, but did not have severe skills deficits.

Ideally the SkillSum application requires users who are not already enrolled at college. In practice, however, we had some difficulty in finding such people. Community workers we contacted are protective of the people they work with and wary about involving them in experiments. To get around this problem, we work with people who are enrolled in community or FE college courses where we have existing contacts and have built up trust. We feel that although this is not ideal, the people we work with do, in fact have problems with basic skills and are thus well-placed to test SkillSum; to comment on any difficulties they have with using it and to make suggestions for improvements.

Method. Each participant answered some questions about his/her background (e.g. age and college course) and completed one of the on-line SkillSum assessments; half did literacy and half did numeracy. SkillSum then generated a report and participants were asked (a) some simple questions about it to check comprehension (e.g. *"What was your score?"*) and (b) for their comments on SkillSum in general. Participants were then asked to read aloud the two texts; the texts were presented in random order. We recorded the speech digitally using a high-quality but unobtrusive lapel microphone.

Analysis of recordings: Recordings were analysed accurately using speech signal processing and annotation software to mark beginning and end of reading the entire text and the beginnings and ends of reading errors. Reading time and rate for the entire text was calculated as well as total time spent making reading errors.

We found a very strong practice effect on reading rate (the second text was almost always read significantly more quickly than the first) even when we attempted to factor this out using repeated measures statistics, so we decided to only use data from the first text read by each subject.

Following Sanders and Noordman [2000], we decided to exclude as outliers subjects whose reading speed was more than twice the standard deviation from the mean; in practice this meant that we excluded three very poor readers (reading speed less than 110 words/min) from the analysis. We also excluded six sets of recordings that were too noisy to analyse.

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. 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 only measured numbers of errors, ours measured error numbers and time spent making errors.

Results: Oral reading rate results for the remaining fifty-one people are shown in Table 1. The ER version was read on average 16 words per minute faster than the control text (9% faster) and this result is statistically significant at p<0.05.

Text	n	Mean oral reading rate (words/minute)	Sig. (indep. samp t-test)
control	25	173	0.040
ER	26	189	

Table 1 – Results for oral reading rate

An analysis of reading error times gave the results shown in Table 2. Types of errors made were typically substitution errors, where one word had been substituted for another e.g. *correct* for *contact*, pauses (hesitations) that were not at phrase boundaries and insertion errors, where readers had uttered words, or parts of words that were not present in the texts, e.g. *par* before *parliament*.

Text	n	Mean error time (milliseconds)	Sig. (indep. samp t-test)
control	25	1588	0.058
ER	26	874	

Table 2 – Results for reading error times

The table shows that subjects spent on average an extra 714ms making errors on the control text, that is they spent 82% more time making errors on the control text than on the ER text; this is weakly significant at p = 0.058.

5 Discussion and future work

Our evaluation experiment suggests that the ER choice rules at the discourse level do enhance the readability of generated texts for low-skilled and moderate-skilled readers. However, the effect is not as large as we believe SkillSum is capable of. We suspect this is partly because we only looked at a few linguistic choices, and in particular because we

have not yet included lexical choice. This will be addressed in our future work, see below. Our reviewers pointed out that layout could also affect readability. This is another factor to investigate in future work.

Another reason why the effect of our ER models on readability is not so large is because we are using only one choice model for people with a wide range of skills and (dis)abilities. Our experiences suggest that there are major differences between moderately-low-skilled readers (such as most subjects in our evaluation experiment) and very-low-skilled readers (such as the outlier subjects excluded from the experiment); and between people who are dyslexic, non-native speakers, or simply have not had the chance to learn to read well. More generally, adults with poor reading skills are often said to have "spiky" ability profiles; for example one person may have good vocabulary but poor grammar, and another may have the reverse. ARCS [2005] divides poor readers into 10 subgroups; building models for each of the ARCS subgroups would be one way of making our models less generic and more focused.

Ultimately, we would like to explore building choice models for *individuals*, that is groups of 1. Building such models of course requires spending considerable effort in acquiring data about individual readers. However, if such models are significantly more effective than generic or subgroup models in terms of generating readable texts, they may be worth exploring, since the benefits of making health information (for example) more accessible to people with limited reading skills could be very large.

5.1 Future work on lexical choice

Including lexical choice in our models will not be easy because true synonyms and paraphrases are rare [Edmonds and Hirst, 2002]. Hence we cannot simply select between N different lexicalisations that have exactly the same meaning, instead we have to determine which synonyms or paraphrases are appropriate given the text's content and the system's goals (as well as readability preferences).

For example, since many people have problems with the technical term *grammar*, in our pilots we tried paraphrasing *problems with grammar* as *problems writing good sentences*. However, many subjects interpreted the latter phrase as saying that they had problems with spelling (not grammar), and this confused students whose spelling was in fact fine. Hence this paraphrase is probably not appropriate, at least for students with poor grammar but good spelling. We initially avoided lexical choice because of this issue (except for choice of cue phrases, where this problem in fact arose, as mentioned in Section 3.2).

Recently, we carried out a small study on technical terms in English and maths (such as *grammar*) to find out what kinds of explanation basic skills students would understand. We elicited their own explanations of the terms as well as trying out uses of the terms themselves, paraphrases of the terms and illustrative examples of the terms. We asked basic skills students to think aloud while solving simple English and maths problems and then explained their errors to them using technical terms. In think-aloud, they varied a great deal:

- Some people did not use technical terms at all.
- Some people used technical terms inaccurately, (e.g. *spelling* or *capital letters* were described as *grammar*).
- Some people used technical terms correctly.

We concluded that it is dangerous to use technical terms when they are likely to be misunderstood. When explaining terms to people, however, they seemed to comprehend the illustrative examples best of all and we will try these in the next version of SkillSum.

5.2 Future work on content selection

Other future work will be to improve content selection and personalisation in SkillSum to help people understand their strengths and weaknesses. This poses a challenge because the topic of SkillSum's reports is very sensitive indeed, e.g. telling vulnerable people that they have problems with their literacy and/or numeracy can be hurtful! To date, one of our biggest difficulties has been to generate feedback for people who did not answer any of the questions correctly. What should SkillSum say to them? This is particularly difficult because SkillSum does not know what might have gone wrong during the test: perhaps there was a problem with using the computer or the mouse; or perhaps the user had difficulties with reading the screen because of poor eyesight; or perhaps he/she has severe learning difficulties.

Knowledge acquisition for SkillSum is difficult because the task of generating adult basic skills summary reports is novel, poorly-understood and complex. See Reiter *et al.* 2003b] for a detailed review of KA problems specific to NLG. Like the smoking cessation letters generated by the STOP system [Reiter *et al.*, 2003a], adult basic skills reports did not previously occur naturally. That is, tutors do not tend to write down feedback and advice for basic learners. A related text type is school reports. However, Education literature on writing school reports is unhelpful because adult basic learners have often had bad experiences with school and school reports in the past [FENTO, 2004].

We are currently involved in knowledge acquisition to derive improved content selection rules. We elicited tutor-authored reports for ten case studies in literacy and ten in numeracy. We gave tutors test results and a short user profile containing background material, e.g. age, gender, course enrolled in and ambitions (these were built using anonymised data from actual people who took part in earlier pilots and from [Swain *et al.*, 2004]). An analysis of the tutor-authored reports demonstrated that they have similarities in high-level content structures but individual author differences in lower-level content. Some issues that we are currently considering include:

- Should reports mention students' mistakes as well as their correct answers?
- Should reports congratulate students for doing well, when perhaps their performance was worse than normal, or, on the other hand, should they commiserate

with students when perhaps their performance was better than normal?

- Should reports refer to students' ambitions (e.g. in terms of qualifications or career)?

- How much advice should be given?

- How much motivational content should be included?

We need to reconcile what tutors tell us with what students tell us. Tutors tend to agree that reports should be encouraging, focus on positive aspects and not mention mistakes. On the other hand, when we talk to basic skills students, they often want to know what their mistakes were (it can be frustrating to score twenty-six out of twenty-seven and not know which one was wrong!).

Not knowing an individual and how much effort he/she has put into the test is another problem. Evaluative comments in a report such as *"this is very good"* are meaningless without such knowledge and, indeed they are meaningless without reference to some scale (but tutors have advised against mentioning the core curriculum scale as students are not familiar with it).

We propose to elicit self-assessments and some background information about users' ambitions from an initial questionnaire in SkillSum. The former might help with choice of evaluative comments. But combining the latter with advice and motivational content e.g. *"you may be able to get a plumbing qualification if you do a course to improve your English"* could be dangerous because the system has no way of knowing what an individual's potential might be and the resulting content could be highly inaccurate.

5.3 SkillSum final experiment

Lastly, we are planning our final experiment in which we will evaluate SkillSum to find out if it meets the commercial and research goals of the project partners. More specifically, these goals are:

- *Basic research*: Does SkillSum generate texts that can be read and understood by people who have some problems with literacy but not severe learning difficulties? In particular, is the ER linguistic choice model making the right microplanning choices for such readers? We will show low-skilled readers reports generated with SkillSum using different linguistic choice models, and see how the choice models affect readability.

- *Applied research*: Do NLG reports help students to understand their strengths and weaknesses, and whether their skills are adequate? We will test this with two versions of SkillSum (with and without the NLG component) with students who wish to do a course at an FE college; hence "whether skills are adequate" will be judged relative to the needs of the student's intended course.

6   Conclusion

SkillSum is an on-going project that is just starting to make some progress on generating readable texts for low-skilled readers and there is much more that can and should be done. Nevertheless, we would like to think that the choice preferences we used (Section 3.2) could be practically useful to people building systems that produce texts for low-skilled readers, and indeed systems that produce texts for the general public (since good readers don't seem to be affected by these choices, there is no harm in using low-skilled preferences for all readers). We also hope that our work encourages other researchers to think about this topic, as making information more accessible to low skill readers would have major benefits for society.

Acknowledgments

References

[Bouayad-Agha *et al.*, 2001] N. Bouayad-Agha, D. Scott, and R. Power. The influence of layout on the interpretation of referring expressions. *Multidisciplinary Approaches to Discourse.* L. Degand, Y. Bestgen, W. Spooren and L. van Waes (eds.), Stichting Neerlandistiek VU Amsterdam and Nodus Publikationen Münster, 2001.

[ARCS, 2005] Adult Reading Components Study. www.nifl.gov/readingprofiles/FT_ARCS.htm

[Carlson *et al*, 2002] Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In *Current Directions in Discourse and Dialogue*, J. van Kuppevelt and R. Smith eds., Kluwer Academic Publishers, 2002.

[Cawsey *et al*, 2000] Alison Cawsey, Ray Jones, and Janne Pearson. The Evaluation of a Personalised Information System for Patients with Cancer. *User Modeling and User-Adapted Interaction,* 10(1):47-72, 2000.

[Degand *et al*, 1999] L. Degand, N. Lefèvre, and Y. Bestgen. The impact of connectives and anaphoric expressions on expository discourse comprehension. *Document Design: Journal of Research and Problem Solving in Organizational Communication*, 1:39-51, 1999.

[Devlin *et al*, 1999] S. Devlin, J. Tait, Y.Canning, J. Carroll, G. Minnen and D. Pearce. The Application of Assistive Technology in Facilitating the Comprehension of Newspaper Text by Aphasic People. In C. Buhler and H. Knops (eds.) *Assistive Technology on the Threshold of the New Millennium.* IOS Press. 1999.

[Di Eugenio et al, 2001] Barbara Di Eugenio, Michael Glass, Michael J. Trolio and Susan Haller. Simple Natural Language Generation and Intelligent Tutoring Systems. *Proc. of Artificial Intelligence in Education*, 2001.

[Edmonds and Hirst. 2002] Philip Edmonds and Graeme Hirst. Near-Synonymy and Lexical Choice. *Computational Linguistics* 28(2):105-144, 2002.

[FENTO, 2004]. Further Education National Training Organisation (FENTO). Including Language, Literacy and Numeracy Learning in all Post-16 Education. Guidance on Curriculum and Methodology for generic initial teacher education programmes, www.nrdc.org.uk, 2004.

[Harley, 2001] Trevor Harley. *The psychology of language from data to theory*. Psychology Press Ltd, 2001.

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

[Knott 1996] Alistair Knott. *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. PhD Thesis, University of Edinburgh, 1996.

[Lavoie and Rambow, 1997] Benoit Lavoie and Owen Rambow. A Fast and Portable Realizer for Text Generation. *Proc. of ANLP-1997*, pages 265-268, 1997.

[Mann and Thompson, 1987] W. Mann and S. Thompson. Rhetorical Structure Theory: A Theory of Text Organization. Information Sciences Institute, Reprint Series no. ISI/RS-87-190, 1987.

[McKeown et al, 1993] Kathleen McKeown, Jacques Robin, and Michael Tanenblatt. Tailoring Lexical Choice to the User's Vocabulary in Multimedia Explanation Generation. *Proc of ACL*, pages 226-234, 1993.

[Millis and Just, 1994] Keith Millis and Marcel Just. The Influence of Connectives on Sentence Comprehension. *Journal of Memory and Language*. 33:128-147, 1994.

[Milosavljevic and Oberlander, 1998] Maria Milosavljevic and Jon Oberlander. Dynamic Hypertext Catalogues: Helping Users to Help Themselves. *Proc. of Hypertext 1998*, pages 123-131, 1998.

[Moore et al., 2004] Johanna D. Moore, Kaska Porayska-Pomsta, Sebastian Varges, and Claus Zinn, Generating Tutorial Feedback with Affect, in *Proc. of the 17th International Florida Artificial Intelligence Research Society Conference*, AAAI Press, 2004.

[Moser 1999]. Claus Moser. *Improving literacy and numeracy: a fresh start (report of working group)*. http://www.lifelonglearning.co.uk/mosergroup/. 1999

[Paris 1988] Cécile Paris. Tailoring Object Descriptions to a User's Level of Expertise. *Computational Linguistics* 14(3):64-78, 1988.

[Power, 2000] Richard Power: Planning texts by constraint satisfaction. *Proc. of COLING 2000*, pp. 642-648, 2000.

[Power et al, 2002] Richard Power, Donia Scott, and Nadjet Bouayad-Agha. Document Structure. *Computational Linguistics* 29(2):211-260, 2002.

[Reiter and Dale 2000]. Ehud Reiter and Robert Dale. *Building Natural Language Generation Systems*. Cambridge University Press, 2000.

[Reiter et al., 2003a]. Ehud Reiter, Roma Robertson, and Liesl Osman. Lessons from a Failure: Generating Tailored Smoking Cessation Letters. *Artificial Intelligence* 144:41-58, 2003.

[Reiter et al., 2003b]. E. Reiter, S. Sripada, and R. Robertson. Acquiring Correct Knowledge for Natural Language Generation. *Journal of Artificial Intelligence Research* **18**:491-516., 2003.

[Sanders and Noordman, 2000] Ted Sanders and Leo Noordman. The Role of Coherence Relations and Their Linguistic Markers in Text Processing. *Discourse Processes*, 29(1):37 – 60, 2000.

[Scott and de Souza, 1990] Donia Scott and Clarisse de Souza. Getting the Message Across in RST-Based Text Generation. In R Dale *et al.* (eds), *Current Research in Natural Language Generation*. Academic Press, 1990.

[Siddharthan 2003] Advaith Siddharthan. *Syntactic Simplification and Text Cohesion*. PhD Thesis, University of Cambridge Computing Lab, 2003.

[Sripada et al., 2003]. Somayajulu G. Sripada, Ehud Reiter, Jim Hunter and Jin Yu. Generating English Summaries of Time Series Data using the Gricean Maxims. Proc. of KDD 2003, pp. 187-196, 2003.

[Steeds, 2001] Steeds, Andrew (Ed.). Adult Literacy core curriculum including Spoken Communication. Produced by Cambridge Training and Development Ltd. on behalf of The Basic Skills Agency. ISBN 1-85990-127-1, 2001.

[Swain et al., 2004] Swain, Jon., Elizabeth Baker, Debbie Holder, Barbara Newmarch and Diana Coben. Beyond the daily application: Making numeracy teaching meaningful to adult learners. National Research and Development Center for adult literacy and numeracy. www.nrdc.org.uk. 2004.

[van Hasselt, 2002] van Hasselt, Clare. Oral Reading Achievements, Strategies and Personal Characteristics of New Zealand Primary School Students Reading Below Normal Expectation. Probe Study, National Education Monitoring Project (NEMP), University of Otago, New Zealand, 2002.

[Williams, 2004] Williams, Sandra, H. *Natural Language Generation of discourse relations for different reading levels*. PhD Thesis, University of Aberdeen, 2004.

[Williams et al. 2003] Sandra Williams, Ehud Reiter and Liesl Osman. Experiments with discourse-level choices and readability. *Proc. of the 9th European Workshop on Natural Language Generation*, pp. 127-134, 2003.