SkillSum: basic skills screening with personalised, computer-generated feedback

Sandra Williams[1] *and Ehud Reiter*[2]
[1]The Open University, [2]University of Aberdeen

Key words: *e-learning, literacy assessment, numeracy assessment, natural language generation*

Abstract

We report on our experiences in developing and evaluating a system that provided formative assessment of basic skills and automatically generated personalised feedback reports for 16-19 year-old users. Development of the system was informed by literacy and numeracy experts and it was trialled 'in the field' with users and basic-skills tutors. We experimented with two types of assessment and with feedback that evolved from long, detailed reports with graphics to more readable, shorter ones with no graphics. We discuss the evaluation of our final solution and compare it with related systems.

1 Introduction

O*ne in six* school leavers in the U.K. lack sufficient literacy skills to cope with the vocational college courses that are necessary for pursuing their chosen careers, and worse, *one in two* lack sufficient numeracy skills. The Skills for Life Survey commissioned by the U.K. Government in 2003 [4] assessed the basic skills (literacy and numeracy) of 8730 people. The survey authors estimated that around 16% of 16-19 years-olds were at *Entry Levels* or below in literacy and 50% of 16-19 year-olds were at similar levels of numeracy (see Table 1).

Entry Level 1 is the national school curriculum equivalent of attainment at age 5 - 7. Adults with skills below Entry Level 1 may not be able to write short messages to family or select floor numbers in lifts.	**Entry Level 2** is the national school curriculum equivalent of attainment at age 7 - 9. Adults with skills below Entry Level 2 may not be able to describe a child's symptoms to a doctor or use a cash point to withdraw cash.	**Entry Level 3** is the national school curriculum equivalent of attainment at age 9 - 11. Adults with skills below Entry Level 3 may not be able to understand price labels on pre-packed food or pay household bills.	**Level 1** is equivalent to GCSE grades D – G Adults with skills below Level 1 may not be able to read bus or train timetables or check the pay and deductions on a wage slip.	**Level 2** is equivalent to GCSE grades A* - C Adults with skills below level 2 may not be able to compare products and services for the best buy, or work out a household budget.

Table 1: Levels of literacy and numeracy, source: 2008 Skills for Life Survey [2]

A subsequent U.K. Government survey in 2008 [2] assessed the progress of the literacy and numeracy initiatives. Progress in numeracy was *substantially worse* than literacy, since they estimated that only 10% of adults who had GCSE qualifications[1] below grades A* to C in mathematics gained qualifications in numeracy compared with the 18% of adults with poor GCSE grades in English who gained literacy qualifications. An obvious first step towards persuading more people to sign up for courses is to assess their current skills. With

[1] General Certificate of Secondary Education (GCSE) are exams taken at U.K. schools, normally at 16 years.

between a sixth and a half of U.K. 16-19 year-olds requiring basic skills training, *manual* basic-skills testing, marking and feedback becomes impractical and this was the motivation underlying our research into *automatic* assessment and feedback in the SKILLSUM project.

SKILLSUM was a joint project between the University of Aberdeen and TRIBAL CTAD. It integrated a feedback generator from Williams' Ph.D. system [5] with CTAD's commercial tools for automatic assessment of basic skills (literacy and numeracy) to produce a *formative* application, i.e., one that assesses a learner's current skills and gives feedback and recommendations for further study. Both components of the system communicated with a secure relational database that stored assessment questions, users' login details and answers to the questions. All three were installed on a web server so that many users could access it simultaneously over the Internet (see fig. 1).

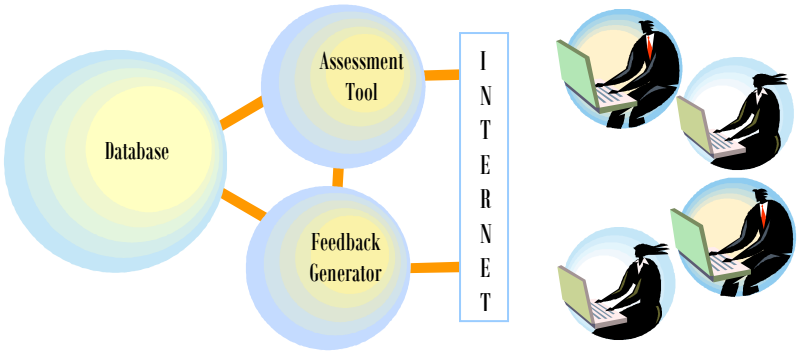

Figure 1: the web-based assessment and feedback application developed in SKILLSUM

Automatic, on-line assessment and feedback are speedy (a user does not have to make an appointment and wait for the date to arrive), question marking is accurate, and feedback reports appear almost instantaneously. Furthermore, feedback can be printed out as a permanent record, perhaps to discuss with a friend or family member. Privacy can be another important benefit especially in literacy assessment. Indeed, a hospital study on patients' attitudes to literacy assessments for health education found that *'all participants described exposure of their reading limitations as a risky situation, to be avoided whenever possible'* [1]. If users are able to access on-line tests from home, or some other private location, then they will avoid such exposure.

Whilst technical details of the application have been published elsewhere [7], this paper focusses on three fundamental questions that were addressed in the project and that we feel will be of interest to developers of other assessment and feedback systems:

1. Which groups of users can use software tools to assess their basic skills without help?
2. What types of basic-skills assessments are appropriate for use independently at home or in a college with a tutor present?
3. What is appropriate content and length for feedback reports about basic skills assessments?

These questions are addressed in sections 4-6; but first we introduce the SKILLSUM assessment tools (section 2) and the generated reports (section 3).

2 Basic skills Testing

At the beginning of the project, we used a detailed diagnostic tool, but we soon moved on to a less accurate but shorter screener tool. Both were objective, i.e., with only one correct answer for each question.

2.1 Diagnostic Assessments

Detailed diagnostic tools assessed the literacy and numeracy skills listed in Table 2. The literacy assessment contained 90 questions overall and the numeracy assessment 155; however, a user would *not* normally see all questions. An algorithm administered harder or easier sets of questions depending on a user's performance on previous sets. At the beginning of an assessment, very low-level questions were administered; that is, alphabet recognition in literacy assessment (see fig. 2, left-hand screenshot), or digit recognition questions in numeracy assessment. In these initial questions, if a user gave incorrect answers to five questions in sequence, or received a score of seven out of ten, or less, then the system would assume that there was a problem and would exit.

LITERACY SKILLS	NUMERACY SKILLS
Letter Recognition	Whole Numbers
Sentence Completion	Fractions
Punctuation and Capital Letters	Decimals
Word Ordering	Percentages
Spelling	Money
Form Filling	Time
Skimming and Scanning	Measures
Listening	Shape and Space
	Handling Data

Table 2: Basic skills that were tested and diagnosed by the diagnostic tools

Question formats were multiple-choice (fig. 2 left-hand screenshot), drag-and-drop (fig. 2 right-hand screenshot), number entry, multiple-select and timed viewing of short passages of text. Algorithms scored all skills in table 2, each was assigned a level which accorded with the score achieved; for instance, a score of more than seven out of ten for digit recognition would receive Entry Level 1 (a higher level for digit recognition alone was not possible) and a set of questions at a slightly higher level would be administered. A final level for overall literacy or numeracy was calculated from the scores attained for each question set attempted.

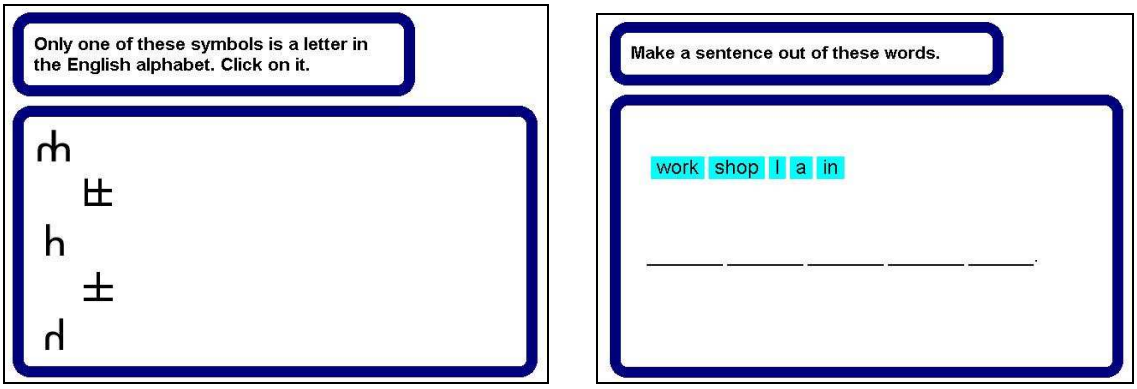

Figure 2: Screenshots from the literacy diagnostic tool

2.2 Screeners

Level	Screener Score
Working towards Entry Level	1 – 8
Working towards Level One	9 – 13
Competent at Level One	14 – 23
Competent at Level Two	24 – 27

Table 3: Mappings between Screener Scores and levels of basic skills (see Table 1)

These are quick and simple skill-checker tools, one for literacy and one for numeracy, each with twenty-seven questions. Whilst covering a fairly wide range of basic skills, they are not diagnostic assessments, since they do not give enough information about proficiency in

each skill. Nevertheless, they can indicate skill levels well enough for rough mappings between overall scores and the levels of basic skills, as shown in Table 3.

Questions were similar formats to the diagnostic assessments (see fig. 3). Like the diagnostic assessment tool, the screener tool exits at a point where it is considered that a user is having severe problems, i.e., whenever a user supplies five sequential incorrect answers.

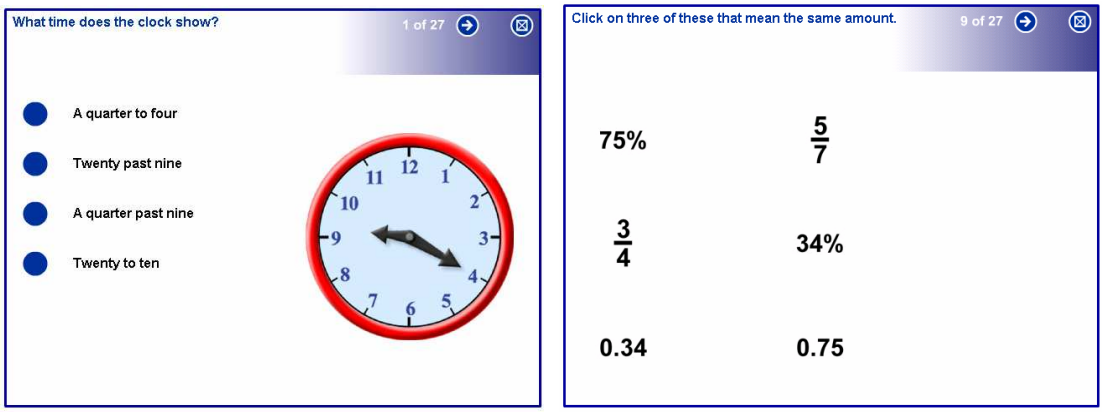

Figure 3 : Screenshots of questions from the numeracy screener tool

3 Feedback generation

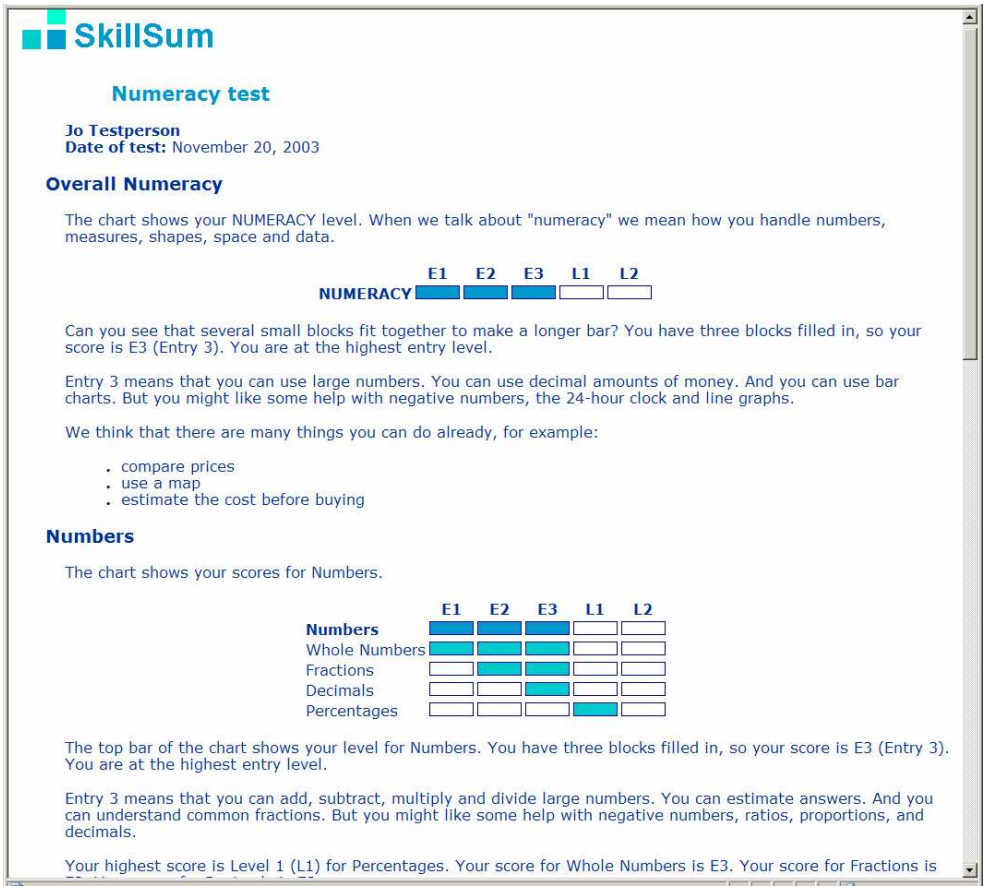

Fig 4: A long report produced by the feedback generator (about ¼ of the report is displayed)

Technical details about the Natural Language Generation (NLG) system used to produce feedback reports have been published elsewhere [7][5]. We trialled three types of feedback:

- canned messages (no NLG technology)
- long, detailed feedback reports – see fig. 4

- short feedback reports – see fig. 5

Canned messages were output by the screening tools before they were linked to the feedback generator. Very short messages with the overall score and level achieved were inserted between sentences of thanks and instruction to talk to a supervisor, for example:

Thank you for doing this test.
You scored 19.
You are OK at Level 1 literacy.
Talk to your supervisor.

As for generated feedback, long, detailed feedback reports were generated at the beginning of the project for the diagnostic assessments (see fig. 4), but later, shorter feedback reports were generated for the screener tests (see fig. 5).

Content of *long reports* was heavily based on information from the Basic Skills Core Curriculum (BSCC) [3], including results for the skills that were assessed, with levels attained and simplified example activities at the current level and the next level.

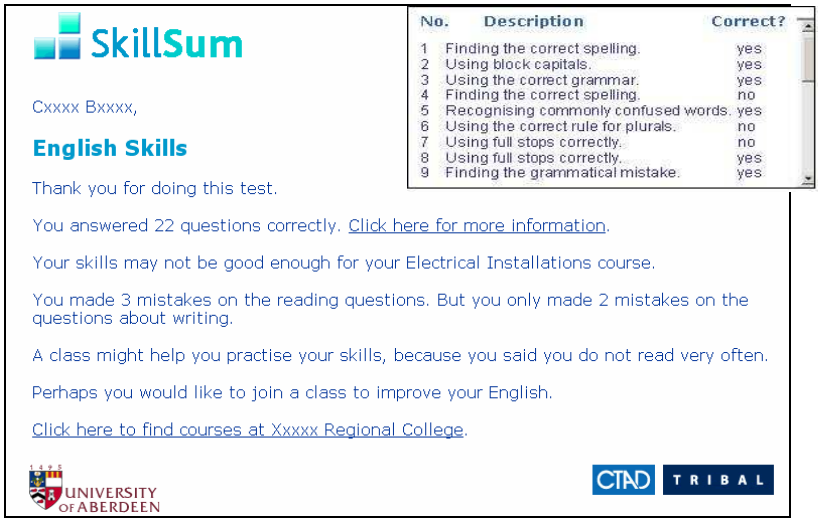

Fig 5: A short report produced by the feedback generator (inset showing detailed results)

Short reports had *no* references to BSCC levels and activities. They contained the overall score, an optional list of correctly/incorrectly answered questions, an opinion on whether the user's skills would be sufficient for his/her course and advice about improving skills. Each user completed a short Web form before the test. It asked for the user's course, his/her reading or calculation habits, and his/her confidence in his/her skills. This information was used to personalise the generated report.

4 Which groups of users can use software tools to assess their basic skills independently?

During trials, a number of types of users tried the system:

A. **Users with special needs (physical disabilities and dyslexia)**: Five people with physical disabilities tried the application in May 2004. Many of them could not use the system without special equipment; for example, some were unable to use a mouse or keyboard. A few people suffered from dyslexia which prevented them from skim-reading short text passages fast enough (passages were displayed for only 20 seconds).
B. **Users with very low levels of literacy and numeracy**: People with very low levels of literacy and numeracy tended to feel let down when the *screener test* closed down if they answered five sequential questions incorrectly. Basic-skills experts stressed that such people should not take the test independently, only with support from professionals. Also, reading and understanding the text of the questions was

challenging for them (this problem was avoided in the *diagnostic assessment* where an audio file with recorded human speech was played to read the question texts).

C. **Users with moderately poor literacy and numeracy:** People with moderately low levels of literacy and numeracy, on the other hand, tended to cope very well with using the application independently.
D. **Users who had never used a computer**: In March 2005, a study with 14 older basic-skills students included some who had never use computers before. Following some instructions on how to use a mouse, they coped well with the application.
E. **Users who were extremely computer literate**: The majority of our other studies (one with 8 school leavers on skills-for-work courses, one with 10 school leavers on the same course and one with 60 16-19 year-olds) were with younger people, usually people in their late teens, who were all highly computer literate and had no problems with accessing our application [7].

Because A, B and D types had difficulties with our application, our criteria that people could handle the tool were that they belonged to both C and E groups. That is, people who are computer-literate and not at the very lowest level of basic skills, but nevertheless with skills that may fall below the requirements of vocational training courses (hairdressing, building, nursery care, etc.), i.e., people at Entry Level 3 or above (see table 1).

Our final evaluation of SKILLSUM was in collaboration with a Further Education College during induction week with 192 newly-enrolled 16-19 year-old students – the entire year's intake of students took part unless they had timetable clashes or tutors already knew their skill levels were below Entry Level 3. Around 20 students at a time took the screener test under exam conditions with minimal supervision. Students with special needs were given support.

5 What types of basic-skills assessments are appropriate for the home or for colleges?

Experts in basic skills advised us that the *diagnostic assessment* was unsuitable for unsupported, independent self-assessment because it was too long (one person took four hours to complete it and most users took at least an hour) and the results were hard to understand without help. So, detailed diagnostic assessments are more suitable for use *within a college or centre* where basic skills courses are taught and where they can be administered by a trained tutor. In fact, CTAD already market diagnostic assessments that are integrated with basic skills teaching modules and learning plans for use by tutors in colleges.

Since the motivation underlying our project was to persuade more people to sign up for basic-skills courses, it was important that the skills test we used should be quick to complete even if it gave only a rough indication of skill levels. These considerations, together with the difficulties encountered with the longer assessment, led the experts to recommend that *shorter, screener tests* would be more suitable for *independent use without a tutor present*. This fitted well with our chosen group of users. As will be seen in section 6, the screeners and feedback reports did indeed convince students with low skills that they needed help.

6 What is appropriate content and length for feedback reports?

When deciding on the content and length of feedback reports, we had to reconcile expert knowledge from basic skills professionals with often conflicting evidence from studies with real basic-skills learners and tutors 'in the field'.

We acquired knowledge about appropriate content from experts in literacy and numeracy via interviews, from sample feedback that they had written, and from corrections they made to reports the prototype had generated. Initially, experts suggested that feedback should be based on information in the BSCC [3], in particular that there should be explanations of the levels achieved and examples from the BSCC of real-life activities both at the users' current level

and at the next level. These suggestions were implemented in the feedback generator that output *long reports* (fig. 4). However, many participants in our studies (people with low basic skills) had problems these [6], and in interviews with people who received them, many said that although they liked being presented with evidence of things that they could do already, they did not like being told they could learn activities at the next level that they felt they already knew. Other users said that they did not understand the BSCC levels or the graphics. Faced with this evidence, we decided that *shorter reports* would be more readable and that content referring to levels and activities from the BSCC and graphics should be removed.

Experts were against telling users which questions were answered correctly and incorrectly because they thought that users would also need explanations. However, in the light of evidence that many users asked about their scores, in *short reports* we added a link to a list of questions with scores (fig. 5) allowing users to decide whether to view them or not.

We compared *short reports produced by the feedback generator* (see fig. 5) with *canned messages* (see the example at the beginning of section 3) in a pilot June 2005 study. We asked 15 basic-skills students (who had completed the screener test and had seen both types of feedback) which of the two they preferred. Their answers showed an overwhelming preference for generated feedback (13 people) over canned messages (2 people), significant with $p<0.008$ in a binomial test. People who preferred generated text said: *"It explains better. Tells you more about what you got, what you were wrong on." "you get more information - it gives you a reason"* Those preferring canned message said: *"Short and simple.", "to the point."* [7]

In *short reports* we also personalised content as described in section 3. Our final evaluation with 192 students in September 2005 investigated whether those receiving personalised *generated reports* increased their understanding of their skill levels compared to people who received *canned messages*. This was measured by asking them to answer the same question both before and after taking the screener test and receiving one of the types of feedback: *"Do you think your English/Maths Skills are good enough for your course?"* They indicated their answers on a scale ranging from "yes" to "no". We recorded whether participants changed the slider in the right direction (towards "yes" if their skills were at least equal to the college requirement for their course, or towards "no" otherwise). Significantly more people who read *generated feedback reports* moved the slider in the right direction compared to those who read the *canned messages* (significant at $p<0.02$ in a χ^2 test) [7]. We believe that this was because generated feedback was personalised by mentioning whether a user's skills were good enough for his/her course, whereas canned-messages only mentioned the overall score and level.

7 Related Applications

At the time of our project (2003-2005) two other systems were available:

- *iAchieve At Home*[2]
- *Read Write Plus* National Test application[3]

iAchieve reports mention the score and compare it with the peer group. A list of questions has ticks or crosses to indicate correct and incorrect answers. Another screen shows a complex diagram indicating the level on a rainbow-coloured scale of "achievement" shown in parallel to a scale of "difficulty" with correct/incorrect questions arranged along it. Two lists below this diagram indicate "pleasant surprises"— questions that had been answered correctly *above* the learner's current level, and "further work necessary"— questions that had been answered incorrectly *below* the current level. Such a complex diagram would be hard for students to understand and adverse peer-group comparisons could upset some students.

[2] See www.iachieveathome.com.au
[3] See www.dfes.gov.uk/readwriteplus/learning

Read Write Plus reports contain vast amounts of data that is poorly explained. The student's report is 5 pages; it classifies skills as *emerging*, *consolidating* or *established*, but it does not explain what these terms mean. It mentions areas to work on in a separate learning plan that can be ten pages long. A tutor's report is even longer, e.g., thirty-two pages. One tutor advised us that this amount of data for each student is information overload!

8 Conclusions and Future Work

Referring back to the three research questions that we posed in the introduction, we can sum up our findings by saying:

1. The group of users who can use software tools to assess their basic skills without help are people who are computer-literate, who have specially-adapted computers if they have physical disabilities, and who have moderately low, but not very low levels of basic skill (Entry Level 3 or above on the BSCC scale).
2. The type of basic-skills assessment tool that is most appropriate for use independently at home is the quick skills-checker or screener. The type that is best for use in a college with a tutor present is the in-depth diagnostic assessment tool.
3. Appropriate content and length for feedback reports about basic skills assessments depends on the environment in which the report is to be read. For the scenario of our project, we would recommend short reports containing overall scores, advice about basic skills courses, optional lists of question scores and simple, personalised information that is relevant to the situation of the user

Subsequent to the SKILLSUM project, we have begun research into one aspect of Natural Language Generation that is somewhat related: communicating numerical information. A new project, NUMGEN[4] is investigating how to generate as many linguistic variations as possible for proportional quantities (e.g. "0.4987", "almost one in two", "nearly half", and so on) and to link the choice of numerical expression to users' level of numeracy so that different users will be able to understand them and obtain more information than they might otherwise.

References

[1] Brez, S.M. and Taylor M. (1997) Assessing literacy for patient teaching: perspectives of adults with low literacy skills. Journal of Advanced Nursing, 25:1040–1047.

[2] Burr, Tim (2008) Skills for Life: Progress in Improving Adult Literacy and Numeracy. Comptroller and Auditor General, National Audit Office, London.

[3] Steeds, Andrew (Ed) (2001) Adult Literacy core curriculum including Spoken Communication. Tribal CTAD for The Basic Skills Agency. ISBN 1-85990-127-1

[4] Williams, J., S. Clemens, K. Oleinikova and K. Tarvin (2003). The Skills for Life survey: A national needs and impact survey of literacy, numeracy and ICT skills. The department for education and skills, research report no. 490, HMSO.

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

[6] Williams, Sandra, and Ehud Reiter (2005). Generating readable texts for readers with low basic skills. 10th European Workshop on Natural Language Generation, 140-147.

[7] Williams, S. and E. Reiter (2008) Generating Basic Skills Reports for Low-Skilled Readers. Journal of Natural Language Engineering, Cambridge University Press Preprints.

Author(s):

Dr Sandra Williams, Computing,
The Open University, MK7 6AA, U.K.
s.h.williams@open.ac.uk

Dr Ehud Reiter, Computing Science,
University of Aberdeen, AB24 3UE, U.K.
e.reiter@abdn.ac.uk

[4] See http://mcs.open.ac.uk/sw6629/numgen