CORE

Metadata, citation and similar papers at core.ac.uk

Provided by Open Research Online

The Open University

Open Research Online

The Open University's repository of research publications and other research outputs

Generating Feedback Reports for Adults Taking Basic Skills Tests

Conference or Workshop Item

How to cite:

Reiter, Ehud and Williams, Sandra (2005). Generating Feedback Reports for Adults Taking Basic Skills Tests. In: Proceeding of the The Twenty-fifth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, 12-14 Dec 2005, Cambridge, UK.

For guidance on citations see FAQs.

© [not recorded]

Version: [not recorded]

Link(s) to article on publisher's website:
http://www.bcs-sgai.org/ai2005/

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online's data policy on reuse of materials please consult the policies page.

oro.open.ac.uk

Generating Feedback Reports for Adults Taking Basic Skills Tests

Ehud Reiter and Sandra Williams
Dept of Computing Science, University of Aberdeen
{ ereiter,swilliam}@csd.abdn.ac.uk

Lesley Crichton
Cambridge Training and Development Ltd
lesleyc@ctad.co.uk

Abstract

SkillSum is an Artificial Intelligence (AI) and Natural Language Generation (NLG) system that produces short feedback reports for people who are taking online tests which check their basic literacy and numeracy skills. In this paper, we describe the SkillSum system and application, focusing on three challenges which we believe are important ones for many systems which try to generate feedback reports from Web-based tests: choosing content based on very limited data, generating appropriate texts for people with varied levels of literacy and knowledge, and integrating the web-based system with existing assessment and support procedures.

1. Introduction

There are a growing number of short assessment tests available on the Web, which people can use to assess their health, education, entitlement to benefits, and so forth. Users fill out a form (typically multiple-choice questions), and submit this to a server, which returns to them a numerical score and a fixed text explaining the score. For example, someone using the nicotine addiction test on www.healthcalculators.org will be told whether he or she has a low, medium, or high level of nicotine addiction, together with some explanatory text. Such tests are popular because people can use them at any time, and in complete privacy; we expect that their use will continue to grow, and indeed they will be regarded as essential tools of life in the 21st century.

Currently people using such tests get limited feedback, typically just a level (as in the nicotine addiction example) accompanied by a fixed explanatory text and often a suggestion to contact a professional doctor, lawyer, tutor, etc in order to learn more. The goal of our research is to try to develop a system which produces more detailed and personalised feedback, using Natural Language Generation (NLG) technology, in the belief that better feedback will make such tests more useful and effective.

This paper discusses SkillSum, an NLG system which generates short feedback reports for adults who have just completed a screening test of their basic literacy or numeracy skills. We focus on the following issues, which we believe are relevant to feedback-report-generation applications in general, not just SkillSum:

- Selecting content based on very limited data.
- Generating texts which are easy to read, for people with varied levels of reading ability.
- Integrating report-generation systems into the overall assessment process.

Overall, generating high-quality feedback reports from the results of short tests is more difficult than we first expected; but we believe that it is possible, and that this technology could be both commercially important and beneficial to society.

2. Background

2.1 Natural Language Generation

Natural Language Generation (NLG) systems automatically generate texts in English and other human languages, typically based on some non-linguistic input data, using AI and NLP techniques (Reiter and Dale, 2000). For example, the STOP system (Reiter, Robertson, Osman 2003) generates personalised smoking-cessation leaflets based on a smoker's responses to a questionnaire about her smoking habits, beliefs, and so forth; and the ILEX system (O'Donnell *et al*, 2001) generates descriptions of museum exhibits based on a knowledge base that contains information about items in the museum.

This paper focuses on SkillSum as an application, not on technical NLG issues. For general information on NLG, see Reiter and Dale (2000).

2.2 Basic Skills Assessments

Poor adult literacy and numeracy is a major problem in the UK. The Moser study (Moser et al, 1999) reported that one in five adults in the UK 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 £2 when buying a 68p loaf of bread and two 45p tins of soup. Such people have difficulty finding and keeping jobs, and also have a lower quality of life; poor literacy and numeracy are also a major cause of low productivity in the UK economy as a whole. 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 in England by 2007; similar initiatives are in place in Scotland, Wales and Northern Ireland. 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 existing level of literacy and numeracy assessed to give a clear picture of his or her strengths, weaknesses and learning needs. Proper assessment requires the individual 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.

As many people may initially be reluctant to make this time commitment, there is increasing interest in short *screener* tests, which can be completed quickly and give a general indication of the student’s abilities. These can quickly tell students who are concerned about their skills whether they have any problems, and hence whether they should consider enrolling in a class to improve their skills (a detailed assessment test is usually administered as part of such classes). Screener tests are also useful for organisations such as UK Further Education (FE) colleges (similar to American community colleges), which need to determine which incoming students should be asked to attend skills classes.

Screener tests should be as easy to take as possible, which means that they should be short, and also that ideally people should be able to take them anywhere (not just in a classroom) with minimal support from human tutors. Screener tests are already being put on the web, which makes them available anywhere there is Internet access. But if they are going to be used with minimal support from human tutors, they also need to be able to present their results to users in an easy-to-understand and meaningful fashion. This is the goal of SkillSum: to automatically generate a personalised report summarising how well someone did on a basic skills screener, which encourages this person (if appropriate) to agree to more detailed assessment, to accept basic skills support as part of another course, or to sign up for a discrete literacy or numeracy course.

2.3 Related Work

Some existing web-based educational assessment tools, such as iAchieve at home (http://www.iachieve.com.au) (which is intended for children, not adults) provide limited feedback reports. For example iAchieve reports tell students how many questions they got wrong, explain how this performance compares to other children at the same grade level, and also give (fixed) explanations of how questions should be answered. There are also a number of commercial systems which help teachers write reports on their pupils, such as ReportMaster (http://www.carnsoftware.co.uk/report.htm).

We are not aware of any online assessment tools that use NLG technology to generate feedback reports. The Criterion system (Burstein, Chodorow, Leacock 2003) uses sophisticated NL Understanding techniques to analyse writing samples (which students can submit on the web) and identify problems in how a student writes, but it does not use NLG to communicate its analysis to the student.

3. SkillSum

SkillSum's goal is to develop an NLG system which automatically generates useful and understandable feedback reports for people who are taking a short online screening test of their basic skills. It is a collaborative project between the University of Aberdeen and Cambridge Training and Development. It builds on an earlier PhD project at Aberdeen [Williams 2004] which made an initial attempt at building such a system. Essentially the PhD project focused mostly on theoretical issues involved in generating texts for low-literacy readers, and did not seriously try to build a real application. The goal of SkillSum is to explore application issues as well as theoretical issues, and to build a system which is robust and realistic enough to enable us to evaluate whether we can indeed automatically generate useful and helpful feedback texts for real people who are concerned about their basic skills.

An example output (with the name of the student changed) from the current version of SkillSum is shown in Fig. 1. This report is generated from the student's response to 27 assessment questions (mostly multiple choice). A typical question is shown in Fig. 3, together with some background information about the student (Fig. 2).

3.1 Knowledge Acquisition

SkillSum reports are based on knowledge acquisition (KA) activities with domain experts (basic skills tutors) (Williams and Reiter, 2005a) and on pilot experiments. Essentially we asked tutors to write some example reports; analysed these to determine what information tutors were trying to communicate to students and also how they thought this information should be expressed; and then implemented a simplified version of these rules in the software. We then showed reports produced by our software to both tutors and students taking basic skills courses, and revised the reports based on this feedback. This follows the general KA for NLG methodology described by Reiter, Sripada, and Robertson (2003).

One of the most important findings of our KA activities was that reports should be short. Some initial versions of SkillSum generated much longer and more detailed reports, but our pilot experiments showed that users wanted short and simple reports, perhaps with details available on another page (e.g. the more information link in Fig. 1); this may reflect the fact that reading a long report requires considerable effort from people with limited literacy.

Our experiments and KA sessions also suggested that

- reports should focus on diagnosis (what the student can and cannot do) and advice (what the student should do to improve his/her skills)
- reports should be relevant to the student's interests and objectives
- reports should not used specialised terminology

While these points may seem obvious in retrospect, in fact some early versions of SkillSum included background information about basic skills, did not try to tailor the reports to students skills and interests, and used terminology that was meaningful to tutors but not to students.

3.2 Implementation

SkillSum is implemented as a web-based system using J2EE (Java); the NLG system is a server-side system which gets input from web forms and databases, and produces an HTML web page as its output. The system divides the task of generation texts into three stages (document planning, microplanning, and realisation), following the architecture of Reiter and Dale (2000).

The *Document Planner* decides what information should be communicated in the text, and how the text should be organised rhetorically. Conceptually it is based on rules acquired by our KA activities, such as:

> IF the student has said he/she is not confident about his/her English skills (even if their level is in fact OK for the student's intended course)
>
> THEN add a message that he/she should consider taking an English course to improve his/her confidence

The "*But an English class…*" sentence in Figure 1 is based on this rule.

The output of the document planner is a tree whose leaves are messages, and whose internal nodes communicate discourse (rhetorical) relations that relate messages and groups of messages (Williams, 2004).

Three representations of messages were used in different versions of SkillSum. Initially we represented messages as deep syntactic structures, similar to those used by RealPro (Lavoie and Rambow, 1997). However, as we modified the system based on KA activities and pilots, in many cases we simply encoded messages as strings, as that was quicker and also considerably easier for people who had limited linguistic expertise. The current version of SkillSum uses an intermediate representation, essentially strings annotated with choices that must be made by the microplanner. An example of such an annotated string is

> "your [English] skills $HEDGE are$ $okay$ for your XX $class$"

Square brackets ([]) indicate optional fragments which the microplanner can delete if it wishes, and dollar brackets ($$) indicate words which can be replaced by a synonym if the microplanner wishes. There are also some flags; for example HEDGE means that a word can be hedged if the microplanner wishes. For example, if the microplanner processes the above string and decides to include optional fragments, include hedges where possible, and replace $class$ by its synonym "*course*", then the result is one of the sentences in Figure 1, namely:

> *Your English skills seem to be okay for your Art, Design and Media course*.

The *Microplanner* (second NLG module) makes choices on how to express content and structure. Content-expression choices basically are the choices involved in processing annotated structures such as the above; this is currently done using a set of rules suggested by tutors. Structure-expression choices include deciding on the order of messages, on the placement of sentence and paragraph breaks, and on the choice of cue phrases such as "*But*" and "*however*". Structure-expression choices are made using a constraint-based approach which has been described in detail elsewhere (Williams and Reiter, 2005b).

The *Realiser* (final NLG module) generates actual texts based on the decisions made by the Document Planner and Microplanner. The most complex part of the SkillSum realiser is a (much) simplified and cut-down version of RealPro (Lavoie and Rambow, 1997), which is used to convert deep-syntactic messages (if these are present) into text. Otherwise, the SkillSum realiser just addresses capitalisation, punctuation, and HTML issues.

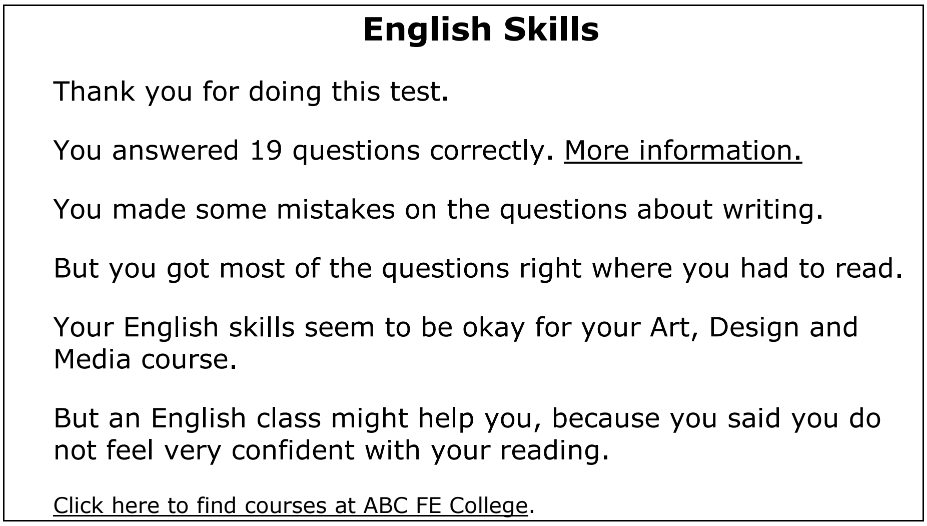
3.3 Evaluation

As mentioned in Section 3.1, we have conducted a number of pilot evaluations of SkillSum. These evaluations involved showing SkillSum reports (or several variations of SkillSum reports) to students who are already enrolled in skills courses, and asking them to do various activities with the reports (such as commenting on them, giving preferences between versions, reading them aloud, and answering comprehension questions). We have conducted 7 of these pilots so far: 6 small ones involving 5-20 people, and one larger one involving 60 people. These evaluations were mostly viewed as knowledge acquisition exercises to improve our system. At our most recent evaluation (in June 2005) we asked 15 students to express a preference between SkillSum reports and the simple reports currently generated by CTAD's software (which just give a score and level, see Figure 4). 13 of the 15 preferred SkillSum reports (significant at $p < .01$ using binomial test), which is encouraging and suggests the system is working reasonably well (we did not see such a clear preference with early versions of SkillSum).

We will conduct a larger final evaluation of SkillSum in September 2005. During this evaluation, we will ask 200 students who are just entering an FE college to take the SkillSum screener test. These students will be divided into three groups:

- *Baseline*: will receive simple reports generated by CTAD's current software. An example baseline report is shown in Figure 4.
- *SkillSum-control*: will receive reports generated by SkillSum using a microplanning choice model which is based on the most common choices observed in two corpora (British National Corpus (BNC) and RST Discourse Treebank Corpus (Carlson 2002)).
- *SkillSum-ER*: will receive reports generated by SkillSum using a microplanning choice model which is based on our KA activities, and which we believe encodes appropriate expression choices for readers with limited literacy.

We will ask students to self-assess their current skill levels and interest in doing a skills course, both before and after they take the test and read the report. We will measure changes in self-assessment accuracy and interest in doing courses in the three groups. We will also ask some of the students comprehension questions about reports, and to read reports aloud; we will measure correctness and time taken to respond to comprehension question, and the time taken and errors made in reading reports aloud. Last but not least, we will ask students to express a preference between different versions of their report, and ask them for general qualitative comments and feedback.



English Skills

Thank you for doing this test.

You answered 19 questions correctly. More information.

You made some mistakes on the questions about writing.

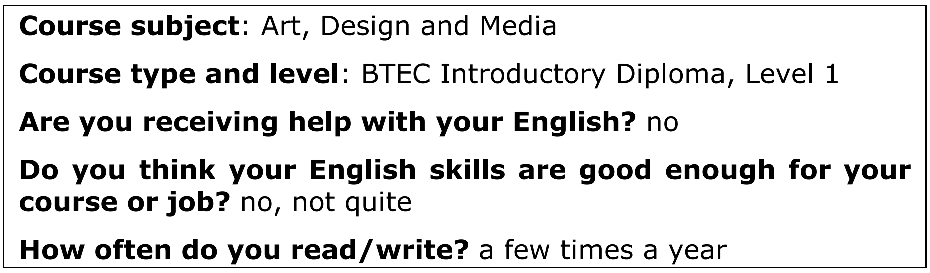
But you got most of the questions right where you had to read.

Your English skills seem to be okay for your Art, Design and Media course.

But an English class might help you, because you said you do not feel very confident with your reading.

Click here to find courses at ABC FE College.

Figure 1: Example report produced by SkillSum



Course subject: Art, Design and Media

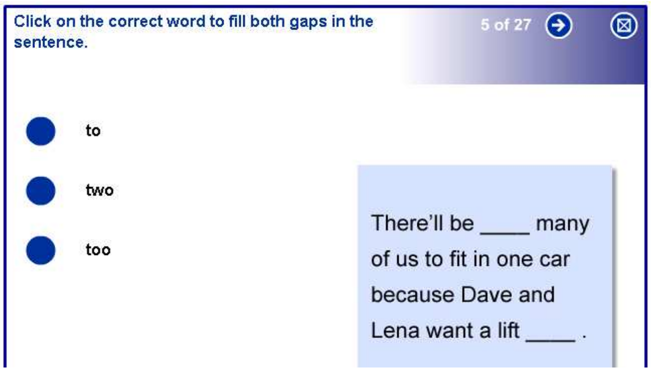
Course type and level: BTEC Introductory Diploma, Level 1

Are you receiving help with your English? no

Do you think your English skills are good enough for your course or job? no, not quite

How often do you read/write? a few times a year

Figure 2: Responses to background questions for Figure 1 student



Click on the correct word to fill both gaps in the sentence.

5 of 27

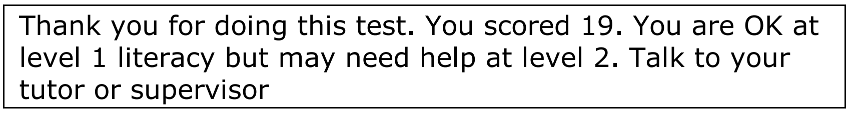
● to

● two

● too

There'll be ____ many of us to fit in one car because Dave and Lena want a lift ____ .

Figure 3: Example SkillSum assessment question



Thank you for doing this test. You scored 19. You are OK at level 1 literacy but may need help at level 2. Talk to your tutor or supervisor

Figure 4: Baseline report for Figure 1 student

4. Main challenges

In this section we discuss three general issues which are important in SkillSum and which we believe will be important in most (perhaps all) systems which attempt to generate feedback reports from SkillSum-like data.

4.1 Choosing Content from Limited Data

One of the main challenges in building a system like SkillSum is deciding what to tell people, especially as our data is very limited. In early versions of the system, we experimented with giving people detailed analyses of how good they are in specific areas such as grammar and punctuation; this was done by associating questions with specific skills. We even at one point discussed detailed diagnosis of incorrect responses; for example, we wanted to tell people not just that they had problems adding and subtracting, but that the problems were due to not being able to carry and borrow. However, basic skills experts we worked with were concerned about this because it was based on very limited data. For example, there are only 3 questions on grammar in the SkillSum literacy test, which is a very small amount data for telling people that their grammar skills are good or bad. Also, different people use different techniques; for example, some people do not carry when adding. Finally, pilot experiments showed that users (understandably!) became annoyed if we told them that they were poor at something which they thought they were good at; and also that people did not interpret words such as "*grammar*" and "*punctuation*" as we expected (for example, some people thought grammar mistakes included punctuation errors).

In other words, the diagnostic inferences we were making about people's skills were not robust, because they were based on very limited data. Communication of these inferences was also error-prone, since people interpreted words in unexpected ways. Finally, the cost of an incorrect inference was high, because incorrectly telling someone they were bad at something could annoy them or diminish their self-confidence, and incorrectly telling someone they were good at something might reduce their interest in getting help. Hence we decided to only give very high-level diagnostic summaries, accompanied (in the latest versions of SkillSum) by a list of the specific questions they got wrong.

We also initially wanted to give people detailed motivational information, explaining why improving their literacy and numeracy would benefit them, given their personal circumstances. We based this on a questionnaire which asked people about possible motivations, such as improving job prospects or helping children with their homework. But again we had major problems because these questions were vague, and hence did not provide much information; and also sometimes people again interpreted questions differently from what we expected. And again getting things wrong could anger people. For example, we originally thought we could say something like "*if you improve your English, you can help your children write reports for school*" if the user had ticked "helping children with homework" as one of his or her motivations; but in fact it is impossible to say this without knowing a lot about the current skills of the user and the user's children, and also the skills being taught by the children’s school. Furthermore, the above phrase might anger someone who thought she was already helping her children to some degree, and just wished she could do a bit more.

The current version of SkillSum, which is targeted towards students attending Further Education colleges, bases its motivational information on the requirements of the specific course that the person wishes to take. For example, it might tell someone that their English skills appear to be adequate for a Level 1 BTEC (British Technical Education Council) Diploma course in Art, Design and Media at ABC FE College. In the text shown in Figure 1, for example, the student’s score on the literacy assessment is above the threshold required for a Level 1 BTEC Introductory Diploma course; hence SkillSum tells her this (“*Your English skills seem to be okay for your Art, Design and Media course.*”). However, the student stated that she did not think her skills were adequate and that she is not receiving help with English, so SkillSum tells her that an English course might help her (indeed, her score indicates that she may have some problems with literacy). On the other hand, if the student had scored below the threshold, SkillSum would tell her that she might need help to bring her English skills up to the level required for her course.

Basing motivation information on the requirements of a course is not ideal because improving skills in order to complete a course is only one type of motivation, and may be less important than intrinsic motivations such as improving self-confidence and self-esteem (Kotler, Roberto, Lee 2002). It also means SkillSum must have detailed knowledge of the requirements of various courses. But from the data perspective, intended course is reliable and easily obtainable data, which is associated with specific literacy and numeracy constraints on students who wish to take the course; whereas most motivations we considered were difficult to obtain reliably, and also difficult to map to specific literacy and numeracy requirements.

The fact that the cost of mistakes is high (because it can annoy and/or demotivate people) certainly makes choosing content in SkillSum considerably more difficult. In contrast, the cost of making mistakes was much lower in the STOP system (also developed at Aberdeen) (Reiter, Robertson, Osman 2003), which generated personalised smoking-cessation leaflets. STOP had to choose which bits of encouragement and advice to include in its leaflets, and did this by analysing questionnaire data to attempt to identify the most useful encouragement and advice; but in general making a mistake in this regard and choosing less-than-ideal advice did not anger people, it just meant that the leaflets were a bit less useful than they might have been.

In short, it is difficult to choose content when the data is limited (and noisy), and the cost of a mistake is high; we suspect this is one of the main challenges for many systems which generate feedback reports based on short assessments. This is one area where human tutors, who have face-to-face dialogues with students and also have extensive knowledge of skills and motivations, do a better job than computer systems, and probably will continue to do so for the foreseeable future. But nonetheless SkillSum shows that it is possible to give more information than just a skill level and associated generic information which is not at all personalised; and we believe that further research will reveal other techniques for suggesting useful content in such contexts. One idea we would like to explore in the future is making the system more interactive, so that users can to some degree tell SkillSum what they want to know. Our experiments certainly show that students vary greatly in what information they want; some in fact want quite detailed analyses of their performance, while others just want reassurance that they are not “thick”.

4.2 Producing Texts for Low-Literacy Readers

SkillSum originated in a PhD project which focused on generating texts which could be easily read by people with poor literacy skills, and this remains one of the challenges of building the system. While this is especially important in SkillSum as most SkillSum users are people with below-average literacy, in fact any system which generates texts for the general public needs to make sure its texts are accessible to people with poor literacy. Unfortunately, 20% of UK adults have a “reading age” of 10 or less (in other words, they cannot read at the level expected of 11 year old children), and 6% of UK adults have a reading age of 6 or less (Moser, 1999). Similar percentages of adults in the US have poor literacy. The situation is a bit better in some other European countries such as Sweden and The Netherlands, but even in these countries about 10% of adults have problems with literacy (Carey, Low, Hansbro 1997).

We have described this aspect of SkillSum elsewhere (Williams and Reiter, 2005b), so we will only give a brief summary here. Basically we focus on microplanning choices, including lexical choice, aggregation (how many messages are realised in each sentence), sentence ordering, and choice of discourse cue phrases. For these choices, we have developed a set of rules (based on pilot experiments and KA with tutors) which we believe are appropriate for low-skill readers, and we experimentally compare texts generated with these rules to texts generated with rules which are based on the most common choices in a corpus.

We continue to actively work on lexical (word) choice in particular. In general, psychological research (Rayner and Pollatsek, 1989; Perfetti 1994) suggests that people find easiest to read those words which they are most familiar with (that is, the words they have used and encountered the most). The simplest way to estimate familiarity is to use frequency in a standard corpus such as the BNC. Unfortunately BNC frequency is not always a good predictor of word familiarity, in part because the BNC includes texts (such as academic research papers) which most English speakers never encounter. Another issue is ambiguity; common words tend to have many meanings, and it is not clear when this is acceptable, and when it is preferable to use rarer words which have fewer meanings and hence are less ambiguous. Skills tutors have also pointed out to us that the ease of reading a word can depend on its visual properties, such as how difficult a word is to sound out from its spelling, and indeed what the overall visual shape of a word is.

Another issue is individual variation. It is clear from our pilot experiments, and indeed from work we have done in other projects (Reiter and Sripada, 2002) that there major differences between individuals that affect lexical choice. These include

- *Variations in meaning*: For example, as mentioned in Section 4.1, different people interpret "*grammar*" in different ways.
- *Variations in connotations*: For example, "*teacher*" is a very common word, but it has bad connotations for students who disliked school; for such students it is better to use an alternative word such as "*tutor*".
- *Variations in language use and exposure*: For example, someone who has learned English as a second language may be more comfortable with semi-technical terms such as "*punctuation*" than a native English speaker who never tried to learn another language.

In fact, a general problem with our approach is that we use the same microplanning choice model for all poor readers, but poor readers are very diverse, and have "spiky" ability profiles (good in some areas, poor in others) (young children, in contrast, tend to have more uniform profiles, with similar performance in all areas). In future research we would like to explore using more specialised choice models, indeed perhaps using choice models which are tailored for specific individuals.

Finally, one point that has emerged very clearly in our pilot experiments with SkillSum is that content and structural choices are very important for readability. As above, we had originally thought of readability in terms of microplanning choices which affected how information is expressed, but not what information is communicated in a text. However, as is perhaps obvious in retrospect, our pilots showed that the length of a text (which is determined by its information content) is also very important; poor readers need short texts that don't overstrain their abilities. Also, poor readers need text that they perceive as being worth reading, otherwise they may not make the effort required to read them. Since people typically start reading a document from the beginning, this means that texts intended for poor readers should start with useful information; hence for example the content-free "*Thank you for doing this test*" sentence in the report shown in Fig. 1 should perhaps be at the end of the report, not the beginning.

4.3 Integration

Our initial vision of SkillSum was that it would be a stand-alone system, which people could access from home or community sites (such as libraries), without needing any support from human tutors or administrators. SkillSum might encourage users to sign up for a formal assessment, but this would be its only link with the existing assessment process.

We have changed our opinion in this respect, and now see SkillSum as an add-on to the existing assessment process, not something which is separate from it. Our current vision is that SkillSum will be deployed at FE colleges and other organisations (large employers, military, etc) that already test people's skills, and be used to provide extra feedback to people who the organisation has already decided to assess, and also to allow other people (who the organisation has not yet decided to assess) to get some information about their skills. In both cases, SkillSum is integrated into the current assessment process of the host organisation, and is seen as a tool for that organisation as much as a tool for the individuals being assessed.

Partially this is due to standard issues: all IT systems are of course more likely to be used if they integrate well with existing processes, and AI systems which may make incorrect inferences (in our case because of sparse data) are often deployed in contexts where human experts are "in the loop". Also, we had decided in any case to focus SkillSum on specific groups, such as students entering a particular FE course, rather than poor readers in general. Such a focus makes it much easier to generate motivational information, as described above. Also, focusing on specific groups reduces variation in subject's ability profiles, which makes it easier both to specify appropriate NLG choices, and to experimentally determine if our choices do indeed make texts more readable (Williams and Reiter, 2005b). In any case, once we had decided to focus SkillSum on groups instead of the general public, which made it much more natural to think of deploying SkillSum within an organisation instead of as a general tool for everyone.

Integration in this sense means that we should support tutors as well as students. We have experimented with generating separate reports intended for tutors, which are much longer and more detailed than reports for students, and which freely use technical vocabulary which students would not know. Another idea we would like to explore is giving tutors some control over the reports produced for their students.

A related point is that if SkillSum is targeted towards specific groups, we need to allow tutors (or developers) to encode the information needed by that group. For example, we will need to tell the system what courses students might undertake, and what level of literacy and numeracy is needed by students in the course. Also, as assessment tests are often adapted for specific groups (for example, a numeracy assessment for Hairdressing might focus on skills needed for Hairdressing, such as dealing with money), we need a mechanism for telling SkillSum about specific tests, and how they should be interpreted.

5. Future Work

There are a number of topics we would like to investigate in future work. Firstly, we would like to further explore individual variations in language use. It is clear that people are very different in their linguistic preferences, linguistic abilities, and linguistic history (what language they have seen and used), and we believe that a system like SkillSum could do a much better job at generating easy-to-read texts if it could adapt texts according to the recipient's linguistic preferences, ability, and history. We would like to empirically explore to what degree people do indeed have different preferences, etc; and empirically measure the impact of adapting texts to an individual's personal use of language.

Secondly, we would like to further explore motivational issues, again probably emphasising individual differences. People have very different motivations for enrolling in skills courses, and SkillSum would probably be considerably more effective if it had a better idea of what a user's motivations were; our STOP system for generating smoking-cessation letters (Reiter, Robertson, Osman 2003) would probably similarly have been more effective if it had a better understanding of users' motivations. Our experiences in both projects suggest that a coarse model of motivation (such as "I want to help my children", or "I want to be healthier") may not be very effective (see Section 4.1). We would like to develop techniques for obtaining a more detailed understanding of motivation, and see if this improved SkillSum's effectiveness.

Last but not least, we would like to make SkillSum more interactive, and also incorporate multimedia (audio and graphics) as well as written text. In theory an interactive version of SkillSum could give users control over what information they see (Section 4.1); audio might be a good media for communicating information to people with limited literacy but good oral English; and graphics might make the system more appealing and friendly to some users.

6. Conclusion

Making it easier for people to assess their literacy and numeracy skills, and indeed other things (such as their health), would be beneficial to society and valuable commercially. We have addressed this problem by trying to build a system which produces a feedback report which is short and easy to read, but nonetheless is more useful than existing "you scored N and are at level X" reports.

SkillSum was harder to build than we at first anticipated, largely because of the fact that we had very limited data about people to work with, and the fact that people are so variable in terms of their skills, motivations, and ability to read. We believe developing techniques for reasoning about people's skills, motivations, etc based on limited data is one of the major research challenges for this class of AI system. However, we also believe that in the longer term more data will in any case be available about individuals, because people are likely to accumulate large amounts of digital data about themselves (Fitzgibbon and Reiter, 2003). Hence while SkillSum-like systems will probably never have as much data about their users as they would like, we suspect that they will have much more data in 10 years time than they do today, because they will be able to access (if permission is given, of course!) large data sets that people have accumulated about themselves. In other words, we believe the long-term prospects of this type of system are very good.

Acknowledgements

Our thanks to the many people who have helped us in SkillSum, including in particular the tutors and students who have participated in our experiments. SkillSum is supported by the ESRC/EPSRC/DTI PACCIT programme under grant ESRC RES-328-25-0026.

References

1. J Burstein, M Chodorow, C Leacock (2003) CriterionSM Online Essay Evaluation: An *Application for Automated Evaluation of Student Essays. In Proceedings* of IAAI 2003, pages 3-10

2. S Carey, S Law, J Hansbro (1997). *Adult Literacy in Britain*. Office of (UK) National Statistics.

3. L Carlson, D Marcu, and M Okurowski (2002). 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.
4. Fitzgibbon and E. Reiter (2003). *Memories for life: Managing information over a human lifetime*. Grand Challenge proposal, published by UK Computing Research Committee (UKCRC).
5. P Kotler, N Roberto, N Lee (2002*). Social Marketing: Improving the Quality of Life (2nd Ed).* Sage.
6. B Lavoie and O Rainbow (1997). A Fast and Portable Realizer for Text Generation Systems. In *Proceedings of ANLP-1997*, pages 265-268
7. C Moser et al (1999) *Improving Literacy and Numeracy: A Fresh Start*. Available at http://www.lifelonglearning.co.uk/mosergroup/
8. M O'Donnell, C Mellish, J Oberlander, A Knott (2001)
   ILEX: An architecture for a dynamic hypertext generation system. *Journal of Natural Language Engineering*, **7**:225-250.
9. C Perfetti (1994). Psycholinguistics and Reading Ability. In M Gernsbacher (ed), *Handbook of Psycholinguistics*. Academic Press.
10. K Rayner and A Pollatsek (1989). *The Psychology of Reading*. Prentice Hall.
11. E Reiter and R Dale (2000). *Building Natural Language Generation Systems*. Cambridge University Press.
12. E Reiter, R Robertson, and L Osman (2003). Lessons from a Failure: Generating Tailored Smoking Cessation Letters. *Artificial Intelligence* **144**:41-58.
13. E Reiter and S Sripada (2002). Human Variation and Lexical Choice. *Computational Linguistics* **28**:545-553
14. E Reiter, S Sripada, and R Robertson (2003). Acquiring Correct Knowledge for Natural Language Generation. *Journal of Artificial Intelligence Research* **18**:491-516.
15. S Williams (2004). *Natural Language Generation of Discourse Relations for Different Reading Levels*. PhD thesis, Dept of Computing Science, University of Aberdeen.
16. S Williams and E Reiter (2005a). Deriving content selection rules from a corpus of non-naturally occurring documents for a novel NLG application. In *Proceedings of Corpus Linguistics 2005 workshop on using Corpora for NLG*, pages 41-48.
17. S Williams and E Reiter (2005b). Generating readable texts for readers with low basic skills. In *Proceedings of the 2005 European Natural Language Generation Workshop*, pages 140-147.