Natural Language Engineering **14** (4): 495–525. © 2008 Cambridge University Press
doi:10.1017/S1351324908004725 Printed in the United Kingdom

Generating basic skills reports for low-skilled readers[*]

SANDRA WILLIAMS[1] and EHUD REITER[2]
[1]*Department of Computing Science, The Open University, Milton Keynes MK7 6AA, UK and*
[2]*Department of Computing Science, University of Aberdeen, Aberdeen AB24 3UE, UK*
e-mail: s.h.williams@open.ac.uk,e.reiter@abdn.ac.uk

(*Received 20 March 2006; revised 8 December 2006; first published online 24 April 2008*)

Abstract

We describe SkillSum, a Natural Language Generation (NLG) system that generates a personalised feedback report for someone who has just completed a screening assessment of their basic literacy and numeracy skills. Because many SkillSum users have limited literacy, the generated reports must be easily comprehended by people with limited reading skills; this is the most novel aspect of SkillSum, and the focus of this paper. We used two approaches to maximise readability. First, for determining content and structure (document planning), we did not explicitly model readability, but rather followed a pragmatic approach of repeatedly revising content and structure following pilot experiments and interviews with domain experts. Second, for choosing linguistic expressions (microplanning), we attempted to formulate explicitly the choices that enhanced readability, using a constraints approach and preference rules; our constraints were based on corpus analysis and our preference rules were based on psycholinguistic findings. Evaluation of the SkillSum system was twofold: it compared the usefulness of NLG technology to that of canned text output, and it assessed the effectiveness of the readability model. Results showed that NLG was more effective than canned text at enhancing users' knowledge of their skills, and also suggested that the empirical 'revise based on experiments and interviews' approach made a substantial contribution to readability as well as our explicit psycholinguistically inspired models of readability choices.

1 Introduction

Most research in Natural Language Generation (NLG) assumes that people who read generated texts will have good reading skills, but many people do not; indeed in the UK, about one in five adults has a reading age of ten or less (Moser, 1999). We believe that tailoring generated texts for such people will make information more accessible and could have important social benefits. It is interesting scientifically

[*] Many thanks to our industrial collaborators at Cambridge Training and Development, to the literacy and numeracy tutors who helped us, to the people who agreed to be subjects in our experiments and to the colleges for allowing us to run experiments with their students. We also thank our colleagues in Aberdeen and Milton Keynes and the anonymous reviewers for their insightful comments and suggestions. This work was funded by PACCIT-LINK grant ESRC RES-328-25-0026.

because low-skilled readers are demanding users. In particular, they are sensitive to linguistic choices that many high-skilled readers would not even notice.

We investigated this problem in the context of SkillSum, an NLG system which generates feedback reports for people who have just completed a screening assessment of their basic literacy and/or numeracy skills. Our hope was that automating the report-generation process would make it easier and cheaper for people to assess the level of their skills, and to seek help if appropriate.

In very general terms, there are two approaches to the problem of generating readable texts for low-skilled readers.

- *Empirical.* Repeatedly try out a system with poor readers. Repeatedly modify it in accordance with advice from domain experts and results of pilot experiments with users.
- *Theory-driven.* Explicitly represent and model the characteristics of readable texts. Build an NLG system that constructs a text which is (near-)optimal under this explicit representation of readability.

We emphasised the empirical approach in the module that determines content and structure (the document planner). The document planning rules do not explicitly model readability, but they have been (repeatedly) modified to incorporate the results of pilot experiments with low-skilled readers and feedback from basic skills tutors. We emphasised the theory-driven approach in the NLG module that chooses linguistic expression (the microplanner), especially when deciding how to communicate discourse structures.

Obviously the theory-driven approach is more attractive in principle, since it is more elegant scientifically and also easier to generalise to other applications. Nevertheless, our evaluations of SkillSum suggest that empirical revision based on experts' advice and empirical experiments also made a substantial contribution. In other words, while explicitly modelling the readability impact of microplanning choices was useful in enhancing the readability of SkillSum texts, revising the content and structure of SkillSum based on advice and pilot experiments was absolutely essential for achieving our readability goals.

Content and structure may of course have more impact on readability than linguistic expression. But some of our most important linguistic choice rules, such as preferring short sentences, were suggested by experts and subjects in our revision exercises. Indeed, as explained in Section 6.1, although our goal was only to change content/structure rules (and not linguistic expression rules) during the revision process, this distinction was difficult to enforce in practice since the changes affected subsequent processing. It would have been more natural to modify all aspects of the system from our empirical work, and we suspect that this would have resulted in a set of linguistic expression rules which were as effective as our theoretically motivated rules. Perhaps this should not be surprising, since current psycholinguistic knowledge of the readability impact of different choices is inadequate.

In our evaluation, users who read SkillSum reports had a better understanding of how good their basic skills were compared to users who read baseline (canned text) output; so SkillSum achieved its application goal of helping users to understand their skills. However, the evaluation also suggested that in at least some cases, bad news (such as reports which suggested that the user's skills were worse than he or she had expected) should be conveyed by a person, not a computer. For this reason, we recommend that SkillSum should be used in contexts only where users can also talk to a human tutor; it is not appropriate to put SkillSum on the Web and let people use it from their homes or from a library, which was our original vision.

In the rest of this paper, we give background information about basic skills assessments and related work on generating easy-to-read texts; we describe SkillSum's document planner and how we revised it in accordance with pilot experiments; we describe SkillSum's microplanner and its explicit model of readability constraints and preferences; and we summarise the results of our evaluations.

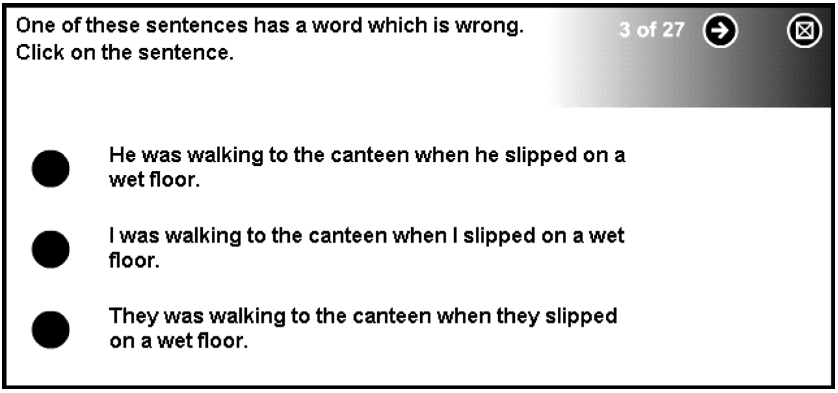
2 Background

2.1 Basic skills assessment

Poor adult literacy and numeracy is a major problem in most developed countries (Binkley, Matheson and Williams 1997). In the UK, the Moser study (Moser, 1999) reported that one in five adults is not functionally literate; for example, if given the alphabetical index to the Yellow Pages, they cannot locate the page reference for plumbers. One in four adults is not functionally numerate; for example, they cannot calculate how much change to expect from £1 when buying a 68p loaf of bread. Such people have difficulty finding and keeping jobs. Poor literacy and numeracy are a major cause of low productivity in the UK economy and also affect quality of life. Recognising these problems, the UK government launched the *Skills for Life* strategy, and is committed to raising the basic skills of 1,500,000 adults by 2007. Information and Communication Technology (ICT) is seen as a key element in these efforts.

The first step in improving an individual's basic skills is for that person to acknowledge that he or she may have a problem, and to come forward to have their level of literacy and numeracy assessed to give a clear picture of strengths, weaknesses and learning needs. Proper assessment requires him/her to complete a detailed assessment instrument, such as Cambridge Training and Development's *Target Skills: Initial Assessment* (http://www.targetskills.net). Such assessments must be taken in a formal setting, with the results analysed and explained by a basic skills tutor. They require a substantial time commitment on the part of the student, who must come to a scheduled session which may last several hours.

Because many people may initially be reluctant to make this time commitment, there is an increasing interest in short *screener* tests, which can be completed quickly and give a general indication of people's abilities and whether they should consider enrolling in a class to improve their skills. Screener tests are also useful for organisations, such as UK Further Education (FE) colleges (similar to American community colleges), to determine which incoming students require skills support. Screener tests should be as easy to take as possible – i.e. short, and available anywhere with minimal support. They are already on the Web, which makes them available wherever there is Internet access. Minimal support was the original goal of



One of these sentences has a word which is wrong.
Click on the sentence.

● He was walking to the canteen when he slipped on a wet floor.

● I was walking to the canteen when I slipped on a wet floor.

● They was walking to the canteen when they slipped on a wet floor.

Fig. 1. An example literacy screener question.

SkillSum; however, for vulnerable users we recommend that human tutors should be available to offer support.

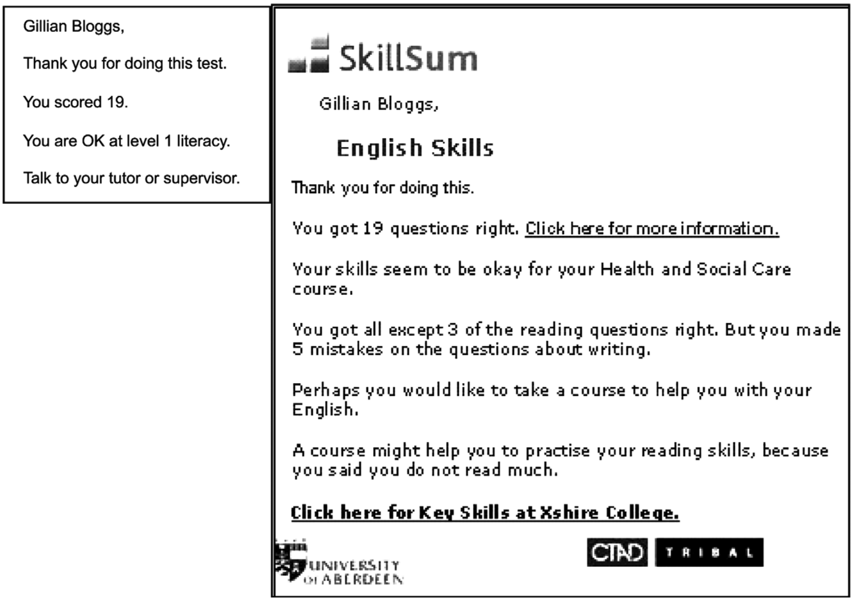
Screeners need to present their results to users in an easily understood manner. This was the main goal of SkillSum: to automatically generate a personalised report which summarises how well someone did on a basic skills screener, and which encourages them to complete a more detailed assessment, when appropriate, and to accept basic skills support.

2.2 SKILLSUM

SkillSum is a Web-based application which integrates basic skills testing and feedback report generation. Intended users of SkillSum are adults aged 16 years and over with low basic skills, but not with severe learning difficulties. Users test their literacy or numeracy by completing a short screener test consisting of at most twenty-seven multiple-choice questions. (Figure 1 shows an example literacy screener question.) Users are then shown a personalised report, which is generated by the SkillSum NLG system. Figure 2 shows an example of a SkillSum-generated report on the right-hand side.

SkillSum was a collaborative project between a commercial partner, Cambridge Training and Development Ltd. (CTAD), and NLG researchers at the University of Aberdeen. CTAD developed the basic skills testing module and Aberdeen the feedback report generator. The skills testing software was derived from an existing system that produced canned text reports, such as the one shown on the left in Figure 2; the generator was developed from a PhD project (Williams, 2004). Part of our evaluation was to assess the usefulness of generated reports compared to existing canned reports (see Section 6).

Figure 3 shows the architecture of SkillSum's NLG module. This follows the pipeline architecture with three sequential processes – document planning, microplanning and realisation – as used in many NLG systems and described in Reiter and Dale's book (2000).



Gillian Bloggs,

Thank you for doing this test.

You scored 19.

You are OK at level 1 literacy.

Talk to your tutor or supervisor.

SkillSum

Gillian Bloggs,

English Skills

Thank you for doing this.

You got 19 questions right. Click here for more information.

Your skills seem to be okay for your Health and Social Care course.

You got all except 3 of the reading questions right. But you made 5 mistakes on the questions about writing.

Perhaps you would like to take a course to help you with your English.

A course might help you to practise your reading skills, because you said you do not read much.

Click here for Key Skills at Xshire College.

Fig. 2. Example of CTAD's original canned output on the left and SkillSum generated output on the right (the user's name has been changed).

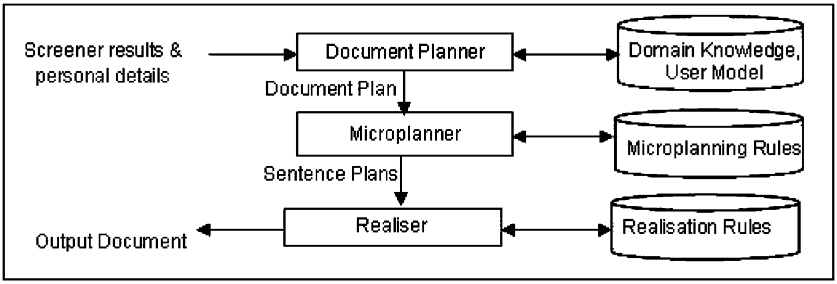


Fig. 3. Architecture of SkillSum's NLG module.

Document planning determines the content and discourse structure of the document. It produces a tree, in which core messages are related by discourse relations, such as *explanation* or *concession*, taken mostly from rhetorical structure theory (RST) (Mann and Thompson, 1987); i.e. the theory that rhetorical relations, such as *concession*, *condition* and *elaboration* connect statements in a document. In RST, rhetorical relationships are hierarchical and represented by rhetorical structure trees. The hierarchical nature of rhetorical relationships means that text spans in a rhetorical relation can be arbitrarily long, and that some text spans can themselves contain rhetorically related statements.

For example, Figure 4 shows an RST analysis of part of a human-written feedback report. The RST tree in Figure 4 shows a hierarchical arrangement of two discourse relations with *concession* at the root and *condition* at the next level down the tree. The paragraph is split into three discourse segments. Reading left to right, the second

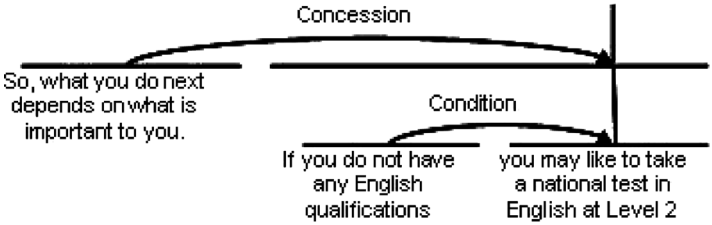

Concession
So, what you do next depends on what is important to you.
Condition
If you do not have any English qualifications
you may like to take a national test in English at Level 2

Fig. 4. RST analysis of an advice paragraph from an expert-authored report.

segment, *if you do not have…* forms the satellite of *condition*, and the third *you may like to take…* is the nucleus of *condition*; while the first segment, *so, what you do next…* forms the satellite of *concession* and the entire *condition* relation forms the nucleus of *concession*.

In SkillSum, older messages (leaves of an RST tree) from the original PhD system are represented as deep syntactic structures loosely based on RealPro (Lavoie and Rambow, 1997), but newer messages are represented as string-based templates (Reiter, Williams and Crichton 2005).

The microplanner chooses ordering, discourse connectives, aggregation, punctuation and lexical items. Microplanning is achieved by constraint satisfaction techniques followed by preference rules, and produces sentence specifications.

The final process, linguistic realisation, converts deep syntactic structures (or string-based templates) into English sentences. The realiser also adds hypertext links and final document formatting.

More detailed descriptions of the document planner and microplanner are given in Sections 4 and 5. The realiser is described in more detail in Williams (2004) and Reiter *et al.* (2005).

3 Related work

Zukerman and Pearl (1986) and Scott and Souza (1990) were among the first to propose that NLG systems should incorporate specific techniques to facilitate readability (particularly ease of comprehension). Scott and Souza (1990) suggested some psycholinguistically motivated rules for expressing discourse relations, particularly recommending the use of discourse connectives (short words and phrases, such as *and*, *for example* and *however*) which make the underlying rhetorical structure more explicit to the reader. They stressed that connectives would aid the comprehension of a document's rhetorical structure and hypothesised that readers would be 'unlikely to retrieve the rhetorical structure of a message unless it is stated explicitly' (p. 50). Although they did not test this experimentally, they proposed that discourse connectives should be generated whenever possible. We followed this advice in the development of our readability discourse model.

With regard to tailoring texts for different readers, a number of previous NLG systems tailor texts according to whether the reader is a domain expert or a novice, (Paris, 1988; Bateman and Paris, 1989; McKeown, Robin and Tanenblatt, 1993; Milosavljevic and Oberlander, 1998). Other systems tailor content according to users' likes and dislikes, e.g. the restaurant-recommender dialogue system of Walker *et al.*, (2003).

Few systems tailor output texts according to users' reading ability. Perhaps the best known is PSET (Practical Simplification of English Text) (Devlin, Canning, Tait, Carroll, Minnen and Pearce 2000), which parsed articles from *The Sunderland Echo* and simplified them for aphasic readers. The system substituted common words for uncommon ones, activised passive sentences, resolved references, and reduced multiple-clause sentences to single-clause sentences. Psychologists believe that all of these revisions assist aphasic readers; accordingly, we have taken up some of these ideas in SkillSum, even though aphasics may have slightly different problems from people who have never developed competence in reading.

Lack of detailed evaluation in PSET was a major limitation, since the psycholinguistic hypotheses that inspired its design were never fully tested with the application itself but only with manually prepared texts. One published evaluation was a pilot study where nine aphasic patients read original articles from *The Sunderland Echo* and manually simplified versions of the same articles. A comparison of performance on comprehension questions on the two kinds of article indicated that seven patients performed better on simplified texts (Devlin and Tait, 1998). A small pilot with six aphasic users found indications that manual anaphor resolution improved reading rate and comprehension (Canning, 2002). A larger evaluation with sixteen aphasics had the same finding (Canning, 2002). Unfortunately, no evaluations of the system itself were carried out with aphasic users.

Two text simplification systems are reported by Siddharthan (2002; 2003) and Chandrasekar and Srinivas (1997). Both reduced multiple-clause sentences to single-clause sentences. Siddharthan's system was aimed at poor readers, but not evaluated with them. Chandrasekar and Srinivas's system's intended users, on the other hand, were not human at all, but other Language Technology systems. The aim was to simplify texts before they were supplied as input to the parser of a Natural Language Understanding system, as a pre-parsing process. Okumura (2000) devised a revision system for enhancing the readability of concatenated extracts produced by an automatic text summarisation system. It is unclear which types of revision were actually implemented, but at least some resembled those implemented in PSET. Some limited evaluation of readability using human judges was also attempted. Inui, Fujita, Takahashi, Tetsuro, Iida and Iwakura (2003) proposed simplifying texts for deaf people by a combination of statistical and rule-based approaches. So far as we are aware, the system did not reach a stage where it could be evaluated with users.

To summarise, a number of algorithms for text simplification have been proposed and at least partially implemented. They work by simplifying human-authored texts and applying rules based on psycholinguistic ideas about readability. Unfortunately, there has been little evaluation of these algorithms with realistic user groups.

Other work on language technology and readability includes the REAP project, which used a language modelling approach to predict readability of short texts (Collins-Thompson and Callan, 2004). This technique proved as good as standard readability calculators, such as Flesh–Kincaid. Such language modelling could potentially provide an alternative knowledge source for generating readable texts.

The REAP system also retrieves documents for personalised graded reading practice (Brown and Eskenazi, 2005) using estimates of users' vocabulary based on word histograms derived from data on documents that they have read and words that they know. Another system (Eddy, 2002) selected microplanner solutions according to readability criteria (but the research interest was document style, not reading age).

SkillSum's application domain is education. Other NLG applications in this domain are intelligent tutoring systems, e.g. (Moore, Porayska-Pomsta, Varges, and Zinn, 2004) and (Di Eugenio, Glass, Trolio and Haller, 2001) but these are interactive dialogue systems that address students' immediate difficulties with a task, whereas SkillSum summarises students' overall skills.

SkillSum incorporates psycholinguistic evidence on readability. One of the strongest findings is that short, common words are easier to read (Harley, 2001). Other relevant findings are that short sentences are more readable (Coleman, 1962), and that including discourse connectives improves comprehension (Degand, Lefèvre and Bestgen, 1999), (Leijten and van Waes, 2001) and (Sanders and Noordman, 2000). All these findings have been implemented in SkillSum. SkillSum chooses discourse connectives from discourse relation data only. It does not use rhetorical features such as those proposed by Knott (1996), Knott and Sanders (1998) and Miltsakaki, Dinesh, Prasad, Joshi and Webber (2005), among others.

4 Determining document content: the document planner

The communicative purpose of basic skills summaries is to help people understand their strengths and weaknesses and encourage them (if necessary) to get help. We faced a number of general problems in adapting SkillSum's document planner to achieve this purpose.

(1) *Corpus.* We did not have a naturally occurring corpus of human-written summaries, since tutors normally give feedback orally. We briefly considered creating and analysing a corpus of school reports, but it would have been difficult to acquire such a corpus; we also suspected that adult basic skills learners might react badly to anything which reminded them of school, since many of them had bad experiences there (Hunter and Howard, 2004). Consequently, we collected our own corpus of expert-authored texts (Section 4.2).

(2) *User Modelling*. We lacked detailed knowledge about users. In many cases it would have helped immensely to know more about users' backgrounds, motivations, specific skills abilities and deficits, and so forth, but we were limited to information from the literacy or numeracy screener and from a short questionnaire. With regard to motivation in particular, a Masters student working with us found that it was hard to include effective motivational information in SkillSum texts without better information about users (Tintarev, 2004); hence we decided to include little motivational material in SkillSum reports.

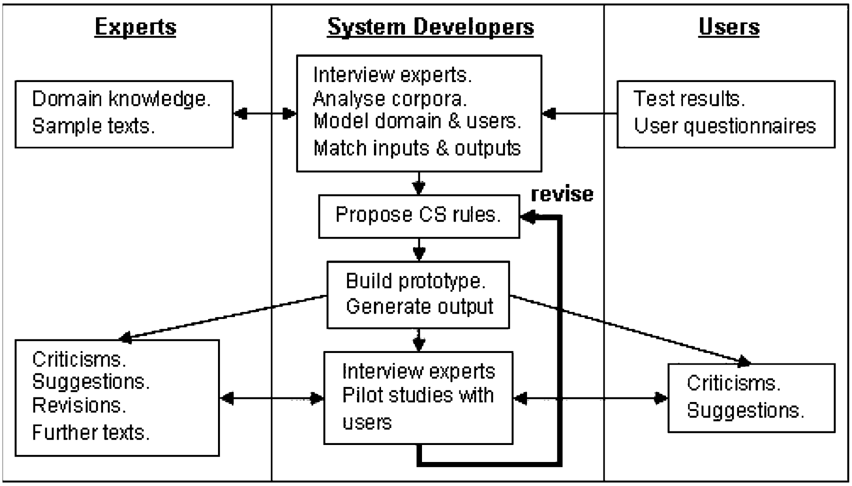


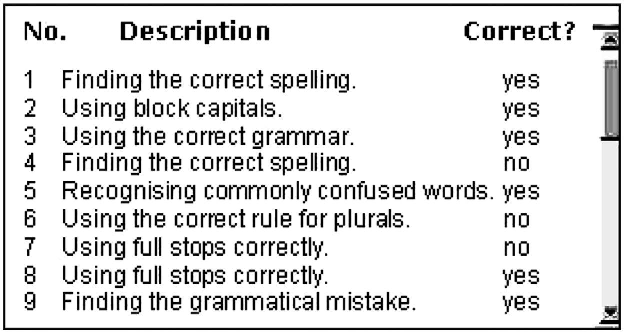
Fig. 5. Flowchart illustrating our methodology for deriving content selection (CS) rules.

(3) *Risks in communicating bad news.* The cost of getting content wrong could be high, since basic skills is a sensitive topic – telling people with low self-confidence that they have problems with literacy and numeracy might obviously be hurtful. For some NLG applications, if the content is wrong, mistakes of this kind might not matter much, but in SkillSum, inappropriate content might easily anger or upset users and discourage them from improving their skills. Hence we regularly produced and evaluated prototype systems, and used feedback from user questionnaires and interviews to improve document content (Section 4.4).

(4) *Users' attitudes and technical problems.* A final problem was that we did not know how seriously people took the literacy and numeracy screener tests, nor whether there had been technical problems during the test. This meant that when users answered only one or two questions correctly, we could not tell whether they genuinely had skills problems, or whether they had just clicked randomly on answers because they were chatting to a friend, or whether they had had computer or network problems. Therefore reports for such users are very short and simply advise talking to a tutor.

4.1 Methodology for deriving document content

Our methodology for deriving document content is summarised in Figure 5. We started off with knowledge acquisition and creating and analysing a corpus. We then created domain and user models, along with a model of the kinds of messages to be included in the generated texts. Next we entered an iterative process in which we proposed a set of specific content rules, developed a prototype system, generated some reports, asked experts and users to evaluate these reports, and identified ways to improve the system. We repeated this process six times. Our goal was not just to mimic experts, but also to consider the requirements of users and to make



No.	Description	Correct?
1	Finding the correct spelling.	yes
2	Using block capitals.	yes
3	Using the correct grammar.	yes
4	Finding the correct spelling.	no
5	Recognising commonly confused words.	yes
6	Using the correct rule for plurals.	no
7	Using full stops correctly.	no
8	Using full stops correctly.	yes
9	Finding the grammatical mistake.	yes

Fig. 6. Pop-up window giving more details about screener performance.

reports useful and relevant to them. For further details, see Williams and Reiter (2005).

4.2 Acquiring knowledge about content from an expert-authored corpus

We asked two experts (basic skills tutors) to write basic skills summaries from nine case studies in literacy and nine in numeracy. An alternative method would have been to record tutor–student feedback sessions, but because of their sensitive and confidential nature, we preferred a method that was less intrusive; we also wanted the experts to consider the issues involved in producing written reports. We gave them test results and short user profiles which were built with anonymised data from people who took part in pilots.

The resulting expert-authored corpus was small, and suffered from two problems: data sparsity (i.e. corpus texts covered only a small fraction of the possible permutations of inputs to the system), and disagreements amongst experts about content. Because our experts had never tackled the task of writing this kind of document, and disagreed with each other about it, we did not regard the corpus as a gold standard for the kind of content that SkillSum should generate or be evaluated against, but rather as initial suggestions to be discussed and revised.

We interviewed experts and asked them to criticise the generated reports, as detailed above. When experts disagreed with each other, we discussed this with them to decide on the best solution. An example was the inclusion of material about individual screener questions. Experts did not want to tell students about individual questions (and some students did not want such details), but many students told us that they did want to know exactly which questions they got right and wrong. It can be frustrating to score twenty-six out of twenty-seven and not know which one was wrong! Our solution was to add this information in a pop-up window (see Figure 6) which students could look at if they wanted to. This was not an ideal solution as the question descriptions are only brief summaries rather than complete questions. However, during evaluation, students were able to ask tutors for further explanation if they could not remember which questions the summary referred to. Finding that experts disagree is not new and has been discussed by many authors, for example, Reiter, Sripada and Robertson (2003).

Table 1. *Human-authored corpus report divided into sections*

Section	Sample text
Initial	Thanks for doing this.
Summary	You answered 15 questions correctly.
Diagnosis	You only made mistakes on a couple of questions where you had to read. You said that you like reading – so that does not seem to be a problem for you at all. Do you agree? The mistakes you did make were more to do with writing. It may be that you would like to improve your spelling and punctuation.
Advice	What you do next depends on what is important to you. If you do not have any English qualifications you may like to prepare for the national test in English at Level 1.

4.3 Deriving document content rules

Our analysis of the expert-authored corpus showed that texts were similar in high-level content structures but dissimilar in lower level detail. Our high-level analysis essentially followed the methodology of Geldof (2003). Most reports included an initial section (Initial), a summary of results (Summary), an interpretation of the results (Diagnosis) and advice on what to do next (Advice). Table 1 shows a sample from the expert-authored corpus, broken into sections.

Most expert-authored reports followed a similar basic structure. Sometimes ‘thanks’ in the initial section was omitted. The summary section was always present. The diagnosis section was not present in reports for students who had answered fewer than five questions correctly. As the overall score increased, the length of the diagnosis and advice sections also increased, and sometimes these sections were interleaved (‘Diagnosis, Advice, Diagnosis, Advice’ and so on). We had previously acquired 1,500 sets of test results (i.e. sample input data for SkillSum); this gave us an idea of the range of inputs that SkillSum needed to cover. However, because the corpus was small and data were sparse, we had to manually extrapolate content rules derived from the corpus to account for cases that it did not cover. Rules are expressed as *if-then* rules and sometimes, for instance, *if-then* data might be present but the *else* data might be missing from the corpus. Extrapolating a rule would mean supplying the missing part. For example, ‘you should have the reading skills to be able to cope with your sports course’ was present (one occurrence) in the corpus, but there were no data about what to say when skills were inadequate. We asked experts what to say in these cases and extended the rules to include any content that they suggested. For this particular rule, we revised it by adding an *else* part. The entire rule follows.

```
IF
    The user is to begin a Level 1 course at college
    AND
    his/her English skills are at least Level 1,
THEN
    Add content to advise that his/her skills are adequate for
    his/her course.
```

```
ELSE,
    Add content to advise that his/her skills are inadequate
    IF
        he/she is not already receiving help with basic skills,
    THEN
        add content that he/she should try to improve his/her
        skills, e.g. by taking a course.
```

The above rule incorporates domain knowledge about courses and levels of skill, knowledge about the user (i.e. what course the user is about to take and whether he/she is receiving help with basic skills), and expert knowledge (i.e. to advise the student to take up a basic skills course). This turned out to be one of the most important rules (important in the sense that it is deployed in the generation of every report where the user's course and the level of the course are known) even though only part of it actually occurred in the corpus.

To create actual content rules, we needed to convert general rules, such as *Add content to advise that his/her skills are adequate for his/her course* into rules which added specific messages, such as a representation of 'Your skills seem to be okay for your Health and Social Care course'. We did this by finding such messages in the corpus, and creating templates based on them. The templates in some cases were RST trees (see Section 2.3) which combined several messages.

4.4 Revision

We revised the system by piloting different versions of the SkillSum system as a whole (i.e. both the basic skills screener and report generator), holding discussions with experts and making modifications. We revised the document planner and the lexical selection part of the microplanner, but not the discourse-level planner containing the control and readability models (described in Sections 5.1, 5.2 and 5.3) which remained unchanged throughout. We also modified the basic skills assessment and the presentation format (e.g. by experimenting with hypertext). All pilots used the latest version of the system and after each pilot, revisions were fully implemented before the next pilot. In the following, we focus on three pilots that had the greatest influence in shaping the system. Different colleges and different participants were used in each pilot.

4.4.1 Pilot experiment, April 2004

Participants: Eight 16–19-year olds, four males and four females, attending a course to support them in their search for jobs and improve their basic skills. All were computer-literate, but with poor literacy and/or numeracy.

Method: Each participant was asked to take CTAD's long *Target Skills* assessment of over eighty questions and to read their own report generated by the system. Each was then tape-recorded during free recall comprehension and interviewed about the content and relevance of the report.

Results: At this time, reports were much longer, around 820 words, with full details of the long assessment. Recall was minimal, all participants remembered their overall score, but only three remembered other details. Even assuming that each sentence only contained one item of information, recall was only 1 to 5 per cent. Half of the participants commented that they found the report hard to read; the rest said it was 'easy' or 'fine', but could not explain what some of the terms meant. Two people said that they did not normally read anything much longer than titles of TV programmes and that they found the length of the reports daunting.

Discussion: Free recall results were poor. Typically in free recall experiments, the results vary enormously, for example in an experiment with good readers, participants recalled 13 to 18 per cent of the information (Lorch and Lorch, 1996), whilst in another experiment, readers of average skill recalled 43 to 51 per cent (Mason and Morris, 2000). Variations in recall depend on many factors, an obvious one being text length.[1] In light of these, we did not expect high recall results, but even so, they were strikingly low. Most participants had poor communications skills, which meant that interviewing was difficult. Half of the participants overestimated their own reading abilities.

Revisions: Following interviews with tutors and basic-skills experts, SkillSum was modified to use CTAD's shorter *screener* test. Our hope was that the screener could be used in an unsupported environment more successfully than the Target Skills assessment, which took too long to complete without guidance. Inputs to the NLG system therefore changed significantly, and we had to make corresponding changes to the content selection rules.

At the suggestion of experts, we also shortened reports and simplified the language by removing technical terms (e.g. *subject and verb agreement*, *pronouns* and *critical reading*); we added more personal pronouns to address users more directly; and we introduced short lists of motivational activities from the basic skills curriculum (Steeds, 2001) that were related to overall screener score, and were examples of what a typical person might be able to do at that level and what they might attempt if they progressed to the next level, e.g. 'Write a letter to a friend'.

4.4.2 Pilot experiment, May 2004

Participants: Five participants with disabilities, who had previously completed CTAD's long basic skills assessment.

Method: As used in April 2004, except that this time CTAD's *screener* tests were used.

Results: With the new shorter reports of around 140 words, an average of 38 per cent of items were recalled. Participants commented that they found some items in

[1] Lorch and Lorch (1996) used long texts of around 1,750 words, while Mason and Morris (2000) used very short texts of around 140 words.

the lists of motivational activities inappropriate, but they liked the short reports and the simplicity of the overall structure.

Discussion: Free recall results were better. Many participants received poor scores in the screener because they could not cope with the interface, which was not adapted for people with special needs. This highlighted the problem of what to say in reports when students could not answer any questions correctly.

Revisions: CTAD's screener was revised by adding two more easy questions, so that people with poor basic skills would have the chance to get more questions right. For the NLG module, we sought advice from experts about what to say to people with very low scores, and then added appropriate content rules. Motivational activities were modified slightly to make them less specific, following advice from an expert. We also decided to try breaking the document into shorter sections linked by hypertext buttons, to reduce the amount of text on a screen and thus break down the reading task into more manageable chunks.

4.4.3 Pilot experiment, June 2004

Participants: Eight participants with poor literacy enrolled in basic skills courses at a college.

Method: Participants read a report generated for a person at their level (assessments were done previously). This time, the document was presented in four very short parts, a main part and three subsections linked by hypertext buttons. Browsing behaviour was recorded by the experimenter. Afterwards, the participants did free recall and interviews as before.

Results: Hypertext – all the participants browsed the document by clicking the buttons in the same order (top to bottom), and they all looked at all parts of the document. Free recall – on average, 32 per cent of items were recalled. In the interview, as before, some participants commented that some motivational activities were inappropriate.

Discussion: Recall results were similar to the previous pilot. Hypertext worked well, but there was little variation in browsing behaviour and all participants viewed all parts of the document. Lists of activities were intended to be motivational, but failed to achieve this goal because they were not adapted to individuals but rather to generic levels related to candidates' overall scores. For instance, some people said they could already do some of the activities that had been suggested as objectives for the next level.

Revisions: As a result of this pilot and further small pilots, along with our discussions with experts and the investigation of motivation mentioned above (Tintarev, 2004), we removed content selection rules that generated lists of motivational activities and replaced them by rules that personalised content with information that could be easily obtained from a short questionnaire (described below). Thus, the output documents were shortened even further. Because of this shortening coupled with the lack of variation in browsing behaviour (with all users accessing all parts of the document), we removed hypertext links and generated documents as single blocks of text.

Table 2. *Reports generated by two versions of* SkillSum *from the same user's data (early version at the top and later, revised version, at the bottom)*

Vertion	Sample text
October 2004	**English skills** You scored seventeen. You did very well on finding the main point. But you did not do so well on capital letters, full stops, commas, question marks and apostrophes. It could help you to do a course, if you want to improve your reading and writing skills. You could contact your local college to find out about English courses.
October 2005	**English skills** Thank you for doing this. You got 17 questions right. Click here for more information. Your skills may not be OK for your construction course. It looks as if you find punctuation quite hard. You got all except 2 of the reading questions right. But you made 8 mistakes on the questions about writing. Perhaps you would like to take a course to help you with your punctuation. An English course might help you, because you said you do not feel that your reading is very good. Click here for Key Skills at Xshire College.

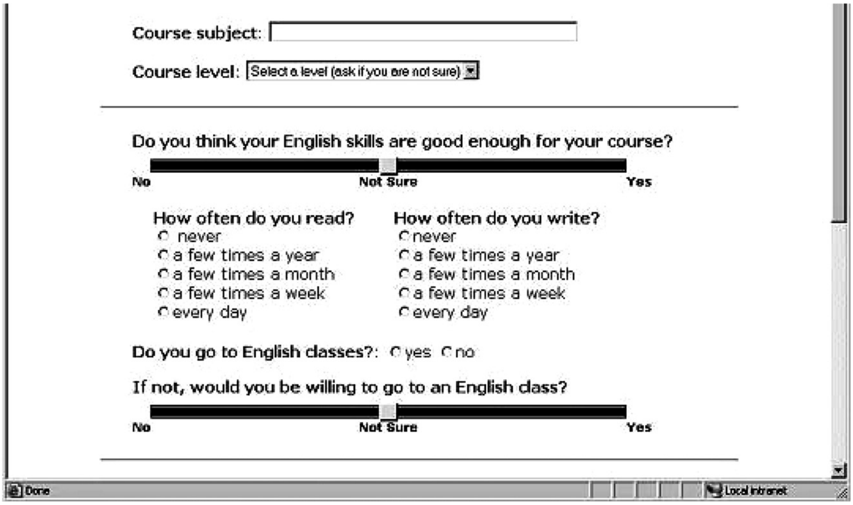
4.4.4 Discussion

Our revisions of the system played a major part in enabling SkillSum to communicate information in a readable way. Section 4.3 above gives more technical detail about how rules were revised. To illustrate how content and structure in SkillSum evolved by revision, Table 2 shows reports generated from the same user's data by the October 2004 version of SkillSum (the upper report) and by the November 2005 version (lower report).

Comparing the two, it is immediately obvious that the content of the 2005 report is more personal. It thanks the user and it mentions the course that the user wants to take, the college at which he/she is enrolled, and the fact that he/she does not rate his/her own skills very highly. The 2005 version contains more information – summaries of results on reading and writing, and hyperlinks to information about scores and local courses. Furthermore, it explicitly states that the user's skills might not be good enough for the course, rather than merely implying it. All of these content revisions came from expert and user interviews and questionnaires.

Some lexical choice rules (Section 5.4) were also changed during revision (we did not change other types of microplanning rules). For example, the non-technical phrase *capital letters, full stops …* (October 2004) was replaced by the more technical term *punctuation* (November 2005) after a pilot study showed that users preferred it.

We also used pilot studies to refine the background information we obtained from users (which was limited to a single-screen questionnaire). In the final version of SkillSum, we asked users to tell us what course they were doing (for users at an FE college), to self-assess their skills, and to tell us how often they read and write (see Figure 7).



Course subject:

Course level: Select a level (ask if you are not sure)

Do you think your English skills are good enough for your course?

No — Not Sure — Yes

How often do you read?	How often do you write?
○ never	○ never
○ a few times a year	○ a few times a year
○ a few times a month	○ a few times a month
○ a few times a week	○ a few times a week
○ every day	○ every day

Do you go to English classes?: ○ yes ○ no

If not, would you be willing to go to an English class?

No — Not Sure — Yes

Fig. 7. Part of pre-test questionnaire to elicit information about users.

Although we did not run experiments that compared different versions of SkillSum, we had a very strong qualitative impression (from free recall and comprehension question studies in pilots) that students found later versions of SkillSum easier to understand. In our early experiments (e.g. April 2004), many students struggled to understand the reports; in experiments carried out half-way through the project (e.g. October 2004), most students understood the reports but a few still struggled; while in experiments at the end of the project (e.g. November 2005), almost all students seemed to understand the reports (although not all agreed with what the report said). We believe that SkillSum achieved some success in helping people evaluate their own skills (Section 6.2) largely owing to the revision process. That is, because of improvements to the document planning rules that choose document content and structure.

5 Choosing linguistic expression: the microplanner

The SkillSum microplanner explicitly represents the readability impact of different microplanning choices, and reasons about which set of choices would lead to the most readable text. Again, we used corpus analysis (with a different corpus) to see which microplanning choices (and sets of choices) were possible, and then created preference rules (largely based on psycholinguistic evidence) which found the optimal set of choices from a readability perspective. We used different mechanisms for discourse-level choices (cue phrases, ordering, aggregation) and lexical choice (for content words), but both mechanisms used the above strategy.

5.1 Discourse-level choices

We focused on three types of discourse-level choices (partially inspired by Moser and Moore, unpublished data (1997) and Moser and Moore (1995)).

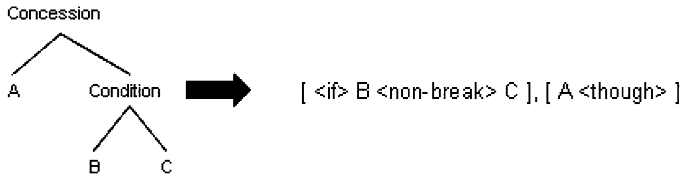


Fig. 8. A rhetorical structure tree mapped to a flat list of sentence structures.

Discourse connectives. Which connective (or multiple connectives), if any, is present and where it is positioned. For example,

- *If you do not have any English qualifications, you may like to take a National test in English at Level 2* (one connective, *If*, placed before the satellite).
- *If you do not have any English qualifications, then you may like to take a National test in English at Level 2* (two connectives, *If* and *then*, *if* placed as before and *then* before the nucleus).

Ordering. Which order the discourse segments occur in. For example,

- *you may like to take a National test in English at Level 2 if you do not have any English qualifications* (nucleus first);
- *if you do not have any English qualifications, you may like to take a National test in English at Level 2* (nucleus second).

Punctuation and aggregation (sentence structure). What punctuation (if any) is used between discourse segments, and whether discourse segments are in separate sentences. For example,

- *What you do next depends on what is important to you; you may like to take a National test in English at Level 2 if you do not have any English qualifications.* (Single sentence, semi-colon separation.)
- *What you do next depends on what is important to you. You may like to take a National test in English at Level 2 if you do not have any English qualifications.* (Two sentences.)

The job of the microplanner is to map RST trees (produced by the document planner) to flat, ordered lists of sentence structures, by making the above choices. Figure 8 shows one possible mapping. A, B and C represent discourse segments, and the output is a list of sentence structures. The first sentence aggregates B and C. The connective *if*, is placed before B, which is followed by non-breaking punctuation (e.g. a comma), then C. The second sentence contains A followed by the connective *though.*

5.2 Modelling hard constraints on discourse-level choices

Not all combinations of the choices in Section 5.1 are legal. For example, we cannot say, ‘Then you may like to take a National test in English at Level 2, if you do not have any English qualifications’ (both *if* and *then* connectives, nucleus first).

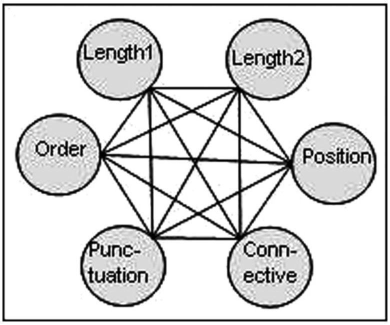


Fig. 9. CSP graph representing a discourse relation.

We built a model of pairs of legal choices by analysing a portion of the RST discourse treebank corpus (RST-DTC) (Carlson, Marcu and Okurowski 2003). Essentially, a pair of choices was deemed legal if it was observed in an RST-DTC text, and illegal otherwise. This analysis was done for the seven most common discourse relations in our report: concession, condition, elaboration-additional, evaluation, example, reason and restatement. Full details of this analysis are given in Williams (2004).

We represented the model as hard constraints in constraint satisfaction problem (CSP) graphs (implemented using JCL 2.1 (Torrens, 2002)); our approach is similar to the constraint-based microplanning of Power (2000) and Power, Scott and Bouayad-Agha (2003). Seven CSP graphs were built – i.e. one for the corpus analysis results of each discourse relation. The structure of the graphs is exactly the same for each relation, with six nodes, and fifteen connections linking each node to all the others. This structure is illustrated in Figure 9.

The nodes in the graph in Figure 9 are CSP domain variables. Estimated lengths in words of the discourse segments to be generated (*Length 1* and *Length 2*) are inputs, the other four nodes are outputs. Constraints between each pair of variables were represented as 'good lists', i.e. pairs of values for the variables that were seen in the RST-DTC, and hence are legal. For instance, in the graph for *elaboration*, the connection between *Length 1* and *Punctuation* contains the pair <*short, sentenceBreak*> in its good list, meaning that if the length of the first segment in this relation is short, it is 'legal' to place sentence-breaking punctuation, such as a full stop, between the segments. And so on for other pairs. We used pairs because our corpus analysis was too small to provide reasonable data for triples, quadruples, etc. Further details can be found in Williams (2004).

A nice aspect of the CSP approach is that it makes different kinds of choices simultaneously. Di Eugenio, Moore and Paolucci (1997), in contrast, used machine learning to determine the sequential order of making choices. This is very restrictive; for example, if the decision to include a particular connective is made first, it could mean that later on it is illegal to choose a sentence break. Using CSP allows us to generate all possible solutions and afterwards use preference rules to choose between them, rather than to make choices too early and run the risk of losing potentially good solutions.

Table 3. *Rules for scoring CSP solutions*

Feature	Readability scoring rules	Control scoring rules
Ordering	If order is nucleus-satellite, add one.	Add points equivalent to the ordering's frequency in the relation.
Connective position	If position is before-satellite, add one.	Add points equivalent to the connective position's frequency in the relation.
Punctuation	If punctuation is sentence-breaking, add 20. If it is non-breaking, add 2.	Add points equivalent to the punctuation's frequency in the relation.
Connective	If a connective is present, add 10. If connective requires an NP argument, take away 5. Add percentage frequency of connective in the relation, (with length correction).	Add points equivalent to the connective's frequency in the relation, (with correction for the percentage of relations in which the connective occurred).

5.3 Modelling preferences on discourse-level choices

We developed two scoring functions to choose between legal text specifications produced by the CSP; these are shown in Table 3. The readability scoring rules represent our belief about which choices are best for texts intended for low-skilled readers. The control rules are primarily based on frequency in the RST-DTC, with a correction for ambiguity applied to the connective choice rules.

Control rules add points to the score for each feature in the solution, according to the percentage frequency of that feature found in the RST-DTC for the relation being processed. For instance, in our corpus analysis, we found 99 per cent of *restatement* relations with *nucleus-satellite* ordering. Therefore, if a CSP solution for *restatement* has *nucleus-satellite* ordering, 99 points are added to its score.

The readability rules favour solutions with short, common discourse connectives, and punctuation between discourse segments that shortens sentences. They also apply a penalty if the connective requires an NP argument, e.g. *without*, which would result in a gerund. These rules are qualitatively based on psycholinguistic findings and advice from domain experts; the exact numerical weights were determined by trial and error.

Our control rules for scoring connectives included a correction (derived by trial and error) to reduce ambiguity. For example, *but* is highly ambiguous, since it occurred in six of the seven relations investigated, or 86 per cent. To score *but* for *concession*, the percentage of *but* in *concession* (33 per cent) is divided by the percentage of relations found with the connective (86 per cent) and multiplied by 100 to give a score of 38. Applying the correction to the less ambiguous *though* results in a much higher score of 73. With frequency scores alone, *though* would score only 9 and *but* would score 33. The ambiguity correction was an attempt to investigate the trade-off in readability between short, common, but semantically ambiguous connectives (favoured by the *readability* scoring rules), and semantically precise, but less common connectives (favoured by the *control* rules).

5.4 Choosing lexical items

The SkillSum microplanner also performed lexical choice – that is, it decided which words should express a concept. This was done in a conceptually similar fashion to discourse-level choices: we first enumerated legal possibilities based on corpus analysis (of the small expert-authored corpus), and then used two preference models (control and readability) to choose between these possibilities. However, implementation of lexical choice was much simpler because we did not look at interactions between pairs of lexical choices. This meant that we could (offline) pre-compute concept-to-word mappings for the different preference models; we did not need to dynamically solve a CSP whilst generating the text.

The first step was to enumerate lexical possibilities – that is, to list the words that could be used in a context to communicate a concept to the reader. We did this by analysing our corpus of expert-authored reports to determine which concepts were being conveyed, and which alternative words and phrases were used in the corpus to communicate each concept. For example, all corpus reports told the student how many assessment questions he or she had answered correctly, but this was expressed in different ways. We observed that three different sentences were used to communicate this meaning:

- *You answered N questions correctly* (used by tutor A);
- *You scored N* (used by tutor B);
- *You got N correct answers* (used by tutor B).

Note that the phrasings are idiosyncratic in the sense that each tutor has one or two preferred phrasings which she sticks to. In discussions with the tutors, we suggested *right* as an alternative for *correctly*, and the tutors agreed this was a reasonable candidate. Hence we came up with the following sets of lexical alternatives:

- *answered*, *got*, *scored* (verb that communicates numerical performance on assessment);
- *questions*, *answers* (noun that refers to responses to assessment questions);
- *correct*, *correctly*, *right* (modifier that indicates response is the correct one).

Note that while *questions* and *answers* of course have quite different meanings in general, in the context of this message, either word can be used to communicate the concept; hence we consider them to be lexical alternatives (in this context).

We were concerned that corpus authors might use words that readers might not know, or might interpret in unexpected ways (Reiter and Sripada, 2002). We tried to identify such words via pilot experiments with readers, and eliminate them from our sets of possible lexicalisations. For example, pilots showed that some people interpreted *grammar mistake* to include mistakes in capitalisation and some did not; hence we tried to avoid using this term in generated texts. Since there is no synonym for *grammar mistake* which is more understandable to readers, we used the more generic term *writing mistake* (which of course refers to many other types of errors as well) in our generated texts.

Jucks and Bromme (2007), who analysed doctor–patient communication, pointed out that patients often interpret technical terms differently from doctors; for example, a patient may use *migraine* to refer to any painful headache, whereas a doctor may use *migraine* to refer to a particular type of headache which recurs and is caused by a specific set of biological mechanisms. Hence, doctors should be cautious about using medical terminology even if patients seem to accept (and even use) it. This seems similar to our observations. Our subjects all realised that *grammar mistake* referred to some kind of mistake in a sentence (similiar to the patients realising that *migraine* referred to some kind of headache), but some of them did not know which specific kinds of mistakes the term referred to. Hence SkillSum, like doctors, should be cautious about using technical terms.

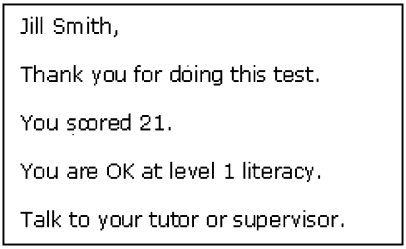
The second step in lexicalisation was to choose which lexicalisation to actually use for a concept (if there was more than one possibility). We developed two preference functions to make this choice. The control preference function was simply based on frequency in the British National Corpus (BNC); given several possible lexicalisations of a concept, it preferred the one that was most common in the BNC. The readability preference function was also based on frequency, but it used frequency in the spoken portion of the BNC (instead of the full BNC), as we thought that this would better represent the language that our low-literacy subjects heard and used. It also favoured shorter words, since our experts thought that shorter words would be easier to read. The actual formula was

$$ScoreForWord(W) = \frac{FreqInSpokenBNC(W)}{LengthOfWordInChars(W)} \tag{1}$$

For example, when deciding whether to lexicalise the concept of INCORRECT-RESPONSE as *error* or *mistake*, *mistake* was preferred by the readability formula, while *error* was preferred by the control. This is because *error* is much more common in the written BNC than in the spoken BNC, in part because *error* is used in technical statistical phrases such as *sampling error*. For this particular example, incidentally, pilot experiments showed that both users and experts agreed that SkillSum reports should use *mistake* instead of *error* (in other words, they agreed with the readability formula). In general, however, there were few such differences between the two preference functions; in most cases they choose the same alternative (see example texts in Figure 11).

One finding from several of our pilot experiments was that people did not always agree with either of our preference functions. In particular, in one experiment, 92 per cent (23 out of 25) subjects preferred ‘You got N questions *correct*’ over ‘You got N questions *right*’ (significant at $p < .001$). Since *right* is much more common than *correct* in both the full BNC and the spoken BNC, and is also shorter, this preference contradicts both of our preference functions.

The problem may be that *right* should not have been considered as a lexical alternative for this concept in the first place. *Right* did not occur in our corpus; as mentioned above, we added it to the set of lexical alternatives for this concept (because it seemed a plausible way of communicating it using a high-frequency word). Although the tutors agreed to this, in retrospect this was a mistake, and perhaps we should have been more cautious about adding new lexical alternatives that were not in our corpus.



Jill Smith,

Thank you for doing this test.

You scored 21.

You are OK at level 1 literacy.

Talk to your tutor or supervisor.

Fig. 10. Example baseline report (no NLG).

6 Evaluation

Our evaluation of SkillSum had two hypotheses.

- **Readability.** SkillSum reports are more readable for low-skilled adults when the readability model was used in microplanning rather than the control (corpus frequency) model.
- **Usefulness.** SkillSum reports help low-skills readers more than the simple canned text reports produced by existing assessment software (see example in Figure 10).

We explored these in two larger experiments. Like the pilots, these were conducted using the SkillSum system as a whole with the most up-to-date screener and NLG components available at the time.

- **First evaluation experiment** (October 2004). Sixty subjects focused on readability. See Section 6.1.
- **Final evaluation experiment** (September/October 2005). 230 subjects tested both readability; see Section 6.1, and usefulness; see Section 6.2.

Like the pilots, both experiments were conducted with students at different Further Education (FE) colleges (using different participants in each experiment). We tried to conduct experiments with other types of subjects, but this proved difficult. Although there are, of course, large numbers of people with poor literacy in the UK, they tend to have low self-confidence and also dislike being reminded of their literacy problems; hence it is not easy to recruit them as experimental subjects. In contrast, FE college students in general were willing to be subjects if their tutors encouraged them to take part in our experiments. We were fortunate in finding a number of FE college tutors who were excited by our project and willing to encourage their students to be subjects.

Both experiments were carried out at the FE colleges in classrooms (in other words, we could not conduct experiments in a controlled laboratory environment in our university). Participants were new students who had just started a course.

A major problem in our experiments was variability among subjects. Low-skilled adult readers are an extremely diverse group. This depends partly on the reason for poor skills; for example, dyslexics, non-native speakers and people who attended poor schools have different profiles. But even within each of these groups, there were major differences among individuals. This made it difficult for us to get statistically

Table 4. *First experiment: results for oral reading rates (2004 version of* SkillSum*)*

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

significant results, since there was a lot of 'noise' due to inter-subject variability. It also made us wonder whether it would be more sensible to build readability models focused on particular groups or even individuals, instead of trying to create a general readability model which works for all low-skilled readers (SkillSum's goal); we discuss this further in Section 7.2.

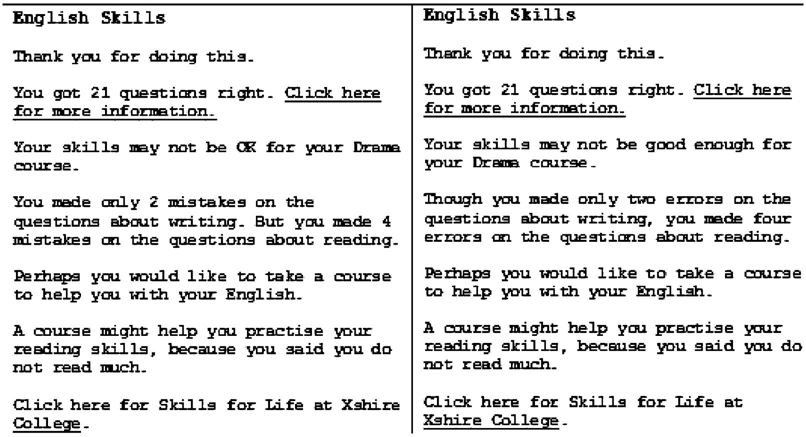
6.1 Readability evaluation

In our first experiment in 2004, we asked subjects to orally read texts generated using the SkillSum readability and control preference models in the microplanner (these models, described in Section 5.2, remained the same across different versions of the NLG system, but content and discourse structuring rules in the document planner changed radically, as described in Section 4.4). We measured oral reading rate and oral reading errors; we also asked comprehension questions and measured response correctness. These measures are commonly used by psychologists (Kintsch and Vipond, 1979) and educationalists (see the Adult Reading Components Study, www.nifl.gov) to measure reading difficulty. We found a statistically significant effect in oral reading speed (Table 4); a text produced using the readability model was read on average 16 words per minute (9 per cent) faster than a text produced using the control model. There were no significant differences in the other measures. See Williams (2004) for full details and analysis of this experiment.

In our final experiment in late 2005, we again compared texts generated with the readability model to texts generated with the control model (with the same content); see Figure 11 for an example. We measured

- oral reading speed and speech errors (as in the first experiment);
- correct responses to comprehension questions (as in the first experiment);
- preferences (show subjects both versions, ask which of the two they prefer);
- silent reading speed (ask subjects to read texts and respond to a comprehension question, measure time taken to do this – pilots showed that if we simply asked subjects to silently read a text and press a button when finished, many would press the button right away without actually reading the text).

We did not obtain any significant effects in any of these measures. We believe that the reason we did not reproduce the results of the first experiment was that our revision-based improvements to the content selection and discourse structuring algorithms (Section 4.5) substantially reduced the effect of later discourse-level planning by the microplanner. In fact, the differences between texts produced by the control and readability models (in the final version of SkillSum) was quite small, as can be seen by comparing the texts in Figure 11.



English Skills

Thank you for doing this.

You got 21 questions right. Click here for more information.

Your skills may not be OK for your Drama course.

You made only 2 mistakes on the questions about writing. But you made 4 mistakes on the questions about reading.

Perhaps you would like to take a course to help you with your English.

A course might help you practise your reading skills, because you said you do not read much.

Click here for Skills for Life at Xshire College.

English Skills

Thank you for doing this.

You got 21 questions right. Click here for more information.

Your skills may not be good enough for your Drama course.

Though you made only two errors on the questions about writing, you made four errors on the questions about reading.

Perhaps you would like to take a course to help you with your English.

A course might help you practise your reading skills, because you said you do not read much.

Click here for Skills for Life at Xshire College.

Fig. 11. Reports generated with the Readability model on the left-hand side and the Control model on the right-hand side (November 2005 version of SkillSum).

Table 5. *Readability statistics over reports generated from 191 students' data*

SkillSum version	Preferences model	Mean words per sentence	Mean characters per word	Flesch Reading ease	Flesch–Kincaid grade
Oct 2004	Control	15.6	4.4	75.5	6.6
	readability	10.0	4.3	82.3	4.2
Nov 2005	Control	10.6	4.5	81.9	4.4
	readability	9.4	4.4	83.8	3.9

An interesting perspective on the differences between the October 2004 and November 2005 SkillSum systems comes from computing the standard Flesch and Flesch–Kincaid readability statistics on the outputs of these systems (see de Vries' (1999) overview of readability formulae). Table 5 shows averaged readability statistics for 191 texts (generated from real student data), generated using both Readability and Control models in both versions of the system. This shows a marked difference between Control and Readability models for the October 2004 system, but not for the November 2005 system. This is essentially because the November 2005 document planner produced simpler and shallower rhetorical trees than the October 2004 system; the simpler structures were a consequence of document structure revisions (see Section 4.4).

Writing guides such as www.plainenglish.co.uk suggest that sentences should on average be 15–20 words long; and the control model of the October 2004 version of SkillSum produced sentences of this length. Our readability model produced much shorter sentences (ten words on average), and this reflected our belief that poor readers find it easier to read sentences that are shorter than writing guides recommend. By November 2005, both models produced almost equally short sentences even though the models had not changed. What had changed was the inputs to these models: we had changed the document-structuring templates in the document planner during the revision process using progressively simpler structures

Table 6. *Chi-square test: self-assessment slider movement, baseline literacy report versus* SkillSum *report*

Report received	*n*	Wrong direction	Right direction	No change	Pearson chi--square	Asymp. Sig. (two-sided)
Baseline	63	19	30	14	8.0	0.018
SkillSum	60	6	34	20		

which experts and users both favoured. This had the effect of decreasing the depth of rhetorical paragraph 'trees' to such an extent that only a single leaf node remained in many paragraphs. Discourse planning would of course be ineffective on such paragraph structures and only differences resulting from lexical choices would be seen in the output. In retrospect, we should perhaps have included comparisons of the readability of texts produced by different versions of SkillSum in our evaluations. However, the low figures achieved by the control model in the November 2005 NLG system in Table 5 demonstrate that once such radical revisions had been made to the document planner, it constructed such simple discourse structures that a simpler algorithm for making discourse-level choices (i.e. based on corpus frequencies alone, like the control model) would have sufficed in the microplanner.

6.2 Helpfulness of reports for users

In our final experiment in late 2005, we investigated whether NLG technology was effective compared to CTAD's existing canned text feedback method. We thus compared versions of SkillSum with and without NLG technology by attempting to find out whether subjects who received generated reports increased their understanding of how good their skills were compared to people who received baseline reports (CTAD's canned text, e.g. see Figure 10). This was measured by asking subjects to self-assess their literacy skills before they used SkillSum (see the slider in Figure 7 with the question 'Do you think your English Skills are good enough for your course'), repeating this question after they had taken the SkillSum assessment and had read either the SkillSum or baseline report, and seeing whether subjects had changed the slider in the right direction ('right direction' was determined by their performance on the assessment; the college told us what performance they expected for each course). Significantly fewer people who read SkillSum reports moved the slider in the wrong direction compared to those who read baseline reports, and more people who read SkillSum reports moved the slider in the right direction; see Table 6. Perhaps this was because SkillSum reports explicitly state whether a user's skills are good enough for his/her course, whereas the baseline reports merely state the user's overall level.

After subjects had finished the self-assessment exercise, we showed them the other version of their report and asked which of the two they preferred. Only 55 per cent preferred SkillSum reports (not significant), which was surprising because in a pilot in June 2005, 87 per cent had preferred SkillSum reports (significant at $p < 0.01$). Both sets of participants saw their own reports; the only difference between the participants was that those in the June pilot had reached the end of their college courses (so information on whether their literacy or maths skills were good enough to complete their courses was not really relevant) whereas students who participated in the final experiment were starting new courses (so the information was relevant). We explored this difference in results further in a smaller follow-up experiment, where we asked another set of students which report was most useful as well as to state their preference (this time students saw printouts of anonymised reports for other people, rather than their own, since we did not have computer access at their particular college). We found that 92 per cent (23 out of 25) students believed that SkillSum reports were more useful ($p < .0001$); however, only 72 per cent (18 out of 25) actually preferred the SkillSum reports ($p = 0.023$). This type of finding is not unique, athough it interesting that some people prefer reports that they regard as less useful.

Qualitative comments from the students were also interesting. Those who preferred baseline reports said that they thought SkillSum reports were 'not nice' and might upset people or make them feel bad; indeed, two subjects in the final experiment had been distressed by their reports (this was unexpected because we had not received any comments in the pilots, which suggested that the reports had upset participants). On the other hand, some of the students who preferred SkillSum reports said that they thought SkillSum reports were nicer and less upsetting than the baseline reports because they gave more information and context.

Obviously telling someone that they cannot read very well can have a significant emotional impact. Our experiences (in other projects as well as SkillSum, e.g. Reiter, Robertson and Osman (2003)) suggest that the best way to present such 'bad news' to someone depends on their personality. Until we have good computational models of personality, perhaps it is best for bad news to be delivered by human tutors instead of by a machine, especially as many people probably prefer to have bad news delivered by a person in any case. This suggests that SkillSum should not be used without a human tutor present.

7 Recommendations, future work and conclusions

7.1 Recommendations

We suggest that anyone building an NLG system for subjects with poor literacy should keep the following points in mind.

- *Texts should be short.* SkillSum reports for users with very low literacy scores were no more than twenty-five words in length, increasing to a maximum of around ninety words for users with higher scores. This may inevitably mean losing some information, but we found that people with poor literacy will not read long texts.
- *Texts should have a very simple structure.* Use very short sentences, paragraphs and shallow rhetorical structure trees. During our revision process, many paragraphs which included two or three discourse relations were replaced by multiple paragraphs, each of which included a single relation or discourse

segment. SkillSum sentences average 10.6 words in length, hence they are much shorter than recommended by guides, such as www.plainenglish.co.uk.

- *Do not use technical terms that users may not understand.* From a lexical perspective, the key challenge is to avoid words that users do not understand. This can be determined by conducting comprehension experiments on representative users, which can be a time-consuming process, but unfortunately we do not know of any reliable shortcuts. Certainly, the simplistic approach of using raw BNC frequency to predict which words are correctly understood does not work. For example, SkillSum users understood *punctuation* more reliably than *grammar*, even though *grammar* is 10 times more common in the BNC than *punctuation*. Indeed, the fact that *grammar* is relatively common, and hence used in many non-technical contexts (such as *grammar school*) may mean that users are less likely to interpret it correctly when it is used in a technical sense.
- *Be very careful when communicating emotionally depressing information.* Indeed, perhaps it is best to leave this task to a human.
- *Test your system with experts and users.* The most important lesson of them all: pilot your system with experts and users, and keep on doing so until they seem satisfied. Of course, this is a good advice when developing any IT system, as advocated by the HCI community, but it is perhaps especially important when developing a system for users who have low self-confidence.

Whilst it is true that some of these recommendations echo the kind of advice found in general guidelines on readability, we would like to emphasise that they have in fact been evaluated empirically with respect to the concrete task reported in this paper and with representative users.

7.2 *Future work*

SkillSum tried to create a single readability choice model which would work for all poor readers, and as described above this was not entirely successful, in part because in practice it did not differ much from a control model which simply picked the most common choices. But perhaps this is an inevitable consequence of trying to create a model which covers such a diverse and heterogeneous group. We would like to try creating readability choice models which are focused on smaller groups or even on individuals (the ideal case). In other words, we would like to try to build models of the skills, deficits, vocabulary and preferences of groups with specific reading impairments (such as dyslexia or non-native speakers), or (even better) of individual readers, and tailor texts to such models. We believe this would have a significant impact on readability.

The interactions between different kinds of revisions and different kinds of choices in the readability model should be investigated in more depth, as well as the best computational architecture for finding an optimal text when many kinds of choices are being considered. If a gold standard corpus of expert-authored texts should be developed in the future, we could use it to derive content and document structure; furthermore, the generated texts could be compared to it in evaluations.

Finally, from a more pragmatic perspective, techniques for automating the analysis of our initial corpus could be developed, and also techniques for automatically revising the system. We did this manually, and it was very time-consuming; automatic techniques would make the process of building a SkillSum-like system much cheaper.

7.3 Conclusions

Generating appropriate texts for people with poor literacy is an important challenge for NLG. In SkillSum, we explored two approaches to this problem: an empirical approach that incorporated extensive piloting and revision, and a more theory-driven approach that formulated explicit psycholinguistically inspired models for choosing linguistic expressions that enhanced readability. Our experiments suggest that this combination of approaches worked fairly well in our particular application, although it is difficult to create a good linguistic choice model for a group as heterogenous as adults with poor literacy skills.

In the longer term, we believe that NLG systems will be able to generate more readable texts for their users by taking into consideration the specific reading (dis)abilities and preferences of their users, perhaps basing this on models of the effects of specific reading impairments. In the shorter term, we recommend that anyone building an NLG system for low-skilled readers should extensively pilot and revise the system with its intended users; this is perhaps not very exciting in academic terms, but is absolutely essential to creating a good system for people with poor literacy skills.

References

Bateman, John A. and Paris, Cécile L. (1989) Phrasing a text in terms the user can understand. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, IJCAI, 1989, Detroit, MI, pp. 1511–17.

Binkley, M., Matheson, N. and Williams, T. (1997) Working Paper: Adult Literacy: An International Perspective. Technical Report, National Center for Education Statistics (NCES) Electronic Catalog No. NCES 9733, http://nces.ed.gov.

Brown, J. and Eskenazi, M. (2005) Student, text and curriculum modeling for reader-specific documant retrieval. In *Proceedings of the IASTED International Conference on Human–Computer Interaction*, Phoenix, AZ.

Canning, Y. (2002) *Improved Syntactic Analysis of, and Simplified Text Generation from, Free-Form Text*. PhD Thesis, University of Sunderland, Sunderland.

Carlson, L., Marcu, D. and Okurowski, M. E. (2003) Building a discourse-tagged corpus in the framework of rhetorical structure theory. In Jan van Kuppevelt, and Ronnie Smith (eds.), *Current Directions in Discourse and Dialogue*, Text, Speech and Language Technology, Vol. 22, pp. 85–112. Berlin, Springer.

Chandrasekar, R. and Srinivas, B. (1997) Automatic induction of rules for text simplification. *Knowledge-Based Systems* **10:** 183–90.

Coleman, E. (1962) Improving comprehensibility by shortening sentences. *Journal of Applied Psychology* **46**: 131–4.

Collins-Thompson, K. and Callan, J. (2004) A language modeling approach to predicting reading difficulty. In Susan Dumais, Daniel Marcu and Salim Roukos (eds.), *HLT-NAACL*

2004: Main Proceedings, pp. 193–200. Morristown, NJ: Association for Computational Linguistics.

Degand, L., Lefèvre, N. and Bestgen, Y. (1999) The impact of connectives and anaphoric expressions on expository discourse comprehension. *Document Design* **1**: 39–51.

Devlin, S., Canning, Y., Tait, J., Carroll, J., Minnen, G. and Pearce, D. (2000) An AAC aid for aphasic people with reading difficulties. In *Proceedings of the 9th Biennial Conference of the International Society for Augmentative and Alternative Communication (ISAAC 2000)*, Washington, USA, pp. 10–12.

Devlin, S. and Tait, J. (1998) The use of a psycholinguistic database in the simplification of text for aphasic readers. In John Nerbonne (ed.), *Linguistic Databases*, pp. 161–73. Cambridge: Cambridge University Press, CSLI Publications.

DeVries, H. (1999) *Reading Ease@WWW*. Masters Thesis, Macquarie University, Australia.

Di Eugenio, B., Glass, M., Trolio, M. J. and Haller, S. (2001) Simple natural language generation and intelligent tutoring systems. *Proceedings of Artificial Intelligence in Education*, pp. 50–8.

Di Eugenio, B., Moore, J. D. and Paolucci, M. (1997) Learning features that predict cue usage. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid, Spain, pp. 80–7.

Eddy, B. (2002) Towards balancing conciseness, readability and salience: an integrated architecture. *Proceedings of the International Natural Language Generation Conference*, New York, pp. 173–8.

Geldof, S. (2003) Corpus analysis for NLG. In E. Reiter, H. Horacek and K. van Deemter (eds.), *Proceedings of the 9th European Workshop on Natural Language Generation (ENLG'03)*, Budapest, Hungary, pp. 31–8.

Harley, T. (2001) *The Psychology of Language from Data to Theory*. Erlbaum, UK: Psychology Press.

Hunter, D. and Howard, U. (2004) Including language, literacy and numeracy learning in all post-16 education. Guidance on curriculum and methodology for generic initial teacher education programmes. Technical Report, FENTO (Further Education National Training Organisation), www.nrdc.org.uk.

Inui, K., Fujita, A., Takahashi, T., Iida, R. and Iwakura, T. (2003) Text simplification for reading assistance: a project note. *2nd International Conference on Paraphrasing: paraphrase acquisition and applications*, Sapporo, Japan, pp. 9–16.

Jucks, R. and Broome, R. (2007) Choice of words in doctor–patient communications: an analysis of health-related internet sites. *Health Communication* **21**(3): 267–77.

Kintsch, W. and Vipond, D. (1979) Reading comprehension and readability in educational practice and psychological theory. In L. G. Nilsson (ed.), *Perspectives on Memory Research*, pp. 329–65. Hillsdale, NJ: Lawrence Erlbaum.

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

Knott, A. and Sanders, T. (1998) The classification of coherence relations and their linguistic markers: an exploration of two languages. *Journal of Pragmatics* **30**(2): 135–75.

Lavoie, B. and Rambow, O. (1997) RealPro: A fast, portable sentence realizer. IN *Proceedings of the Conference on Applied Natural Language Processing (ANLP, 1997)*, Washington, USA, pp. 265–8.

Leijten, M. and van Waes, L. (2001) The impact of text structure and linguistic markers on the text comprehension of elderly people. In L. Degand, Y. Bestgen, W. Spooren and L. van Waes (eds.), *Multidisciplinary Approaches to Discourse*, pp. 21–9. Amsterdam: Stichting Neerlandistiek VU, Münster: Nodus Publikationen.

Lorch, R. F. and Lorch, E. P. (1996) Effects of organizational signals on free recall of expository text. *Journal of Educational Psychology* **88**(1): 38–48.

Mann, W. C. and Thompson, S. A. (1987) Rhetorical structure theory: a theory of text organization. Technical Report, ISI/RS-87-190, Document Center, USC/ISI, Marina del Rey, CA.

Mason, J. and Morris, L. (2000) Improving understanding and recall of the probation service contract. *Journal of Community and Applied Social Psychology* **10**(3): 199–210.

McKeown, K., Robin, J. and Tanenblatt, M. (1993) Tailoring lexical choice to the user's vocabulary in multimedia explanation generation. In *Proceedings of ACL*, Columbus, OH, pp. 226–34.

Milosavljevic, M. and Oberlander, J. (1998) Dynamic hypertext catalogues: helping users to help themselves. In *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia (HT, 1998)*, Pittsburgh, PA, pp. 123–31.

Miltsakaki, E., Dinesh, N., Prasad, R., Joshi, A. and Webber, B. (2005) Experiments on sense annotations and sense disambiguation of discourse connectives. In *Proceedings of the 4th Workshop on Treebanks and Linguistic Theories*, Barcelona, Spain.

Moore, J. D., Porayska-Pomsta, K., Varges, S. and Zinn, C. (2004) Generating tutorial feedback with affect. In *Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference*, Miami Beach, FL, pp. 923–8. AAAI Press, Menlo Park, CA.

Moser, C. (1999) Improving literacy and numeracy: a fresh start. The report of the working group chaired by Sir Claus Moser. Technical Report, www.lifelonglearning.co.uk/mosergroup.

Moser, M. and Moore, J. D. (1995) Investigating cue selection and placement in tutorial discourse. In *Proceedings of the 33rd Annual Meeting on Association For Computational Linguistics* (Cambridge, Massachusetts, June 26–30, 1995). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, USA, 130–135.

Okumura, M. (2000) Producing more readable extracts by revising them. *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*, Saarbrücken, Germany, pp. 1071–5.

Paris, Cécile L. (1988) Tailoring object descriptions to the user's level of expertise. *Computational Linguistics* **14**(3): 64–78.

Power, R. (2000) Planning texts by constraint satisfaction. In *Proceedings of the International Conference on Computational Linguistics (COLING, 2000)*, Saarbrücken, Germany, pp. 642–8.

Power, R., Scott, D. and Bouayad-Agha, N. (2003) Document structure. *Computational Linguistics* **29**(2): 211–60.

Reiter, E. and Dale, R. (2000) *Building Natural Language Generation Systems*. Cambridge, UK: Cambridge University Press.

Reiter, E. and Sripada, S. G. (2002) Human variation and lexical choice. *Computational Linguistics* **28:** 545–53.

Reiter, E., Robertson, R and Osman, L. M. (2003) Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, **144**(1–2): 41–58.

Reiter, E., Sripada, S. G. and Robertson, R. (2003) Acquiring correct knowledge for natural anguage generation. *Journal of Artificial Intelligence Research* **18**: 491–516.

Reiter, E., Williams, S. and Crichton, L. (2005) Generating feedback reports for adults taking basic skills tests. In *Proceedings of the Twenty-fifth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*, Cambridge, England, pp. 50–63.

Sanders, T. J. M. and Noordman, L. G. M. (2000) The role of coherence relations and their linguistic markers in text processing. *Discourse Processes* **29**(1): 37–60.

Scott, D. and de Souza, C. (1990) Getting the message across in RST-based text generation. In R. Dale, C. Mellish and M. Zock (eds.), *Current Research in Natural Language Generation*, pp. 47–73. Cognitive Science Series. London: Academic Press.

Siddharthan, A. (2002) Resolving attachment and clause boundary ambiguities for simplifying relative clause constructs. In *Proceedings of the Student Research Workshop, 40th Meeting of the Association for Computational Linguistics*, Philadelphia, PA, pp. 60–5.

Siddharthan, A. (2003) Preserving discourse structure when simplifying text. *Proceedings of the 9th European Workshop on Natural Language Generation*, Budapest, Hungary, pp. 127–34.

Steeds, A. (2001) Adult literacy core curriculum including spoken communication. Technical Report, Cambridge Training and Development Ltd. on behalf of The Basic Skills Agency, ISBN 1-85990-127-1.

Tintarev, N. (2004) *Content Determination for Reports Aimed at Adult Literacy Learners.* Masters Thesis, Uppsala Universitet, Sweden.

Torrens, M. (2002) Java constraint library 2.1. Technical Report, Artificial Intelligence Laboratory, Swiss Federal Institute of Technology.

Walker, M., Whittaker, S., Stent, A., Maloor, P., Moore, J., Johnston, M. and Vasireddy, G. (2003) Generation and evaluation of user tailored responses in multimodal dialogue. *Cognitive Science, Rumelhart Prize Special Issue Honoring Aravind K. Joshi*, **28**(5): 811–40.

Williams, S. (2004) *Natural Language Generation of Discourse Relations for Different Reading Levels*. PhD Thesis, University of Aberdeen, Aberdeen.

Williams, S. and Reiter, E. (2005) Deriving content selection rules from a corpus of non-naturally occurring documents for a novel NLG application. In *Proceedings of the Workshop on Using Corpora for Natural Language Generation*, pp. 41–8. Technical Report, no. ITRI–05–03, University of Brighton: Information Technology Research Institute (ITRI).

Zukerman, I. and Pearl, J. (1986) Comprehension-driven generation of meta-technical utterances in math tutoring. In *5th National Conference AAAI-86*, Philadelphia, PA, pp. 606–11.