Uppsala Master's Theses in Computing Science 292
Examensarbete DV3
2005-01-31
ISSN 1100-1836

Content determination for reports aimed at adult literacy learners

Nava Tintarev

Information Technology
Computing Science Department
Uppsala University
Box 337
S-751 05 Uppsala
Sweden

http://www.csd.abdn.ac.uk/research/skillsum/

This work has been carried out at the University of Aberdeen,
in collaboration with CTAD.

Supervisor: Ehud Reiter, University of Aberdeen
Examiner: Mats Dahllöf
Passed:

Date Signature

Abstract

Using a pipeline NLG (Natural Language Generation) architecture the SkillSum/GIRL project generates individualized reports for poor and good adult literacy learners. The original aim of this project was to extend user models beyond poor and good learners into groups such as “dyslexic” and “aphasic”; and to linguistically adapt generated reports accordingly. Knowledge acquisition was based on cluster analysis (SPSS) of learner demographics and the results of an online screener. Results repeatedly showed two clusters of good and poor learners. Concurrently, the importance of motivation was highlighted in an experiment with real learners. The aim was then changed to motivational profiles, and generating texts in order to encourage more people to acknowledge basic skills problems. The focus of this report is therefore document planning, and more specifically content determination. Five adults diagnosed with dyslexia evaluated the resulting generated reports. Four out of these five assessed the new report to be an improvement compared to the control report.

Contents

Acknowledgements

First and foremost an enormous thank you to my supervisor at Aberdeen University, *Ehud Reiter* as well as *Sandra Williams*, whose work I have extended, for all their support, advice and inspiration. They have always made time to answer questions, and resolve motivational struggles. Working with them has been an honor and a pleasure, in all the different facets of research work; from data analysis to contacting research subjects, to writing up articles and reports.

Also thanks to *Liesl Osman* for invaluable help with statistical analysis and interpretation. I am grateful to my examiner *Mats Dahllöf* at Uppsala University for the support and ease of communication despite the limitation of electronic lines. Thanks to *Uppsala and Aberdeen Universities* as well as CTAD for allowing this type of collaboration, and *CSN* for arranging funding. Moreover, I greatly value the help of *Jan Johnson and members of the Aberdeen Dyslexia Group* for their time and feedback.

And last, but certainly not least, I'd like to extend a great big thank you to *all the other students* at the department for their help and support, be it finding a good editor, a chocolate bar or a break at the Bobbin when I most needed a break.

Chapter 1

Introduction

This project is an extension of the SkillSum System: Automatic Generation of Personalized Basic Skills Summary Reports (Williams, 2004a), an online adult basic skills assessment and automatic report generator. The system is an ongoing collaboration between Cambridge Training and Development Ltd (CTAD, 2005) and the University of Aberdeen, Scotland. The main aim of the system as described on the university website is :

"...to develop an automatic assessment and reporting tool for adult basic skills (literacy and numeracy). The tool will be a web-based system that allows people to take a basic skills assessment in a library or community center, or indeed in their own home. When the test is completed, the tool will produce a report for the user describing his or her skill level and problems, and suggesting actions he or she could take to improve basic skills." (Aberdeen, 2004).

Moreover the main objectives are defined as :

Practical *"Encourage more people to acknowledge basic skills problems and seek assistance. This technology should be commercially attractive to organizations that are concerned about poor literacy and numeracy levels, such as colleges, employers, and the prison service."*

Scientific *"Develop techniques for generating appropriate reports for people with poor basic skills. Such people need "easy to read but not childish" texts, which respect their intelligence but at the same time are understandable and tailored to their literacy and numeracy levels."*(Aberdeen, 2004)

1.1 Original aims

The SkillSum system prior to this project already tailored reports for two levels of readers, good and poor. However in line with the scientific objective (Section 1) there was an interest in more detailed modeling of users. Although the SkillSum system consists of assessments for both literacy and numeracy, the focus of this project is specifically on the literacy assessment. The literacy level, rather than numeracy level, of the learner could be considered more relevant in making linguistic choices for the generated report. The aim of the project was sub-divided as follows:

1. Analyze data sets of adult learners doing two literacy skill assessments (27 questions and 90 questions respectively), looking for groups of learners with similar patterns. This would be roughly similar (but much simpler) than the ARCS analysis conducted by the National Institute for Literacy (National Institute for Literacy, 2004), which is further discussed in Section 2.2.

An example of such groups are dyslexics, aphasics and non-native English speakers. An alternative hypothesis was to suggest a relation between results and learning levels as defined by national or even international standards, such as the National Core Curriculum (Basic Skills Agency, 2001).

2. Build up qualitative profiles of the clusters. This may involve talking to learners and tutors. For example, a profile for the "dyslexic" cluster would describe dyslexia.

3. Based on the profiles, create:

   (a) rules which classify learner into clusters, e.g. rules classifying learners as dyslexics based on their assessment results

   (b) rules which suggest appropriate content for the generated reports, e.g. suggest useful things to tell dyslexics in reports.

4. Implement the rules in the context of SkillSum, using Java.

5. Test the modified SkillSum system with learners

1.2 Modified Aims

The aims were modified due mostly to insufficient user demographics (see Section 3.4). A concurrent experiment described in Chapter 4 lead to the attainment of user information regarding motivation, and the focus of the content determination was changed to motivation. The aim of this report hence changed to the practical objective of the SkillSum project, i.e. to encourage learners to acknowledge basic skills problems and seek assistance.

1.3 Natural Language Generation

The SkillSum project, together with the extension which this project constitutes, is regarded as belonging to the sub field of artificial intelligence and computational linguistics known as Natural Language Processing (NLP). Natural Language Generation, henceforth referred to as NLG, is a further division of this field. NLG systems produce texts that are understandable as well as appropriate in a language, commonly English, from non-linguistic input. Hence the term *natural* refers to the type of language used by humans, rather than formal and programming languages.

NLG differs from *NLU* (Natural Language Understanding) in that while NLG systems map in one direction; from meaning to language, NLU systems map in the other; from language to meaning.

While input varies from one NLG application to the next, it is generally unambiguous, well-specified and well formed. This is unlike the input to NLU systems which tends to be highly ambiguous. In our case, input of learner data and demographics can be coded in an unambigious manner. The main concern for NLG applications is instead *choice* on a number of levels such as content selection, lexical selection, sentence structure and discourse structure for generating the output. Our output is the generated report. *Content selection* regards what information to convey, including defining the boundaries for the range of input and output. *Lexical choice* concerns how to convey this information e.g. what words or phrases to use. Choices on the level of *sentences* decide how these words and phrases are put together in a sentence e.g. where to use commas and phrase order. *Discourse level* choices depict the relation between these sentences. These types of choices will be discussed in further detail in Section 1.4).

1.3.1 Pipeline architecture

There are a number of NLG architectures. The one used in SkillSum is linear and is commonly known as the pipeline architecture. It has three components (Reiter and Dale, 2000), (Reiter and Dale., 1997), (Jurafsky and Martin, 2000):

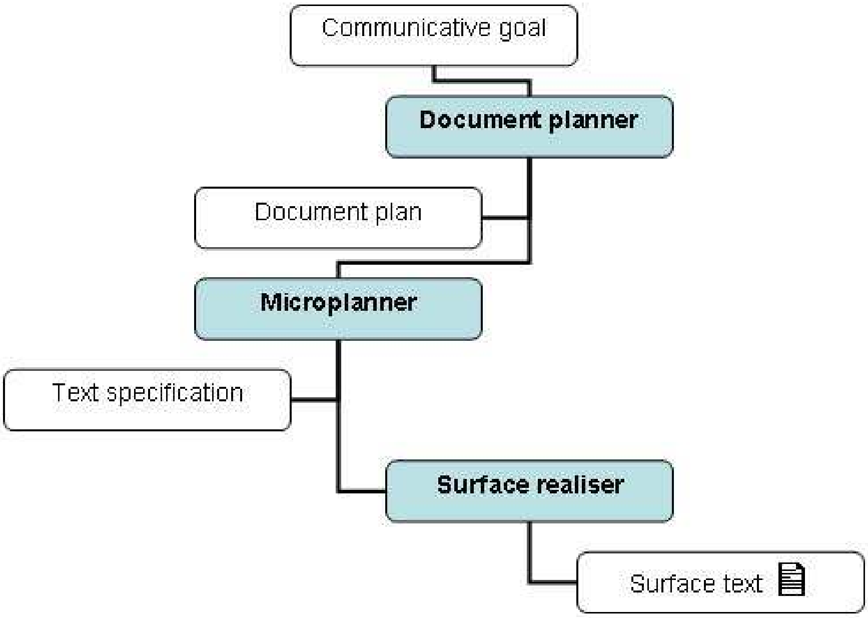


Figure 1.1: Pipeline architecture

1.3.2 Document Planning

The focus of this project lies within this phase. Document planning decides the content and structure of the generated text. Knowledge acquisition (KA), as described in Section 2.4, is a pre-requisite for this phase. In our case it is important to understand the domain of adult learning, and specifically how to communicate with adult literacy learners in order to decide the meaning of the conveyed message. This means that choices made in the document planner are made on both the level of content selection and on the lexical level.

1.3.3 Microplanning/Discourse planning

The microplanner decides how information and structure should be expressed linguistically. This was the focus of GIRL (Generator for Individual Reading Levels) the PhD project affiliated with SkillSum, see 2.1.2. The microplanner transformed discourse representations from hierarchical tree structures into ordered lists of individual sentence structures. This means choices in a microplanner are made on both sentence and discourse levels.

1.3.4 Surface realisation

Generates the actual, and grammatically correct, text from the linguistic structures created in the two previous phases.

1.4 Report structure

Chapter 2 on related work will introduce the SkillSum system in further detail. This includes the motivation for commencing this project and the progress up to date. Special focus has been placed on describing the role of the microplanner in terms of decisions primarily on sentence and discourse levels. Also, different methods for content selection will be discussed, as they are relevant to my extension of the system. For the same reason, relevant background in adult learning and motivation will be supplied as well.

Chapter 3 will include a summary and description of the initial results and difficulties. It will also supply the rationale for the content selection method used in this project. Combining these results with the experiment mentioned in *Chapter 4* will lead to modified aims. These combined results will include rules that define the motivational profiles described in *Chapter 5.1.* These profiles will distinguish groups of learners as well as suggest what type of information should be communicated to each group.

The next chapter, *Chapter 6*, offers an explanation of the Java implementation. This also includes a representitive example of how the input can be used to generate the learner report.Finally this project will be evaluated in *Chapter 7*, and concluded with suggestions for improvement and further work in *Chapter 8*.

Chapter 2

Related Work

2.1 SkillSum

The SkillSum system consists of two components; the assessments and the report generator. For this project the focus is on the literacy rather than the numeracy assessment. Consequently only the literacy assessment and literacy learners will be referred to from here on.

2.1.1 Assessment

SkillSum initially started with *TargetSkills*, a full 90 questions assessment. The literacy assessment contained nine modules with ten questions each; Spelling, Skimming and Scanning, Letter Recognition, Word Recognition, Sentence Completion, Form Filling, Punctuation and Capitals, Word Ordering and Listening.

After initial testing it became clear that this assessment although thorough, could require prolonged testing. For example, some dyslexics needed over two hours, and learners with concentration difficulties found it difficult to sit through the whole test.

Table 2.1: Question types - Screener

Type	**Number**
Spelling	7 (1)
Sentence Completion	1
Punctuation	5 (1)
Grammar	3 (1)
Skimming and Scanning: gist	5
Skimming and Scanning: info	6
Total	24 (27)

A shorter online screener, *SkillSum*, was thus created. Initially it contained 24 questions, later appended with three “non-scoring” questions in order to increase the confidence of the learner. The current test therefore contains 27 questions. Table “Question types - Screener” shows how these questions are laid out. The numbers in brackets represent the three “non-scoring” questions; one in each of the categories spelling, punctuation, grammar. The level of the test is moderately high, i.e. around Level 1, and not aimed at the lowest literacy levels.

See Section 2.2.2 for a short description of the differences between the levels used in this report. Appendix B also contains screen shots of two question used in the literacy assessment.

2.1.2 Report generator

Previous research (Williams, 2004b) investigated the effects of sentence and discourse-level decisions on good and poor adult readers, in GIRL, a Generator for Individual Reading Levels.

GIRL looked specifically at six discourse level features: **ordering** of discourse relation text spans, choice of between-text-span **punctuation**, choice of discourse **cue phrases**, **position** of discourse cue phrases, **length** of the first text span, and length of the second text span.

A *text span*, can be simplified down to a phrase or sentence. Both can be connected via discourse relations. *Discourse* in this case refers to sets of related sentences rather than sentences in isolation. Examples of discourse therefore include both monologues; characterized by communication flowing in one direction, and dialogues; where different participants take turns listening and speaking. Discourse can be both spoken and written (Jurafsky and Martin, 2000).

Discourse *cue phrases* are also known as discourse/linguistic markers, cue words, and discourse connectives. Their function is to tie together sentences. 'For example', 'but', and, 'although' are all discourse cue phrases. Cue phrases may be placed in the beginning and end of sentences, as well as in the middle. The above mentioned choices were based on three corpus analyses (See Section 2.4.2 for more about corpus analysis), including edits made by expert writers, feedback reports by literacy tutors and a larger analysis of a larger text corpus annotated with discourse relations (RST-DTC).

These resulted in a selection of the 8 most common discourse relations, as well as a selection of cue phrases for each. These relations were concession, condition, elaboration-additional, evaluation, example, reason and restatement. Each relations will be explained and exemplified in Section 6.6.2 in the chapter on implementation. To clarify, the main work in GIRL was centered in the microplanner where choices of discourse cues were represented as a constraint satisfaction problem (CSP). A CSP solver was incorporated in the microplanner to generate all "legal" ways for realizing each input discourse relation. The end result of the corpus analyses and pilot experiments was a set of rules for scoring solutions output from the CSP solver.

In the past it has been difficult to decide what information to include in a generated report. The type and quantity of information regarding the formal level of a user as defined by the National Curriculum (U. K.) (Basic Skills Agency, 2001), or IALS (Carey, Low, and Hansbro, 1997) has been an ongoing topic of discussion. Also feedback regarding specific questions, i.e. if the learner should be explicitly informed about what questions they answered incorrectly, and what the correct answer was, has been an question continuously brought up.

Prior to my involvement, the system summarized the result with the total score, and describes one strength, and one area of improvement. If several choices were available, the hardest area was mentioned. The report also suggest generic motivational texts and advice. An example of the "old report" can be found in Appendix B.

GIRL's output was evaluated with thirty-eight users including both good readers and poor readers. The evaluation methodology involved measuring reading speed, comprehension and eliciting judgements. The results, although not statistically significant, indicated that the algorithms produced more readable output and that the effect was greater on poor readers. A later experiment with sixty poor readers (Williams and Reiter, 2005) produced a significant difference in reading times for the poor reader model compared to the good reader model.

2.2 Adult literacy

2.2.1 The extent of the problem

An English study from 1999 suggests that as many as one out of five adults in the country has problems with reading. The literacy level of these adults is lower than what could be expected of an 11-year-old child. It also means that almost 7 million adults cannot locate the page reference for plumbers given the alphabetical index of the Yellow Pages (Moser, 1999). A study by the international *Organisation for Economic Co-operation and Development* confirms this number. The UK also ranked as 14th out of 20 in a list charting the percentage of 16 to 65-year-olds who read a book at least once a month (OECD, 2000). The extent of these problems has been a main motivator for the SkillSum project.

Another English survey by the Department for Education and Skills (DFES) took a look at literacy and numeracy needs in England (Williams, 2003) and found that very few adults regard their reading, writing or maths skills as below average. Even among those with very poor literacy, 54 percent said their everyday reading ability was very or fairly good, and only 2 percent thought that their weak skills hindered their job prospects or led to mistakes at work.

Therefore it seems plausible that many people overestimate their skills. These adults either do not realize the negative effect of their weak skills, have found jobs that demand only the appropriate level of skills, or have developed coping strategies so their limitations are not exposed (Williams, 2003). So although many adults could benefit from basic skills training, many adults do not acknowledge limited skills and fewer are willing to participate in courses. For these reasons the SkillSum project has repeatedly suffered from a lack of representative test subjects.

2.2.2 Levels

Literacy levels can be defined in a number of ways, but in other sections I will restrict references to the levels used in the Adult Core Curriculum as defined by DfES (Williams, 2003), (CTAD, 2005). The national standards for adult literacy and numeracy are specified at three levels: Entry level, Level 1 and Level 2. "Entry level" can be further divided into the three sub-levels; Entry 1, 2, and 3. These are not used by the assessment.

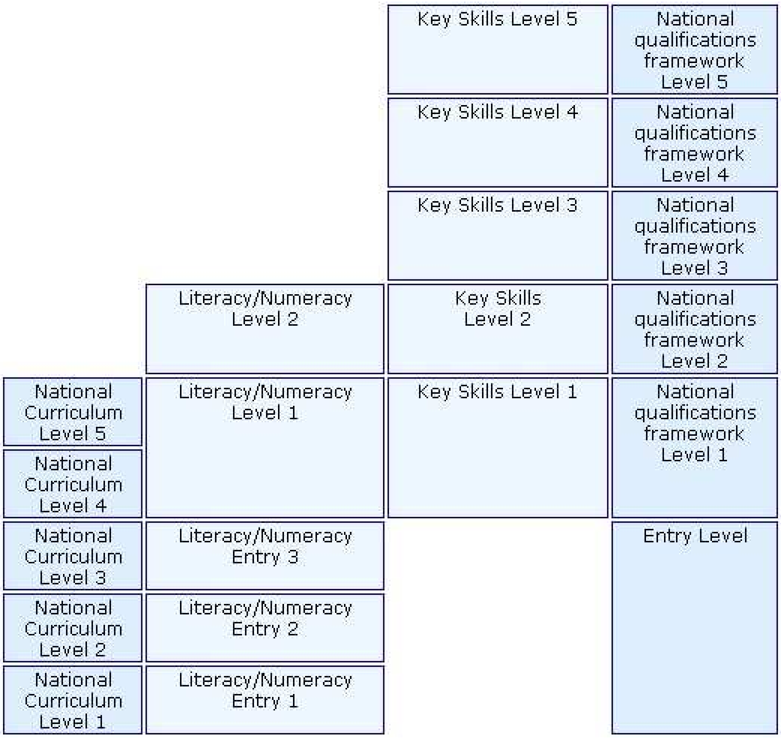
Table 2.2: National Standards for literacy

Level	**At this level, adults will be able, for example to:**	**Age level equivalent**	**Vocational level equivalent**
Entry 1	Read and obtain information from common signs and symbols.	Age 5	(no equivalent)
Entry 2	Use punctuation correctly, including capital letters, full stops and question marks.	Age 7	(no equivalent)
Entry 3	Organise writing in short paragraphs.	Age 9	(no equivalent)
Level 1	Identify the main points and specific detail in texts.	Age 11	Key Skills Level 1
Level 2	Read and understand a range of texts of varying complexity, accurately and independently.	Age 16	Key Skills Level 2

In the DfES study, around 66 percent of adults were found to be at Level 1 or below, i.e. lower than what could be expected of an 11-year-old child. Reading ability at Level 1 is equivalent to being able to read and understand straightforward texts such as simple health and safety information, posters and leaflets. These learners are able to write only personal information or other information on applications forms with reasonable accuracy. These learners are however not able to understand more complex texts, or obtain information from detailed sources. In writing they may have trouble with efficient expression and adjusting to content (Williams, 2003).

Adults at Level 2 can read and understand most health and safety information, working instructions and quality guidelines. These adults can also complete accident report forms, and write letters, memos and reports accurately.

In the previous table I have briefly shown how the levels used in the Adult Core Curriculum correspond to vocational levels (Key Skills). A more thorough mapping between national standards and other definitions of levels can be seen in the figure below:



		Key Skills Level 5	National qualifications framework Level 5
		Key Skills Level 4	National qualifications framework Level 4
		Key Skills Level 3	National qualifications framework Level 3
	Literacy/Numeracy Level 2	Key Skills Level 2	National qualifications framework Level 2
National Curriculum Level 5	Literacy/Numeracy Level 1	Key Skills Level 1	National qualifications framework Level 1
National Curriculum Level 4			
National Curriculum Level 3	Literacy/Numeracy Entry 3		Entry Level
National Curriculum Level 2	Literacy/Numeracy Entry 2		
National Curriculum Level 1	Literacy/Numeracy Entry 1		

Figure 2.1: Literacy levels

2.2.3 Modeling learners

The idea of distinct groups of adult learners was inspired by the Adult Reading Components Study (ARCS) study conducted by NIFL (National Institute for Literacy, 2004). ARCS found eleven distinct reading profiles for 569 adult learners. Each profile showed a different pattern of strengths and weaknesses on reading sub-skills. ARCS used learner demographics, such as native language. Standard test such as PPVT Standard Score, and the WAIS Information Subtest Standard Score were also used. An example profile from ARCS reads as follows:

"Readers in Profile 3 need strengthening in word recognition. Underdeveloped phonological skill is affecting their word recognition ability (low Spelling GE and TAAS). They also need vocabulary enrichment that will enable them to read text at a higher level. They may be able to pass the GED reading/writing exam with their present skills but they will do so at a low level which will limit their opportunities."

It is perhaps a surpise to find consistent profiles as a British study indicates that adult learning differ from children's in that adult literacy profiles often show inconsistent performance across areas (Besser, July 2004).This effect is strongest in very poor learners. So called *spiky* profiles seem to indicate that each adult learner develops differently with strengths in some areas and weaknesses in others. Hence, among adult learners few individuals have identical profiles, while conversely, children who are the same age and have the same test scores often have similar sets of skills and abilities.

2.3 Motivation

2.3.1 Overview

Real life skills such as numeracy can be quite different from the material taught in a purely educational context, e.g. maths (J Swain, 2004), (Williams, 2003). It can be imagined that this, to a certain extent, can be extended to reading and literacy. A learner may therefore be interested in being able to perform well at a given occupation (such as the learners in the interviews, see Section 4), but not be interested in the generic skill, or vice versa. This proved to be true in the case of numeracy and maths. Some learners seemed to enjoy learning maths for it's own sake, or learned in order to improve their self-esteem. Four main motivators have been suggested for adults (J Swain, 2004):

- to prove that they have the ability to succeed in a subject which they see as being a signifier of intelligence;
- to help their children;
- for understanding, engagement and enjoyment;
- to get a qualification for a particular course or improve career prospects

Furthermore, it can be argued that this is the case with literacy as well, a learner may indeed see learning as a goal in itself, as well as a means to an end.

2.3.2 Motivation and goal theory

At this point it might make sense to differentiate intrinsic from extrinsic motivation. An *intrinsically* motivated learner finds interest and satisfaction in what they learn and in the learning process itself, while *extrinsic* motivation describes the behavior of learners who engage in learning because it is a means to an end that has little to do with the content of what is learned (Harlen and Crick, 2003).

An overview of goal theory compares learning goals with achievement goals, although only for children (Arias, 2004). Learners with goals oriented toward learning use more intrinsic motivational strategies while those with goals oriented toward achievement deploy more extrinsic motivational strategies.

The overview argues that though achievement goals are considered less adaptive, and might not lead to the same depth of information processing as learning goals, they are not necessarily maladaptive. Other findings regarding two cohorts of 9 and 13 year-olds respectively, suggest that enjoyment as well as the perception of being good at maths or science correlate positively with higher achievement (S. Thomson and Ainley, 2003). The correlation was even stronger for the older cohort. That is, learning goals, i.e. enjoying studying may correlate with higher achievement.

However, according to trials with learners in the final grades of secondary education, the best results are obtained by learners who in the final phase of execution switch to achievement goals (Arias, 2004). These learners start using achievement goals more than previous grades where learning goals are prevalent. This may be a sign of the learners ability to adapt to the requirements of the educational system, and a result of self-regulated learning. Other findings suggest that the effect of experience on self-regulated learning is that course attendance slightly increases performance, a finding additionally back up by attendees own assessments (Williams, 2003). The summary of previous research in the first mentioned overview suggests that self-regulation is more likely to follow learning goals. To conclude, achievement goals are more likely to be efficient after the use of intrinsic motivational strategies and learning goals.

Therefore it seems that a recommended strategy, if the goal theory is extensible to adults, would be to start with learning goals unless the learner has previous experience and is “ready” to assimilate achievement goals.

2.3.3 Fluid intelligence

Another important factor in motivation is the learner’s view of intelligence. Learners who see intelligence as something fixed and differentiated from effort are more likely to take on achievement goals. Those who consider intelligence as a fluid trait, and modifiable by effort are also more likely to aspire towards learning goals (Arias, 2004). Similarly, learners who attribute success to effort, and who perceive ability to be fluid and controllable are likely to deal with failure constructively, and to persevere with the learning task (Harlen and Crick, 2003).

2.3.4 Social Marketing

Social marketing is the sub field of marketing that looks at using marketing-like techniques to promote social goals such as improving literacy. This area contributes some ideas such This area contributes some ideas such as appealing to “core values”, or in this case intrinsic motivation, rather than pragmatic motivations like getting a job. Another suggestion taken from this field is to target an specific and reasonably sized audience (Kotler, 2002). These theories may help us overcome problems such as limited acceptance of limited basic skills.

2.4 Knowledge Acquisition

Knowledge Acquisition (KA), is the sub field of AI that is concerned with acquiring the information necessary to build AI systems. Usually in software engineering, and especially in expert systems, much effort is put into acquiring *domain knowledge*.

A software engineer must be aware of the context in which the software is meant to function in order to define requirements, which may be functional as well as organizational (Sommerville, 2000). Similarly in NLG systems, specifically in text planning, *communication knowledge* is considered central. There is a need for general and domain-independent knowledge about how to use language in order to achieve communicative goals.

Together these lead to a third type of knowledge requirement, *domain communication knowledge* (Rambow, June 1990), which relates domain knowledge to all aspects of communication, including communicative goals and functions. Domain communication knowledge is necessarily domain-dependent, but is not the same as domain knowledge. It is not needed to *describe* the domain, only to *communicate* about it.

In our case, the domains are adult literacy and motivation. Domain communication knowledge regards the question of how to communicate the results of a literacy test to an adult learner. This requires knowledge about what would encourage a learner to take up studies etc. However, these learners do not require explicit information about adult education, only the motivational text. In our case domain communication knowledge especially regards how to present this information in a highly readable form, while remaining factual.

2.4.1 Gricean maxims

It has been suggested that in natural language the speaker communicates more information than seems to be present in the communicated words (Jurafsky and Martin, 2000). Four very general maxims, or heuristics, have been suggested to play a key role in such communication:

- **Maxim of Quantity -** Give *exactly* the amount of information that is required; Make your contribution as informative as is required (for the current purposes of the exchange), but not more.

- **Maxim of Quality -** Try to make your contribution one that is true; Do not say what you believe to be false or for which you lack adequate evidence.

- **Maxim of Relevance -** Be relevant.

- **Maxim of Manner -** Write clearly; avoid obscurity of expression, ambiguity, be brief (avoid being wordy), and be orderly.

These heuristics have been suggested to apply in practical applications (Sripada et al., August 2003).

2.4.2 KA methods

Knowledge acquisition can be roughly divided into qualitative and quantitative methods.

Qualitative methods

Qualitative methods, also called structured KA, or expert-oriented techniques, are based on expert knowledge. Example of methods include think aloud protocols and structured interviews. In the case of **STOP** (Lennox, 2000), a system for generating computer-tailored smoking cessation letters, KA included "think aloud methods" while experts wrote example letters, sorted questionnaires, and commented on initial intervention letters (this was an iterative process). Group discussions allowed for resolving disagreements and inconsistencies between experts.

Quantitative methods

Quantitative methods refer to a family of techniques based on learning from data sets of correct solutions. A frequently used quantitative method is corpus analysis. Simply put, corpus analysis refers to statistical processing of natural language based on collections of text and speech (corpora). The knowledge gained can vary from grammar rules to discourse models.**SumTime-Mousam** is an NLG system that uses corpus analysis. SumTime generates marine weather forecasts for offshore oil rigs from numerical simulation data. The corpora consist of human-written forecasts (Sripada et al., August 2003).

Comparing methods

Methods, or families of methods, are seldom used in isolation, as each has its own strengths and weakness. Such was the case with the SumTime system for example, where qualitative and quantitative methods were used to balance out the other's weaknesses.

The weaknesses of qualitative methods are coverage and variability. The resulting system may be insufficiently sensitive to unusual or atypical data. This is largely a problem when the system is formed by only a few experts. Experts also often introspect knowledge. That is, the experts become so familiar with the domain that they fail to properly communicate some of it's requirements. Different experts may also be specialized in different areas, and their knowledge may change over time. Experts are seldom able to translate expertise into algorithmic form, which is necessary unless the problem is procedural (rather than algorithmic) in nature. However this type of knowledge can be very valuable for expanding, refining and improving existing prototypes.

Qualitative methods on the other hand are very dependant on correct, consistent data or corpora. Selection of representative data is central in many A. I. and machine learning techniques. Some areas completely lack corpora, while others suffer from individual variations between writers. For many popular corpora author information is not annotated which makes it difficult to resolve such inconsistencies. A choice appropriate for humans may also not be appropriate for NLG systems; sometimes human create sub-optimal texts that are shorter and quicker to read or write (Reiter, Sripada, and Robertson., 2003).

Chapter 3

Cluster Analysis

3.1 Aim

The aim of this project was to model different types or groups of users of the SkillSum system (See Section 1.1). My initial hypothesis was that such groups of users would have unique characteristics, or could be somehow differentiated. It was hoped that these profiles would be similar, but simpler than the NIFL/ARCS study (National Institute for Literacy, 2004) described in Section 2.2. I hoped to find that learners could be divided into more concrete groups such as dyslexics and non-native English learners.

3.2 Method

SkillSum supplies us with information that isn't available in many other types of NLG systems, i.e. information about the users as a result of a literacy assessment. On the other hand, corpora for the domain, e.g. adult literacy reports, is limited and suffers from author variability. These limitations are described in Section 2.4.2. Although there are a number of adult learners enrolled in basic skills training, many adults do not acknowledge limited skills and fewer are willing to participate in studies. Therefore the SkillSum project has repeatedly suffered from lack of representative test subjects as well as representative corpora. Due to these facts, it appeared to make sense to consider statistical methods for user demographics and test results whenever they were available.

The choice of KA method, and therefore the basis for content determination, was exploratory cluster analysis of assessment data. *Clusters* can be defined as groups of individuals which, ideally, are compact and well separated from each other (Cooper and Weekes, 1983). Cluster analysis hence refers to the techniques that may be used to find these groups. Considering the limited amount of data, I was concerned that smaller numbers within each group (compared to the total number) could lead to a loss in statistical power. However, the power of cluster analysis usually compensates for smaller numbers (Darlington and Carlson, 1987).

Statistical analysis was supplemented by the interviews, and data collected in the experiment described in Chapter 4. Viewing these learners as domain experts allowed me to complement the quantitative analysis with qualitative methods. Additionally the analysis was enriched with a background in motivational theory (Section 2.3), adult education (Section 2.2) and informal conversations with tutors (Section 5.4).

The main application used was *SPSS (version 12.0)*. SPSS is a very robust and versatile application that can be used for most types of statistical analysis. It is somewhat complex, and requires a brief introduction.

SPSS data files are composed of two components, the data, and the variables. Each component was represented in their own spreadsheets called a view. The data had to be in a case-based format, e.g. a learner per row, with columns forming variables such as correct or incorrect answer to a given question.

The requirement on data format required a large amount of manual data processing, which was found to be a time consuming process. Data in the supplied database was sometimes raw and irregular (rendering use of macros impractical). It proved practical to convert files to Comma Separated Values (.csv), saving a great deal of time. This format is used by Windows Excel which can be more convenient for modifying data than SPSS. In addition SPSS easily imports .xls files.

The variable view required only minor minor alterations, such as defining the type of variable e.g. numerical. Another useful application is TextPad, a Windows text editor that recognizes regular expressions.

Two Step Cluster Analysis differs from other clustering techniques in that it can simultaneously consider both categorical and continuous variables (LEAD technologies, 2003). *Categorical*, or nominal, variables are ones without inherent ranking, in our case, for example, specified motivation. There is no numerical relationship between “Getting a better job” and “Enjoying reading”. *Continuous* variables on the other hand, do have some built-in ordering. Continuous variables can be either strictly numerical (or scale) or ordinal, i.e. with a natural ordering that isn’t necessarily numerical. Scores on the test are considered numerical, while different literacy levels are considered ordinal, though both are continuous variables (Darlington and Carlson, 1987).

Another notable facet of the algorithm is that it automatically selects the number of clusters. **K-means Cluster Analysis** in contrast requires the number to be specified.

3.3 Results

3.3.1 Full test

The first data set to be analyzed contained the results from the full 90 question screener. It was this data that required the most analysis, although it did not render the most substantial results. One limitation of the analysis was that there were no learner demographics (e.g. dyslexia, problems with short term memory, brain injury etc.) , limiting the analysis to the purely continuous variables of score and duration. Results were grouped into the nine modules suggested by CTAD (see Section 2.1.1). It turned out that only one learner had data for word recognition, so this module was removed. Aside from the question of formatting data, some of the data was from test users, not real learners. As a result, filtering of data was also required. Variables used in the analysis were total score, and scores on groups of questions and duration for the entire assessment. K-means Cluster Analysis was used for cluster sizes 4 and 6. The results are displayed in Table “Full test results”.

The largest cluster contained 27 out of 37 learners, over 70%. Learners in this cluster scored the best, with overall non-zero scores. Areas of weakness were skimming and scanning, and punctuation and capitals. The next two clusters in size contained at most 4 learners. Both had problems with skimming and scanning, listening, and form filling. While cluster 2 had more problems with spelling, cluster 3 had more problems with punctuation and capitals.

Also, the lower the average score for the cluster, the higher the variation in both duration and score. The weaknesses for the remaining clusters varied depending on the number of clusters specified. When the number of clusters was increased to six, the three last learners were dispersed so that the last three clusters contained a learner each.

Table 3.1: Full test results

Cluster	**Weak areas**	**N**
1	Skimming and scanning, Punctuation and capitals	27
2	Skimming and scanning, Listening, Form filling, **Spelling**	4
3	Skimming and scanning, Listening, Form filling, **Punctuation and capitals**	3
4-6	Vary depending on number of clusters	1-3 per cluster

This analysis suggests that:

1. Skimming and scanning is an area that is difficult for most adult learners.
2. Profiles for poor learners are spiky, i.e. inconsistent performance across areas, a view supported by research on adult learning, see Section 2.2.

3.3.2 Screener

The time taken for analysis of the screener did not suffer from formatting to the same extent as the full test. A problem that arose with this test was the unrepresentative distribution of question types and the limited number of questions compared to the modular layout of the full test. A rough division of question types was however defined.

Additionally, some learners had multiple results registered. The test with the earliest date was chosen whenever possible, but if several tests were taken on the same day the selection was inevitably random. After combining and filtering the data sets for duplicates and test users, the full set contained around 549 learners.

Also, older screener data did not contain data for the first three questions. The scores of later tests subsequently were hence stripped to the same 24 question set (see Section 2.1.1), and duration was measured as a total time for the screener in minutes. A total score is therefore out of a maximum of 24 and not 27.

The variables studied were "total" score, scores on groups of questions, scores on individual questions, duration for the entire assessment, and "levels" - as defined below. This analysis used Two Step Cluster Analysis, to see how many clusters were formed.

Table 3.2: Screener scores

Cluster	**N**	**%**	**Duration (StD)**	**Total Score (StD)**
1	203	37	9.68 (4.005)	21.78 (0.859)
2	317	57.7	10.50 (4.191)	17.51 (2.234)
3	29	5.3	8.83 (22.738)	3.86 (4.801)

The first cluster contained learners with the lowest score, but also the shortest duration. This was a very small cluster containing around 5% of the learners. This type of relationship between score and duration is largely due to the way the assessment works. The assessment ends for learners who make 5 mistakes in a row. This cluster could consist of learners that are not interested in the assessment and click randomly, or learners are afraid of doing poorly and use self-handicapping strategies.

The second cluster contained learners with intermediate scores and duration, while the third contained learners with the highest scores and shortest duration. Most learners were in either the second or third cluster. These cluster contained 37% and 58% of the learners respectively. This would imply that learners in this sample could be roughly categorized as very good, fairly good and poor. Also, "better" learners has varied less in their scores and durations, a possible ceiling effect. The variation is especially clear for the third cluster in relation to the first two, but also noticeable between clusters one and two.

CTAD used a rough mapping from score to national literacy level, as illustrated in Table "Categories - CTAD".

Table 3.3: Categories - CTAD

Total Score	**Level**
1 - 5	"working towards Entry Level"
6 - 10	"working towards Level One"
11 - 20	"competent at Level One"
21 - 24	"competent at Level Two"

This meant that three entry levels defined by the national curriculum were given a joint name of either Entry Level or Level One.

Table 3.4: Screener levels

Cluster	**Working towards Entry level (%)**	**Working towards Level 1 (%)**	**Competenent at Level 1 (%)**	**Competent at Level 2 (%)**
1	0	0	0	100
2	0	0	99.7	0
3	100	100	.3	0

Using the mapping used by CTAD the same three clusters were found. Cluster 2 contained learners competent at Level 2, and cluster 3 learners competent at Level 1. To compare, the Skills for Life study (Williams, 2003) found 40% of adults competent on Level 1 and 44% competent on level 2 or above. This might indicate accuracy in CTADs mapping from scores to literacy levels. Consider also that the actual scores are in fact higher for newer data, as scoring for the initial three questions had been removed. However, the third cluster contained learners with scores mapped to the first three levels, working towards Entry Level, working towards Level One and competent at Level One.

Grouping questions into question types did not markedly reconfigure these clusters, suggesting simply that "better" learners do better in every type of question.

Including *scoring for individual questions*, i.e. whether correct or incorrect, in the cluster analysis again led to three clusters. These clusters were slightly different; the cluster with learners Competent at level 2 remained the same. The remaining two clusters each had increasingly longer durations and longer score. Furthermore, while the cluster with higher scores of the remaining two only contained learners Competent at Level 1, the cluster with the lowest scores contained more learners Competent at Level 1 in addition to learners at Entry level or working towards Level 1 than the other two analyses. A weakness in punctuation and spelling for the weakest cluster was also noted.

3.4 Summary and Discussion

The analyses suffered from time demanding data formatting and lack of user demographics. The full screener was the most time demanding, but reinforced the idea that poor adult learners tend to have spiky profiles. Analysis of groups of questions or individual questions did not greatly modify the composition of clusters. The standard deviations for learners with fairly good skills was higher than that for good learners.

Analysis of the screener results suggested that groups could be defined in relation to learning levels, suggesting that CTAD's mapping from score to literacy level is accurate and robust. That is learners can be categorized into: very good, fairly good and poor learners. It also seems that the target group are learners competent at intermediate to high levels, such as Competent at levels 1 and 2. These were the predominant at around 95% of the literacy levels in the samples. Levels 1 and 2 correspond to a range of total scores between 11 and 24. This would suggest a much lower percentage (54%) of people if the sample were evenly dispersed between levels. I.e. there are more learners at these levels than the score range alone would suggest. In line with suggestions from social marketing such a rough definition of audience could be useful. However despite the possible relevancy of these findings, they also strongly limited the possibility for definitions of groups beyond that of good and poor learners, in terms such as "Dyslexia". These analysis already used the data from the experiments performed and described in the next chapter. Fortunately these experiments also gave rise to other types of data as well as new ideas.

Chapter 4

Experiment

4.1 Aim

The aim of the experiments was to collect more learner data. Results from the screener test constituted the quantitative analysis, while structured interviews constituted a more qualitative component. The interviews aimed to find out what motivates adult literacy learners and to collect learner demographics. It was also hoped that observation would result in a deeper understanding of the needs of adult learners. Speaking with learners allows one to explore and to obtain a deeper understanding of the context in which the system is used (T. Greenhalgh, 2004). Special care was taken to use explicit and reproducible methods.

This experiment ran in combination with readability measurements (Williams and Reiter, 2005) of reports generated by SkillSum prior to my extensions of the system, run by research fellow Sandra Williams.

4.2 Method

Subjects Interview sessions were conducted with 60 young adult learners (age 16-25) from the University of Derby College in Buxton, England (University of Derby College, 2004). Of these 10 were male and 50 female. These learners were enrolled in varied courses such as Hairdressing, Care, Travel and Tourism, and Sports. Most users were participating in an enrichment program or Basic Skills training, and were roughly assessed as competent around Level 1.

Materials/Apparatus Desktop computers with Internet access and structured interviews (see Appendix A)

Procedure An experimental session consisted of the two components; running the screener and the structured interview.

Screeners Firstly the learners were informed about the procedure and the purpose of the research. For the first two learners the sessions was conducted simultaneously on two laptops in a quiet test room. One ran the literacy screener and the other the numeracy screener. A coin toss determined the choice of screener. This ensured random selection without imposing too much on the learners. Course tutors were supplied with information sheets.

Due to time constraints, the subsequent screener tests were run in a classroom context instead. For a number of different classes, testing sessions were conducted in the manner described below.

After the initial introduction, learners were divided into two groups. Learners in Group A were requested to run the literacy screener, and learners in Group B were requested to run the numeracy screener. Each learner ran the screener on an individual machine. Whenever possible a learner running the literacy screener had neighbors running the numeracy screener. Conversely, a learner running the numeracy screener had neighbors running the literacy screener. The main motivation for this arrangement was to increase the learner's individual effort; inter-learner discussion was discouraged. Another motivation for this arrangement was to ensure similarly sized data sets for the two types of screeners for the concurrent study.

Both screeners ended the assessment after 5 consecutive incorrect answers. In these cases the learners were told it was a fault of the system, and not related to their performance.

Structured interviews The participants completed the assessment at different times, and were instructed to either answer a series of questions from the questionnaires, or take readability measurements (Williams and Reiter, 2005). The order was based on the availability of resources.

The demographics collected included learner characteristics that could affect their literacy, such as dyslexia, reduced hearing or eyesight, brain injury, or deficiencies in short term memory. Self-assessment of literacy skills was measured on a five point scale ranging from very good to poor. *Motivation* was defined as the first choice from; 'studies', 'work', 'self-confidence', 'enjoy studying', or 'other'. This was complemented by a question about specific skills and interests. In addition, learners specified how often they did certain reading, writing and listening tasks.

4.3 Results

The analysis of test data and motivational data collected from the interviews was only conducted for literacy users, rendering a sample of 33 learners. Of these only two male learners remained. The test group was thus predominantly female, and all were native English speakers from England. Reading and writing mobile text messages (SMS) were removed from this analysis, since almost all the learners seemed to be proficient, i.e. daily usage, in these tasks. Test data was identical to the data used in the previous chapter, i.e. total score, scores on groups of questions, scores on individual questions, duration for the entire assessment, and levels as mapped from scores by CTAD.

Frequency for reading and writing scores was scored on a scale from yearly to daily and respectively summed. A higher sum suggested a more comfortable reader/writer. Assessed skill was also measured on five point scale from very good to poor. As mentioned previously motivation was defined as the primary choice from 'studies', 'work', 'self-confidence', 'enjoy studying', or 'other'.

Self-assessment The relation between learners' self-assessment and score proved to be positive and approaching significance at 0.05. Dissimilarly to previous research (S. Thomson and Ainley, 2003), this sample seemed to have a correct idea of their skill level. However, when probed about motivations, most learners said they had no motivation. They would study if they saw the need or were coerced but often thought that their skills were sufficient. These finding are similar to previous studies that suggest that adults rarely consider their skills to be insufficient, even if their actual skill level is low (Williams, 2003).

Using literacy skills Learners self-assessments coincided with how often they read. That is, learners that read more also tend to think their literacy skills are better.

Very few learners read books or wrote letters often. When asked what they read, most commonly learners claimed to read teenage magazines or newspapers such as the Sun or local advertisers. It seems therefore that the daily usage of literacy skills is very limited in these users. Although the learners claim to exchange SMSs on a daily basis, these messages are likely to be both short and limited in areas such as spelling and grammar. Hence, these learners are neither likely to improve their skills, or realize that their skills could be improved.

Carefulness In this sample a relation between score and duration which here shall be called *"carefulness"* was found. This relation differentiates different groups of users. These may be illustrated by means of exploratory regression methods. For example low carefulness means a low score in combination with short duration.

For durations between five and ten minutes learners don't seem to be taking their time and consequently receiving a lower score. However for durations between zero and five minutes, this is not exactly the case. Most likely this group reflects learners that made many initial mistakes and for which the assessment concluded prematurely. Please note that although the sample used in this study contains very few learners belonging to this group, it is likely that it represents a cluster similar to the first cluster found in 3.3.2.

It could also reflect learners that simply were not interested in the system, in this case they indeed should be considered as less "careful".

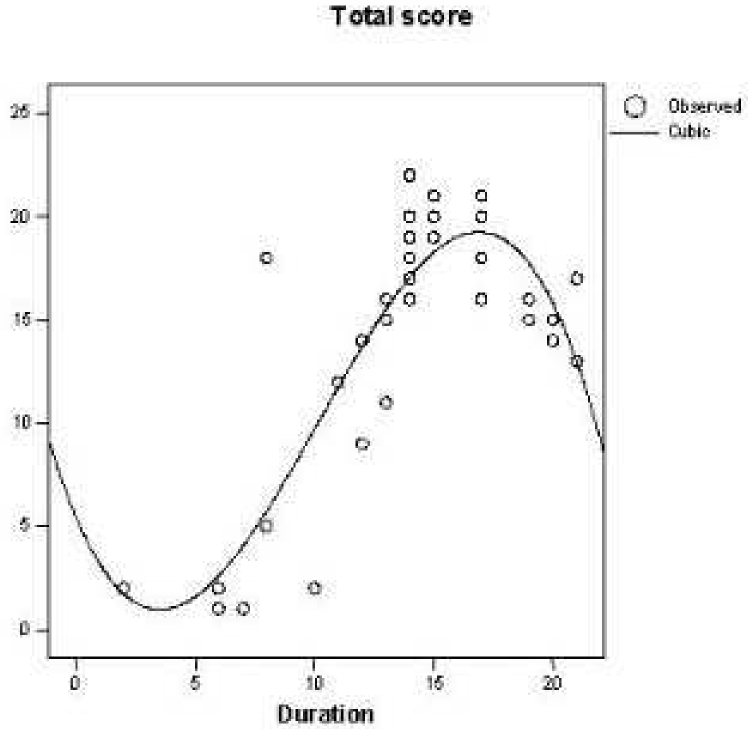


Figure 4.1: Carefulness

Likewise, between five and fifteen minutes, the slope increases with both higher scores and durations. The CTAD cut-off for learners competent at Level 1 is a score of eleven, this reflects a duration of about ten minutes in this figure. This also reflects the choice of ten minutes as duration cut-off in defining regular and low carefulness. Perhaps more surprisingly, for durations longer than fifteen minutes, the scores steadily decline. Some learners take longer though they do not necessarily obtain the highest score. These learners are more careful.

If we instead choose to study the data from the respective of cluster analysis the results are slightly less clear. Using Two Step Cluster Analysis four clusters were formed. Cluster 2 can be said to contain the learners with the highest scores and short durations and Cluster 3 with the lowest scores and least durations.

Table 4.1: Buxton

Cluster	**Duration**	**Total Score**	**Age**	**Self-assessment**
1	17.17 (3.125)	16.00 (2.280)	17.33 (1.033)	3.33 (0.516)
2	15.33 (2.693)	17.44 (3.087)	17.00 (0.866)	3.78 (0.972)
3	6.40 (2.966)	2.40 (1.517)	16.20 (0.477)	2.80 (1.095)
4	14.44 (4.003)	16.56 (3.844)	16.67 (0.500)	3.56 (0.527)

Clusters 1 and 4 both contain intermediate learners. Cluster 1 however describes learners with longer test durations and intermediates scores. Cluster 4 describes learners with slightly shorter durations and slightly higher scores than Cluster 1. These quite possibly reflect the difference between what I have chosen to call very careful and regular learners.

Cluster analysis also indicates that careful learners, i.e. Cluster 1, assess their skills lower than Clusters 2 and 4, and also hold a higher mean age.

Further exploratory analysis such as box plots reveal that these four clusters can be summarized in three categories similar to the ones found in the larger data set described in Section 3.3.2. That is, learners can be categorized into: very good, fairly good and poor learners. These levels were dubbed Level 0 , Level 1 and Level 2. Respective cut-offs were taken from the same box plots and set at scores of 6 and 20. Box plots are a way of graphically summarizing a range with the smallest observed value, lower quartile, median, upper quartile and largest observed value. A box plot may also identify outliers and possibly the mean.

4.4 Discussion

The rationale for a separate analysis for this data was that it made it possible to roughly specify the target group. For example it is highly likely that motivation could be very group specific. This also focused the analysis on a group mostly competent at Level 1, which generally seems to be the target group of the system. Limiting the sample in this manner also circumvented problems with random selection between duplicate results present in the previous samples. Though this also means that the sample is far from large enough for the results to be statistically significant, the found tendencies are well worth exploring.

Aside from quantitative insight, these experiments also paid off with qualitative insight in terms of how interviews are conducted and understanding the learners in their natural setting. Speaking to real learners highlighted the importance of motivation for example. Some learners seemed to need higher self-esteem in their abilities, and others needed more encouragement to study. The former could be influenced by low school grades, as these were often cited in the interviews.

Another side of the same coin is however that the reason that learners self-assessments by and large corresponded to their scores could be that they were already enrolled in education and had learned to assess their abilities within this framework. In this case it might be relevant to again differentiate between real life and academic skills (J Swain, 2004), (Williams, 2003). The latter type of learners could perhaps benefit from understanding that their skills were limited to a certain degree and in turn also limited their possibilities.

It is highly probable that the results were affected by a competitive environment (Harlen and Crick, 2003). Learners could not test the screener in isolation, and could not be completely restrained from discussing questions or results with each other. Unfortunately this was an inevitable result of time constraints, which followed from the experiments being conducted offsite in England. This is another byproduct of the difficulty in finding large samples of adults literacy learners.

Chapter 5

Design

5.1 Overview

The results from the interviews described in the previous chapter suggest that learner profiles could benefit from learner specific motivational data in addition to screener test data. By understanding the importance of motivation in adult learning, the aim of this project changed to the practical objective of the SkillSum project, i.e. to encourage more people to acknowledge basic skills problems and seek assistance.

One way to collect motivation data for any given learner at runtime would be to ask them via an online questionnaire tied into the screener. Due to the target audience, adult literacy users, the questionnaire had to be fairly short, and simply formulated. Lexical selection was guided by the Gricean maxims as listed in Section 2.4.1, but was not performed in a formal manner. See Section 7.1 for an example. Note that the main focus of this project remains content determination.

Another consideration was deciding what general tone an NLG report should use. For example in STOP, a major question was whether to use a positive tone, i.e. "If you stop smoking, your health will improve.", or a negative one, i.e. "If you don't stop smoking, you may get lung cancer." (Lennox, 2000). In children at least, a negative tone can decrease motivation. The effect on self-esteem of those who do not meet their own or other's expectations is often devastating (Harlen and Crick, 2003). The reports therefore held an initial positive tone. This aspect had been especially assessed in later evaluations in pilot studies and by dyslexic adult learners (See Chapter 7).

The additional three questions supplied by the questionnaire finally regarded learner experience, self assessment and explicit choice of motivation for learning. The following sections further describe each component used to build the motivational profile, explaining why it was chosen, and how it was used. The final version of the questionnaire can be found in Appendix B and may favorably be used to complement this text.

5.2 Question 1 - Experience

Rationale In the cluster analysis in the previous chapter, older students score slightly higher, even if they do not get the highest scores and may have required more time. These students also, on average, seem to assess themselves fairly poorly. Also, for students in the final grades of secondary education, the best results are obtained by learners who in the final phase of studies switch to achievement goals, but this does not seem to be as much the case for younger students (Arias, 2004). See Section 2.3.2 for a review of a basic differentiation between types of motivation and goal theory.

I suggest that while achievement goals and extrinsic motivations are not necessarily maladaptive, in the initial phases of learning intrinsic motivations and learning goals could be preferable. I have therefore decided to differentiate between learners that have studied a year or longer, and those who have studied for a shorter period. To clarify, with studies in this case I am referring to adult literacy education. The same type of clarification is found in the online questionnaire. The choice of a year is representative of the data sample, though it is not clear how it would extend to the adult literacy learner population at large.

Application Less experienced users should be redirected toward intrinsic learning and learning goals, while more experienced learners should not only be allowed, but also encouraged, to have extrinsic motivations and aim toward achievement goals. In addition, more experienced learners, especially those who are more careful when completing the screener should receive additional praise for their efforts. E.g. "Keep up the good work". Less experienced learners should be encouraged as well, but possibly in a more general manner. I believe that encouragement can be beneficiary as long as it is factual. Enjoyment as well as the perception of being good at certain tasks correlate positively with higher achievement (S. Thomson and Ainley, 2003). The type and amount of praise also depends on the learner's results, carefulness (effort), and self-assessment.

5.3 Questions 2 - Self-assessment

Rationale The results showed us that our sample in fact was capable of self-assessing their literacy skills. This supplies us with the definition of correct assessment. This definition can be extended to terms such as under and over assessment, by observing the range for correct assessments, and seeing where the learner's score lies in comparison. That is, we use the total *score* from the screener in combination with the self-assessment specified by the individual learner to decide whether an assessment is correct, or an under- or over assessment.

As mentioned in Section 2.3 on Adult learning, over assessment is a recurring event in adults (J Swain, 2004). Therefore, even if this problem is not present in this sample, it is likely to occur in another.

Application A learner that over assesses themselves should be aware that their skills are, in some ways, less advanced than they believed them to be. E.g. *"(You did alright), but your reading skills might not be as good as you thought"*. Notification should not discourage the learner, but nevertheless clarify the situation. Likewise, learners that under assess themselves should be told this as well, but without redundant or patronizing praise. E.g. *"You did alright/great. You should feel more positive about your reading skills"*. If a learner has a good idea of their skills, it might also be worth telling them that their assessment was correct, e.g. *"Keep up the good work."*.

5.4 Question 3 - Explicit motivations

Rationale Each learner has their own reasons for wanting to study, and I believed that the questionnaire should be fairly broad in this respect. When choosing what options to include, I tried to strike a balance between intrinsic and extrinsic motivations, as well as between learning and achievement goals (see Section 2.3.2 for a review of these terms). As a starting point I used previous research on adult motivation in numeracy (J Swain, 2004), see Section 2.3. These were then supplemented with additional intrinsic core values, such as feeling good or better about oneself, as referred to in the section on Social Marketing (See Section 2.3.4). Also, additions as well as alterations were offered from adult literacy tutors at CTAD both via email and a meeting called on November 18th to assess the progress of the system.

Application I posed the question in a general manner, finding out what could inspire the learner to study at all i.e. "Why would you study?", following the difficulties of retrieving such information in the interviews described in Chapter 4. Learners could select as many of the following options as they liked. This was explicitly stated in the questionnaire, i.e. "(You can tick more than one box.)"

Table 5.1: Explicit motivations

1. Because I enjoy reading and writing.	6. To read and write better.
2. To get a better job.	7. To do something new.
3. Because I've been told to.	8. To get a certificate.
4. To make friends.	9. To help my children.
5. To feel better about myself.	

Options 1,4,5,6,7 were considered intrinsic and options 2,3,8,9 extrinsic. Helping children was considered an extrinsic goal even though the definition in that case can be considered ambiguous. The justification is that the learner is aiming at a goal rather than at learning. In another context it may be justified to define this motivation in another way.

In addition, options 1 and 6 were considered specific learning goals. That is, learners that select these options already enjoy intrinsic motivational strategies, and should be complemented on this specific aspect, e.g. *"Its great that you enjoy reading and writing. This will help you learn faster.* A less experienced learner should perhaps be guided towards learning goals, as mentioned in Section 5.2, if they do not aim towards them already, e.g. "Reading more might help you to learn faster too.".

Conversely learners that select option 3 should be guided to study for their own sake, rather than allowing themselves to feel coerced, e.g. *"It could be more fun to learn if you did it for yourself."*.

5.5 Data

The answers to the questionnaire were used as a supplement to the results of score and duration from the screener. An example of how both were used together is deciding if a learners self-assessment is correct or an over- or underestimate (see Section 5.3).

5.5.1 Level

Rationale The total score calculated was out of 27. This means I included the first three questions in calculating the total sum. The actual scores are therefore higher than the ones used in the analysis resulting in the graph depicting carefulness in Section 4.3. However, I used the lower cut-offs found in this graph, as they are more inclusive. The first three questions were also considered “non-scoring” by CTAD and are primarily aimed at increasing the learners self-confidence, and are not considered central in differentiating different levels of learners.

Application Using the cut-offs at scores of 6 and 20 defined in Section 4.3 I assigned learners into one of the three categories dubbed Level 0, Level 1 and Level 2. Deciding content further depended on their self-assessment, and how careful they were. For example learners at Level 0 that assessed themselves poorly, that is correctly, were told not to be discouraged. E.g. *“Do not give up, you could improve your skills with some more practice.”* This reflects the importance of viewing intelligence as a modifiable trait (Harlen and Crick, 2003). This is different from learners at Level 1 that assess themselves correctly, and are told *“You did well/ok”*.

5.5.2 Carefulness

Rationale I chose to define a learner that is *not careful* as one that has not spent enough time on the test (see Section 4.3). Duration is measured in minutes to complete the full screener. Such a learner can hypothetically belong to any of the three levels, but I have focused on Level 0 and Level 1. This is partly due to the fact that experimental sample contained learners mainly from these groups. Also, learners at Level 2 scored well, and might not benefit from the reminder of being more patient. As mentioned previously, learners may score poorly, but they may also be disinterested in the system. I believe this can be reflected in very short durations and very low scores. In the STOP system it proved fairly unproductive to encourage smokers who were disinterested in attempting cessation, therefore I have decided that learners seemingly disinterested in the system should not feel pressured either (Lennox, 2000).

Very careful learners on the other hand, are the learners that may take longer though their scores are not necessarily the highest, and should be encouraged.

Application In order not to force learners who seems to be disinterested in the system, it thanks them, i.e. “Thank you for taking this test!”.

A correct assessment for a learner at Level 2 would be to say that they have “good” or “very good” literacy skills. For learners at Level 1, a correct assessment might be “ok” for scores between 6-10, and “good” if the score was between 11 and 20. A score of 11 differentiated these two types of assessments as this is CTADs cut-off for learners competent at level 1.

Learners at level 0 were considered to assess themselves correctly if they said their skills were “very poor”, “poor” or “ok”. A learner at level 0 does not necessarily assess themselves correctly if they say their skills are “ok”. However, this is not recorded as an over assessment. The reason for this is that telling a very poor learner that they over assessed themselves will probably not add to motivation if they already know their skills are lacking. Instead it may discourage them from further efforts.

Learners that are considered insufficiently careful are instead told they could benefit from more patience, i.e. *“Well done, you might have done even better if you took more time.”*. Careful learners on the other hand are implicitly praised for their patience, i.e. *“Keep up the good work.”*.

Chapter 6

Implementation

6.1 Overview

The initial suggested structure was a pipeline architecture as described in Section 1.3.1, and I chose to continue within this framework. The existing system was implemented in Java, and I continued with Java 2 SDK, Version 1.4.2 03. The structure and main classes remain the same as in the suggested architecture below, though help classes have been added.

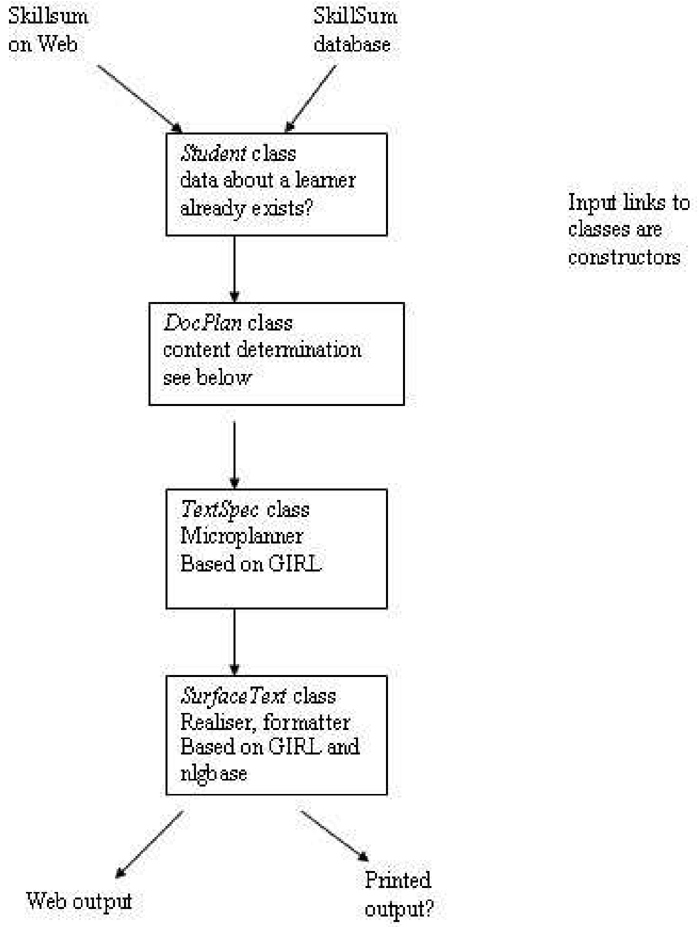

Skillsum on Web
SkillSum database
Student class data about a learner already exists?
Input links to classes are constructors
DocPlan class content determination see below
TextSpec class Microplanner Based on GIRL
SurfaceText class Realiser, formatter Based on GIRL and nlgbase
Web output
Printed output?

Figure 6.1: Suggested implementation architecture

6.2 Skillsum database

The actual database is a Microsoft ODBC database, and is accessed through JDBC. I used the previously written database handler. It consisted of methods such as **startUp()/shutDown(), putStr()/ getStr(), getActivities(),** and **checkID()**. The first two methods are used to open and close a connection to an existing database. *putStr()* writes a single String in the appropriate column in the appropriate record, user id, and table in the database. If the record exists, it updates it, if not, it inserts a new record. Similarly *getStr()* returns a String from the database if the record exists, otherwise it returns null. *getActivities()* reads from the database in a similar manner, but returns all rows from a hard coded table and column. *checkID()* checks if an id exits.

6.3 Skillsum on the web

Skillsum is a web application extending the CTADs literacy screener available at: *http://www.targetskills.net/materials/screener development/skillsum/*. After the user clicks on the "How did I do" button, input from hidden fields such as user id, type of screener (literacy or numeracy) is sent, and results (answers to questions and whether correct or incorrect) are sent to the report generating system. The generating system uses servlets run on Jakarta Tomcat 3.2.1 server. This is a rather old version of Tomcat, but a more recent version was not considered necessary for this purpose. Use of servlets and JDBC has been inspired by Deitel and Deitel (Deitel and Deitel, 1999).

6.3.1 Servlets

Servlets are Java programs that run on a web server and build web pages on the fly. In our case this is useful since data is submitted by the user first implicitly via the screener, and then explicitly via the questionnaire. I.e. the system is first required to generate the web page with the questionnaire and concludes by generating the web page for the report.

StudentFeedback

StudentFeedback.java is the Java servlet that receives the input from the hidden HTML fields. *doPost()* is the main method of this servlet. It reads the information from the HTML form and writes it into the database. It then generates a new HTML page composing the questionnaire (see Appendix A) used to acquire additional user information. The Questionnaire is created by reading a ready HTML file and appending a hidden field with the user ID. If any values are somehow missing or incorrect, an error page is generated instead.

ShowReport

ShowReport.java is also a servlet using *doPost()*. It reads the input to the questionnaire. This input together with the results already stored in the database is then used to build a user model (See 6.4). Finally this user model is used to generate the report as an HTML document. Thus the questionnaire appears after the screener and not before it. This is mostly done to simplify changes to the system, as the initial component is run by the CTADs server(s) and thus needs approval prior to each modification.

6.3.2 StudentReport

StudentReport.java is the main method. It linearly calls DocPlan, TextSpec and SurfaceText, implementing the actual pipeline.

6.4 User Model

6.4.1 Student

Student.java contains the most vital information about the student, like the id and name. *buildLitProfile()* calculates the overall score and score in different categories such as Spelling and Punctuation. *calculateDuration()* calculates the duration for a given test from database entries of start and end date and time.

6.4.2 StudentMotivation

The user model in Student was extended hierarchically with information about the user acquired from the questionnaire. This is where levels and terms such as "careful" and " over-assessment" are defined (see Section 5.3). User specified extrinsic and intrinsic motivations are stored in linked lists. Specific flags are used to specify if the learner has learning goals, is seemingly disinterested in the system or study because they have been told to.

6.5 DocPlan

The document planner, DocPlan.java consists of four components; diagnosis, summary, motivation and advice. Each component consists of a tree of discourse relations, DRTree. Discourse relations and trees are described in Section 6.6.2

6.5.1 DiagnosisDocPlan

This describes one strength, and one area of improvement. If several choices were available, the hardest area was mentioned. If the scores is less than 6, a DRTree equivalent of "You seem to be having difficulty with the questions or you are finding it hard to use the program" is built.

6.5.2 SummaryDocPlan

Currently this is described as a total, i.e. "you scored X out of 27". This is however omitted in DocPlan for the learners at the lowest Level, Level 0 (see Section 5.5.1).

6.5.3 MotivationDocPlan

This class was most central to my aims, and was the most modified of the subclasses in the document planner. It uses the information in the user model to select motivational outputs . The rationale behind the rules used is explained in Section 5.1. To recap; carefulness, self assessment, and specified motivation were considered central variables in content selection. Thus DRTrees were created for each one of the categories. For example, there was a differentiation between learners that were very careful, those that were not careful, and those that were reasonably careful. In addition there was a differentiation between disinterested learners at Level 0 and non-careful learners at Level 1.

It is important to note that in DocPlan, for disinterested users, only the motivation plan is used

For experience learners, defined by over a year at an English literacy course, up to two specified extrinsic and up to two intrinsic motivation are used. This means that the first two ticks in each category are selected. For others with less experience, one intrinsic and one learning motivation are used.

In this case learning motivation is defined by e.g. "Its great (that) you enjoy reading and writing. This will help you learn faster.". This uses the assumption that learners with less experience should focus on learning goals (see Section 5.2). A single default in each category is used where the system attempts to specify intrinsic or extrinsic motivation, but none have been specified by the learner. The default extrinsic motivation is getting a job for Level 2 learners, and a certificate for other learners. The default intrinsic motivation is learning to read and write better.

Rules

To summarize, the rules can be described in the following manner. As data analysis for content selection proved to be time demanding, little time was left for examining scientific methods for choices on the Lexical level. Gricean Maxims (see Section 2.4.1) have however been influential in this process. See Section 7.1 for a brief example.

1. Two cuts-offs at scores 6 and 21 give us three groups, or levels.
2. Set "careful", "assessment"
   - Very-careful → *"Keep up the good work"*
   - Not-careful →
     - Level 0 → *"Thank you for taking this test!"*
     - Level 1 → *"Well done. You might have done even better if you took more time."*
   - Over-assessment (not careful) → *"You did alright, but you did not do as well as you thought you would"*
   - Under-assessment (very careful) →
     - Level 0 → *"You didn't think you would do so well, but you did!"*
     - Level 1 → *+ " great"*
   - Correct-assessment →
     - Level 2 → *"You did very well/well, as you thought"*
     - Level 1 → *"You did well/alright, as you thought"*
     - Level 0 → *"Do not give up, you could improve your skills with some more practice."*
3. Motivation
   - Specific Motivations
     - If the learner studies because they are told to → *"It could be more fun to learn if you did it for yourself."*
     - Learning goals
       * Yes → *"Its great you enjoy reading and writing. This will help you learn faster."*
       * No → *"Reading more might help you to learn faster too"*
   - Experienced based
     - For learners with a year or longer experience, look more at the goals the user specified. Mention one or two intrinsic and one or two extrinsic. Defaults are used if one or the other is not specified. The default extrinsic motivation is getting a job for Level 2 learners, and a certificate for other learners. The default intrinsic motivation is learning to read and write better. → *"Studying might help you to ... ", "It could also help you to ... ".*

- For learners with less than a year of experience focus on learning and intrinsic goals (one of each)

6.5.4 AdviceDocPlan

Returns a DRTree equivalent to the string "You could contact your local college to find out about advanced English courses." for scores over 22 and "You could contact your local college to find out about English courses." for lowers scores.

6.6 TextSpec

Text specification defines the structures of paragraphs and sentences. It converts the plan, a Vector, of DRTrees into flat, ordered Vector of Vectors of paragraphs and sentences. The core of this component lies in the microplanner that is the result of GIRL (See Section 2.1.2). Although my work does not directly relate to the microplanner, it has been relevant to understand it's functionality. Sentences and paragraphs are joined by discourse different relations. Also the preexisting lexeme structure, including definition of lexical features was necessary in order to modify the system.

6.6.1 Lexicon

A Lexeme is a meaningful linguistic unit that is an item in the vocabulary of a language. Examples of Lexemes are words, phrasal and compound words and abbreviations. A lexicon is a collection of such lexemes including verbal conjunctions.

LexFeatures

In LexFeatures.java words are roughly divided into two categories; verbs and non-verbs. Verbs have features such as tense, voice, aspect, taxis, mood, polarity, and morpheme (inflected form). Non-verbs have features such as article, number, position (e.g. pre/post verbal) and person (1st, 2nd, 3rd). These features are a simplification of the deep syntactic structure (DSynt) used with the RealPro realiser .

Example: "has said": tense = past, aspect = simple (rather than continuous), taxis = perfect (rather than nil), mood = indicative, morpheme = said (set in Lexeme.java, see Section 6.6.1). Clarifying examples can be found in the RealPro manual (CoGenTex, 2000).

Lexeme

Lexeme.java contains lexical entries for all lexemes used in this application, i.e. it is the actual lexicon. A lexeme constrains contained a "head" form (e.g. say in the previous example) and lexical features as described in the previous section. If a feature list doesn't already exist for the word, one is created. A Lexeme also has a subject - argument I in RealPro, a direct object - argument II in RealPro, and an indirect object - argument III in RealPro.

Example: *"You did alright."*

"You" is the subject performing the action. The structure of a sentence always revolves around the main verb, in this case "DO". It is past tense, second person, i.e. "did". The attribute "alright" is an attribute of this verb.

Table 6.1: Lexeme

Lexeme		
head = DO		
features =	[**LexFeatures**	
	class = verb	
	tense = past	
	voice = act	
	aspect = simple	
	taxis = nil	
	mood = ind	
	polarity = nil	
	number = sg	
	person = 2nd]]	
subj=	[**Lexeme**	
	head = YOU	
	features =	[**LexFeatures**
		class = personal_pronoun
		article = no-article
		number = sg
		person = 2nd]]
attr =	[**Lexeme**	
	head = ALRIGHT	
	features =	[**LexFeatures**
		class = adverb
		number = sg
		position = sent-final]]

Embedded clauses were implemented by defining additional verbs as attributes to the main verb. *Example: I told Mary that John eats beans. "Tell" (or "told") is the main verb while "Eat" (or "eats") is part of the embedded clause "John eats beans".*

Phrases are put together through discourse relations described in the section below. Thus effort was put into tasks such as deciding to which part of a sentence (PoS) a word most commonly belonged, or whether a verb was regular or defining conjugations/morphemes for different persons and tenses. An invaluable reference for these types of task was "A Grammar of Contemporary English" (Quirk, 1988).

6.6.2 Discourse Relations

Once lexemes and their interrelation within clauses are defined, the next step is defining relationships between clauses. The microplanner specifies the seven most common relations; Concession, Conditional, Elaboration-additional, Evaluation, Example, Reasoning, Restatement.

Examples:

1. *Concession - "A is good, but B is better" or Although a is good, B is better.*
2. *Condition - "If A is good, B is even better.."*
3. *Elaboration - "A. And B".*
4. *Example - "A is good, for example it smells of strawberries".*
5. *Evaluation - "B is good (better than A)."*
6. *Reason - "A is good, because B is good. "*
7. *Restatement - "That is, A is good".*

The GIRL microplanner described in Section 2.1.2 uses a constraint problem solver that generates all legal ways of realising each input discourse relation, looking at six specific discourse level features. These results are then scored by a set of rules extracted from corpus analysis and pilot studies. Choice of cue phrase, ordering etc is therefore not preset, but varies depending on "the context".

DRTrees

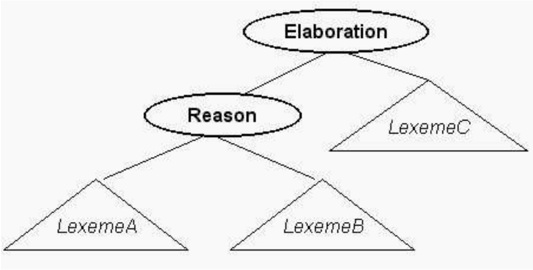


Figure 6.2: DRTree

A discourse relation tree can either be a single lexeme in the root node. This simplest form of a DRTree is a Lexeme. A DRTree can also be a Mother (Root), two Daughters (also DRTrees), and a model for combining the daughters to transform their roots into the mother's root. This model is a specification of how to build the relation, e.g. order of daughters, between-span punctuation, cue phrase, cue phrase position etc.

6.7 SurfaceText.java

The plan the microplanner returns is, as previously mentioned, a Vector of Vectors; paragraphs and sentences respectively. SurfaceText linearly realises this plan paragraph for paragraph, then sentence for sentence, inserting a full stop (.) after each sentence. This framework was in place prior to my involvement in the project. The final result is a text string with the realised expressions embedded within an HTML page. Realization at this stage is thus a translation from lexemic structure (see Section 6.6.1) to sentences. Given time this system might benefit from an alternative solution with a popular lexical dictionary or lexicon such as WordNet (Princeton, 2005) together with a realiser such as RealPro. This is however far from a trivial task, and possibly a project in its own right.

6.8 Example

The purpose of this example is to illustrate how the rules in Section 6.5.3 can be applied to generate a motivational profile, and in turn a report. This example only illustrates the components of MotivationDocPlan and StudentMotivation, i.e. where I have made the largest modifications.

6.8.1 Learner profile

Let us assume a learner, Mary Smith, has gotten a total score of *14 out of 27*. She took her time, and the total duration for the test was *20 minutes*.

Mary had a hard time with peers in school, which led to a disinterest in continuing beyond elementary school. However, she has been attending a Basic Skills course for the past two years. She thinks the course has helped her a lot, and that her literacy skills skills now are about average. She says that although she sees room for improvement, her skills aren't too bad. Mary enjoys reading books and at the moment is aiming towards a nursing certificate. She also has two children, and would like to be able to help them more with their homework.

Table 6.2: Example learner profile

Duration	**Score**	**Level**	**Assessment**	**Careful**	**Experience**	**Motivations**
20 min	14/27	1	Correct	Very	Yes	"Because I enjoy reading and writing", "To get a certificate", "To help my children"

On the first question on the questionnaire she therefore answers yes to the first two sub questions inquiring if she is attending a course where she uses literacy skills, and if the course is specifically aimed at literacy. Since she has been attending the Basic Skills course for two years, she selects "a year or more" when probed about learning experience. Her profile in thus one of an "experienced" learner.

Since she thinks her skills are average, she selects the option in the middle of the scale; "ok". The system translates this to a "correct" self-assessment, as she has scored more than eleven. Her score also means she is classified as a Level 1 learner.

Mary took longer than 15 minutes to complete the screener, she's been very careful.

6.8.2 Generating a report

As explained earlier in this chapter, the SkillSum system first generate Lexemes, and then DRTrees. These are fed through the microplanner and the realiser before they become text. However, for the sake of simplification, I will address generation of Lexemes as if they were the realised text they roughly correspond to.

First the system responds to how careful Mary has been. Since she has been very careful, the system will say something like:

Keep up the good work.

Then the system checks the relation between her self-assessment and her results. A correct assessment for a Level 1 learner eventually leads to generation of the text:

You did alright.

Next, the system uses learner specified motivations. Mary can be considered an experienced literacy learner, and it uses both the intrinsic and extrinsic motivation she selected. She enjoys reading, which is an intrinsic motivation. However, it would not make much sense to say something along the lines of "Studying more will make you enjoy reading even more". The system therefore says, "Its great you enjoy reading and writing. This will help you learn faster.", encouraging existing learning goals. This also means the system perceives that Mary has not specified any intrinsic motivation and the default intrinsic motivation "to read and write better" is used. She has however specified the extrinsic motivations of helping her children and getting a certificate, which the system repeats back to her:

Studying more could help you to read and write better. As you said, studying might also help you to teach your children, and get a certificate. It is great you enjoy reading and writing. This will help you learn faster.

6.9 Running the prototype

The prototype is interconnected with the assessment available online at:
http://www.targetskills.net/materials/screener development/skillsum/

N.B. Running the prototype requires startup of the tomcat server in Aberdeen. Please notify Ehud Reiter: ereiter@csd.abdn.ac.uk prior to testing.

In addition please note that although the server may seem to be running, it may well be the previous unmodified version of the report generator.

Chapter 7

Evaluation

7.1 Piloting

After extensive testing assuring the system was robust and stable, I ran two sets of pilot tests on two PhD students at the Computing Science department at Aberdeen University. Both were working on projects with NLG. The second tester had English as a native language, but not the first.

Modifications ranged from when and how to tell learners they assessed themselves correctly, and if it all was relevant, to what key words to use on the five point scale for self-assessment. Although the two testers did not agree on all points, these pilots resulted in the final version used when running the system with real literacy learners.

The difference of opinion seem to be due to individual differences between testers. Expert variability is a common problem in quantitative knowledge acquisition as mentioned in Section 2.4. Lacking additional evaluational methods I used the Gricean maxims (see Section 2.4.1) to assist choices. For example, "as you said", was used to differentiate between learner specified, and machine specified motivations. I.e. the phrase could be used for learner specified options, but not for machine specified options. This could be said to follow the Gricean Maxim of Quality, suggesting that a machine generated motivation with this phrase is either false or in the best case lacking adequate evidence.

7.2 Testing

7.2.1 Participants

The participants were five members of the Dyslexia Group in Aberdeen. All participants were dyslexic and active adults. The participants worked in variety of occupations, some also took courses. Their level of disability ranged from mild to severe.

7.2.2 Materials/Apparatus

The materials used were desktop computers with Internet access and the evaluation sheet in Appendix C.

7.2.3 Procedure

Appointments were made with each participant according to their schedule and availability. It was made clear that the purpose of the session was not to assess their skills, but to evaluate the report generating system. A test session started with the participant running the screener. This was followed by the online questionnaire and finally concluded with the evaluation. By “Report A” I refer to the report generated by system prior to my modifications, and by “Report B” I refer to the report generated by the modified system. Questions, especially for the online questionnaire, were read aloud whenever requested.

7.2.4 Comments

Participant 1 This participant preferred the Report A because there was less to read. This participant prefers to read as little as possible, as he suffers from visual distortion and studying print makes his eyes hurt.

An external assessment was also that Report B was inaccurate; this participant certainly does not enjoy reading and writing, contrary to what the report says.

Participant 2 The two reports differ in the third paragraph focusing on motivation:

- **Report A**
  “You did very well on grammar.
  You scored twenty out of twenty-seven.
  It could help you to do more things outside your home, if you improve your reading and writing skills.
  You could contact your local college to find out about English courses”

- **Report B**
  “You did very well on grammar.
  You scored twenty out of twenty-seven.
  Keep up the good work. And you did alright. Studying more could help you to read and write better, that, as you said. And to feel better about yourself. And it is great that you enjoy reading and writing. This will help you to learn faster.
  You could contact your local college to find out about English courses”

Participant 2 had a strong preference for Report B.

She especially appreciated the positive tone, and the praise. In addition she perceived the report as factual. However though she did say she enjoys reading and writing, it did not motivate her to read more because she does not have enough time. Report A only elicited the response that she does not feel like she could do *more* outside of home.

When asked how Report B could be improved, Participant 2 said she prefers longer sentences; the paragraph in question contained too many “ands”. She also suggested that the last two sentences be conjoined. An example of this would be: *“..it is great that you enjoy reading and writing, which will help you to learn faster.”*

Also, the participant did not understand what “as you said” referred to in “Studying more could help you to read and write better, that is, as you said.” This seemed to be due to the ordering, in combination of the use of “that is”.

The participant thought that “alright” sounded too much like slang, and suggested it be replaced with fine, well or ok. In “help you to read and write better”, the combination of “help” and “better” seemed redundant. A suggestion was to say something in line with “help your reading and writing”.

Participants 3 and 4 These evaluations were received through a shared email. Both preferred Report B over Report A; “...because it contained more information and instructions about what can be done in individual cases.”

Report A was perceived as very basic and formal while Report B was observed to have “a touch of humor attached to it which made it good to read and think about”.

Participant 5 Participant 5 preferred Report B, as it was viewed as more informative and positive. Additionally this learner noted that the both reports were presented in a font size that was reasonably easy to read. Participant 5 also noted that the text was sufficiently spaced; neither report was too condensed. He appreciated that both reports used short sentences and that the texts were "plain and straightforward".

This participant suggested that text aimed at dyslexics should inform the learner that there are possibilities to apply for extra support, such as one-on-one tutoring or extra help, and that they may be necessary in order to keep up with the rest of the class.

In his case, the suggestion of college courses is not relevant as he currently attends university courses. This seemed to reoccur with the other dyslexics in this evaluation.

7.3 Bugs fixed

During the initial test session the report generator caused the server to terminate after the online questionnaire. Despite attempts to debug at the time of the season, the reason remained unclear. The two participants left without the opportunity to evaluate their reports, or get their results. The problem lay in *calculateDuration()*. All previous times registered in the test database were 12 PM or later, hence only used double digits for the hour. Since the time was retrieved from a database string, this led to incorrect parsing. This bug was fixed, and reports were generated for the two participants and sent via email for evaluation. I was able to retrieve the reply from one, Participant 1.

7.4 Summary: Areas of improvement

This evaluation has suffered from lack of representative test subjects in the same manner as the SkillSum/GIRL project. Though the evaluation does not hold statistical significance, it gave substantial qualitative feedback from real adult learners. This has not only been rewarding in terms of evaluating the system, but also in understand the domain of adult dyslexic learners as they are a highly articulate group. It might also be worthwhile to point out that the modifications I applied inevitably are influenced by the data collection and interviews described in Chapter 4, i.e. the target group of the system is perhaps closer to that of young adults rather than specifically dyslexic users.

Four out of five participants preferred Report B. These participants seemed to appreciate the positive tone, and the way in which it was tailored to them. The general structure, spacing and fonts of the reports were also appreciated. Opinions on sentence length varied. A participant with mild dyslexia preferred longer sentences, while a participant with severe dyslexia preferred shorter. This is in line with the microplanner's adjustment to good and poor learners. However due to some questions being timed, a dyslexic learner may not have time to finish reading and make mistakes where they wouldn't have otherwise. This might consequently lower their score, and fit them a profile lower than their actual ability.

It seems that a restatement relation may use the key words "that is", which together with "as you said", and the ordering defined by the microplanner create an unclear sentence. Perhaps another relation should be used.

Another important modification of the system would be altering the definition, or rather the interpretation of intrinsic goals. In the case of Participant 1, the option "... to read and write better" was selected. Although this is both an intrinsic and learning goal, it does not necessarily imply that this learner enjoys reading and writing. This is the way the system currently works, and should be redefined, perhaps to say something along the lines of *"It is great that you want to read and write better"* rather than *"It is great that you enjoy reading and writing"*.

Chapter 8

Conclusion

8.1 Summary

Although no firm conclusions can be drawn from such a small sample, the evaluation in the previous chapter suggests that the modified report is preferable to the original. Four out of five of the participants thought the report to be improved. They were also articulate about the strengths and weaknesses of the report. Those who appreciated the report specifically preferred the positive tone as well as "simple and straightforward" use of language.

8.2 Further Work

- The changes suggested by the evaluation in Section 7.4 have not been made and should therefore perhaps be the first to be implemented.
- On the implementational side, parts of the code terminate the server upon irregular data or failure, i.e. *System.exit(1);*. It was my intention to scan the code for these terminations, but found myself limited by temporal and geographic constrains.
- As mentioned previously this system might benefit from the usage of popular lexical dictionary or lexicon such as WordNet (Princeton, 2005) together with a realiser such as RealPro (See Section 6.7).

This project and other research such as the STOP project (Lennox, 2000) suggest that it is essential to explicitly address motivational and intentional issues in order to maximize relevance of the texts generated by an NLG system. Tailoring could be more effective if more information about the users personality and background can be obtained. Educational and social marketing research show (J Swain, 2004), (Arias, 2004), (Kotler, 2002) that intrinsic motivations and core values such as enhancing self-confidence are at least as important as extrinsic motivation such as getting a better job or improving health; but existing NLG social-marketing systems nevertheless focus on extrinsic motivations. Future research might therefore aim to improve the ability of social marketing NLG systems to generate effective motivational texts.

References

Aberdeen. 2004. http://www.csd.abdn.ac.uk/research/skillsum/index.html. Electronic document retrieved December 11, 2004.

Arias, Jesús de la Fuente. 2004. Recent prespective in the study of motivation: Goal orientation theory. *Electronic Journal of Research in Educational Psychology*, 2(1):35–62.

Basic Skills Agency. 2001. Adult Literacy - Core Curriculum including Spoken Communication. Produced by Cambridge Training and Development Ltd. on behalf of the Basic Skills Agency.

Besser, S., et al. July 2004. Adult literacy learners difficulties in reading: an exploratory study. Technical report, University of Sheffield.

Carey, S., S. Low, and J. Hansbro. 1997. Adult Literacy in Britain: a survey of adults aged 16-65 in Great Britain carried out by social survey division of ONS. Survey, Government Statistical Service (U.K.).

CoGenTex. 2000. Realpro - general english grammar user manual, August.

Cooper, R. A. and A. J. Weekes. 1983. *Data, Models and Statistical Analysis*. Philip Allan.

CTAD. 2005. http://www.ctad.co.uk/. Electronic document retrieved Janurary 5, 2005.

Darlington, R. B. and P. M. Carlson. 1987. *Behavioral Statistics - Logic and Methods*. The Free Press, A Division of Macmillan, Inc.

Deitel and Deitel. 1999. *Java - How to Program*. Prentice Hall.

East, Get On North. 2000. Supporting skills for life in the north east. Electronic document retrieved Janurary 10, 2005.

Harlen, W. and R. D. Crick. 2003. Testing and motivation for learning. *Assessment in Education*, Vol. 10(No. 2).

J Swain, et. al. 2004. Beyond the daily application: Making numeracy teaching meaningful to adult learners. Technical report, University of Nottingham; Kings College, London.

Jurafsky, D. and J. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing Computational Linguistics, and Speech Recognition*. Prentice Hall.

Kotler, P., et al. 2002. *Social Marketing (second edition)*. Sage.

LEAD technologies, Inc. 2003. *SPSS Base 12.0 : User's Guide*. SPSS Inc.

Lennox, S., et al. 2000. An evaluation of computer-tailored smoking cessation letters in general practice (final report). University of Aberdeen.

Moser, C. 1999. Improving literacy and numeracy: A fresh start. the report of the working group chaired by sir clause moser. Electronic document retrieved Janurary, 2005.

National Institute for Literacy. 2004. ARCS - the Adult Reading Compontents Study. http://www.nifl.gov/readingprofiles/, Electronic document retrieved November 9, 2004.

OECD. 2000. Literacy skills for the world of tomorrow. Electronic document retrieved Janurary 10, 2005.

Princeton. 2005. Wordnet. Electronic document retrieved Janurary 5, 2005.

Quirk, R., et al. 1988. *A Grammar of Contemporary English.* Longman.

Rambow, O. June 1990. Domain communication knowledge. In *Proceedings of the Fifth International Workshop on Natural Language Generation, Pennsylvania Budapest, pp. 127-134.*

Reiter, E and R Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.

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

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

S. Thomson, J. Lokan, S. Lamb and J. Ainley. 2003. Lessons from the third international mathematics and science study - a study commissioned by the australian government department of education, science and training. Australian Council for Educational Research.

Sommerville, I. 2000. *Software Engineering, 6th edition.* Addison Wesley.

Sripada, S, E Reiter, J Hunter, and J Yu. August 2003. Generating English Summaries of Time Series Data Using the Gricean Maxims. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 187- 196.*

T. Greenhalgh, R Taylor. 2004. Papers that go beyond numbers (qualitative research).
http://bmj.bmjjournals.com/cgi/content/full/315/7110/740?ck=nck
Electronic document retrieved October 15, 2004.

University of Derby College, Buxton. 2004. http://buxton.derby.ac.uk/. Electronic document retrieved November 9, 2004.

Williams, J., et al. 2003. The skills for life survey - a national needs and impact survey of literacy, numeracy and ict skills. Technical report, Department for education and Skills.

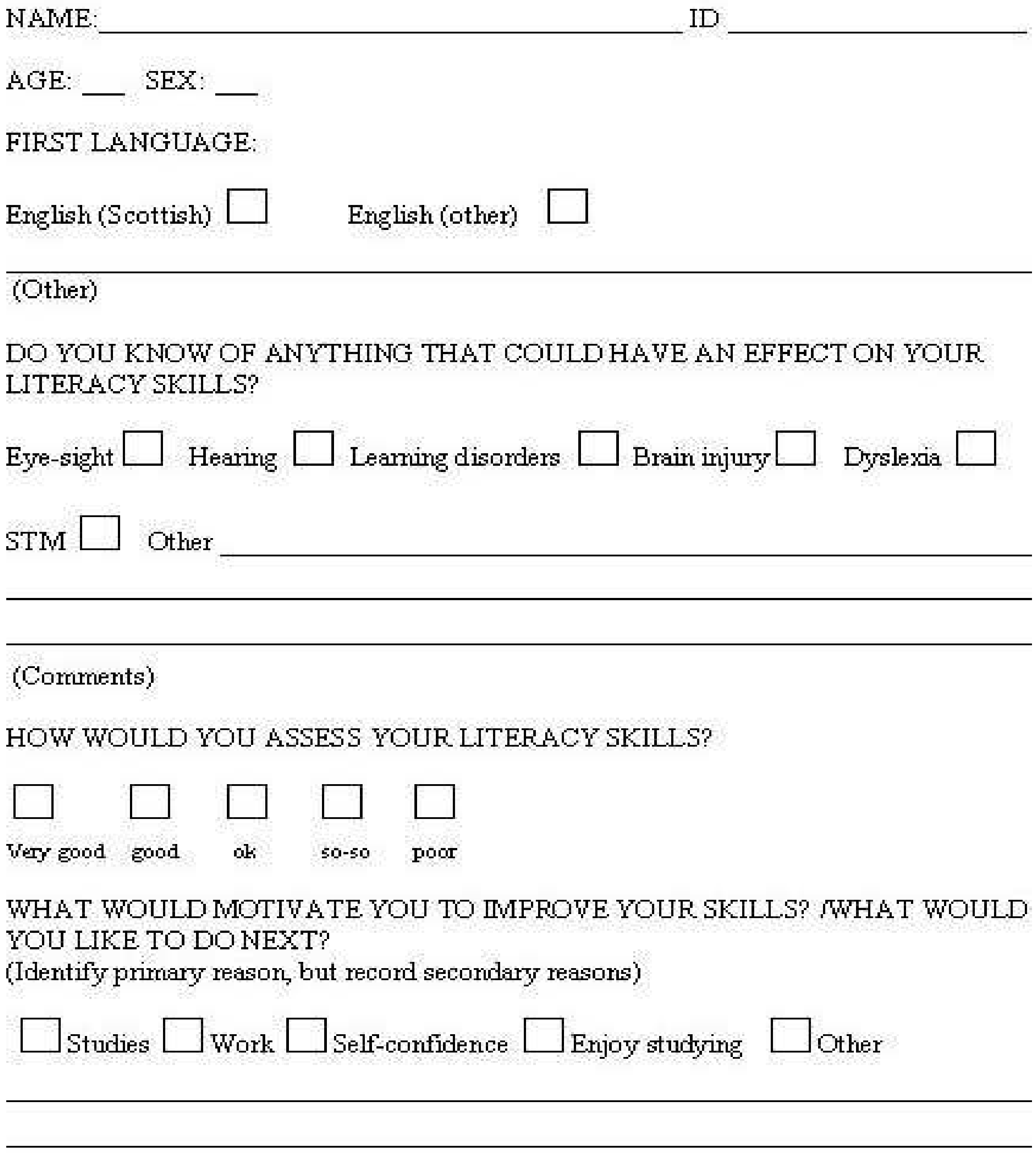
Williams, S. 2004a. Basic skills screener.
http://www.targetskills.net/materials/screener_development/skillsum/
Electronic document retrieved November 9, 2004.

Williams, S. 2004b. *Natural Language Generation (NLG) of Discourse Relations for Different Reading Levels.* Ph.D. thesis, PhD thesis, Department of Computing Science, University of Aberdeen, U.K.

Williams, S and E Reiter. 2005. Generating texts for low-skilled readers. *Submitted to IJCAI-2005.*

Appendix A

Interview structure



NAME:______________________________ ID ______________

AGE: ___ SEX: ___

FIRST LANGUAGE:

English (Scottish) □ English (other) □

____
(Other)

DO YOU KNOW OF ANYTHING THAT COULD HAVE AN EFFECT ON YOUR LITERACY SKILLS?

Eye-sight □ Hearing □ Learning disorders □ Brain injury □ Dyslexia □

STM □ Other ______________________________________

____

____
(Comments)

HOW WOULD YOU ASSESS YOUR LITERACY SKILLS?

□	□	□	□	□
Very good	good	ok	so-so	poor

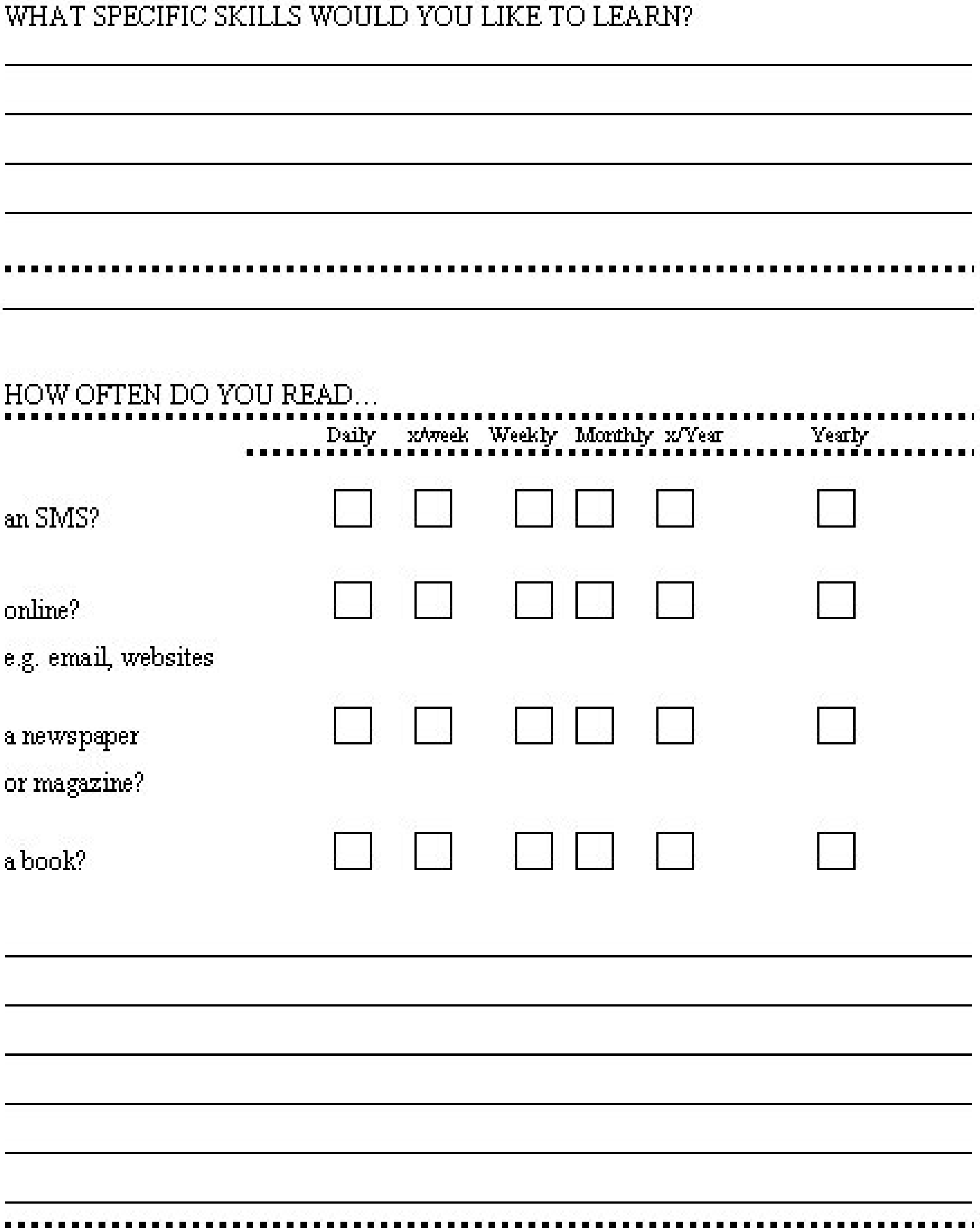
WHAT WOULD MOTIVATE YOU TO IMPROVE YOUR SKILLS? /WHAT WOULD YOU LIKE TO DO NEXT?
(Identify primary reason, but record secondary reasons)

□ Studies □ Work □ Self-confidence □ Enjoy studying □ Other

____

____

Figure A.1: Literacy Questionnaire, page 1



WHAT SPECIFIC SKILLS WOULD YOU LIKE TO LEARN?

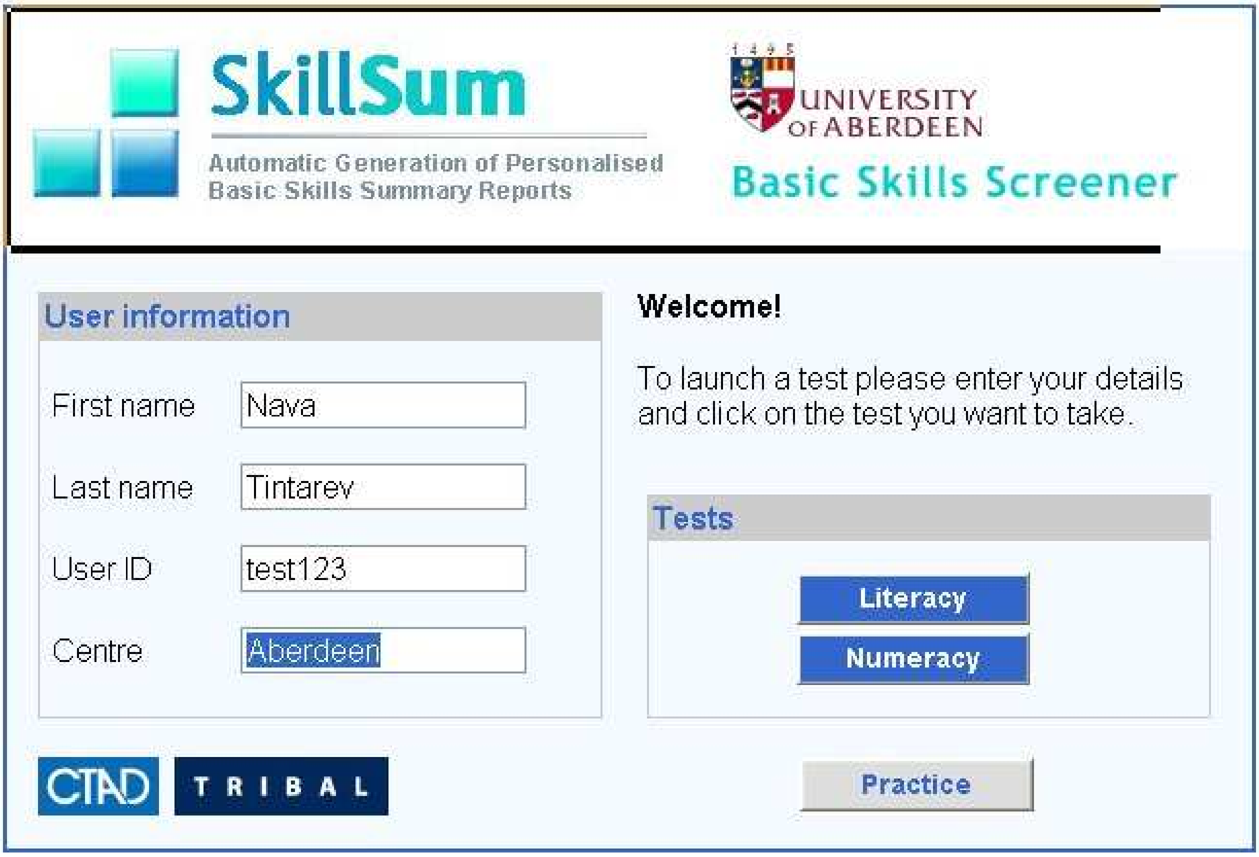
HOW OFTEN DO YOU READ…

	Daily	x/week	Weekly	Monthly	x/Year	Yearly
an SMS?	□	□	□	□	□	□
online? e.g. email, websites	□	□	□	□	□	□
a newspaper or magazine?	□	□	□	□	□	□
a book?	□	□	□	□	□	□

Figure A.2: Literacy Questionnaire, page 2

Appendix B

Screen shots



SkillSum

Automatic Generation of Personalised Basic Skills Summary Reports

UNIVERSITY OF ABERDEEN

Basic Skills Screener

User information

First name: Nava

Last name: Tintarev

User ID: test123

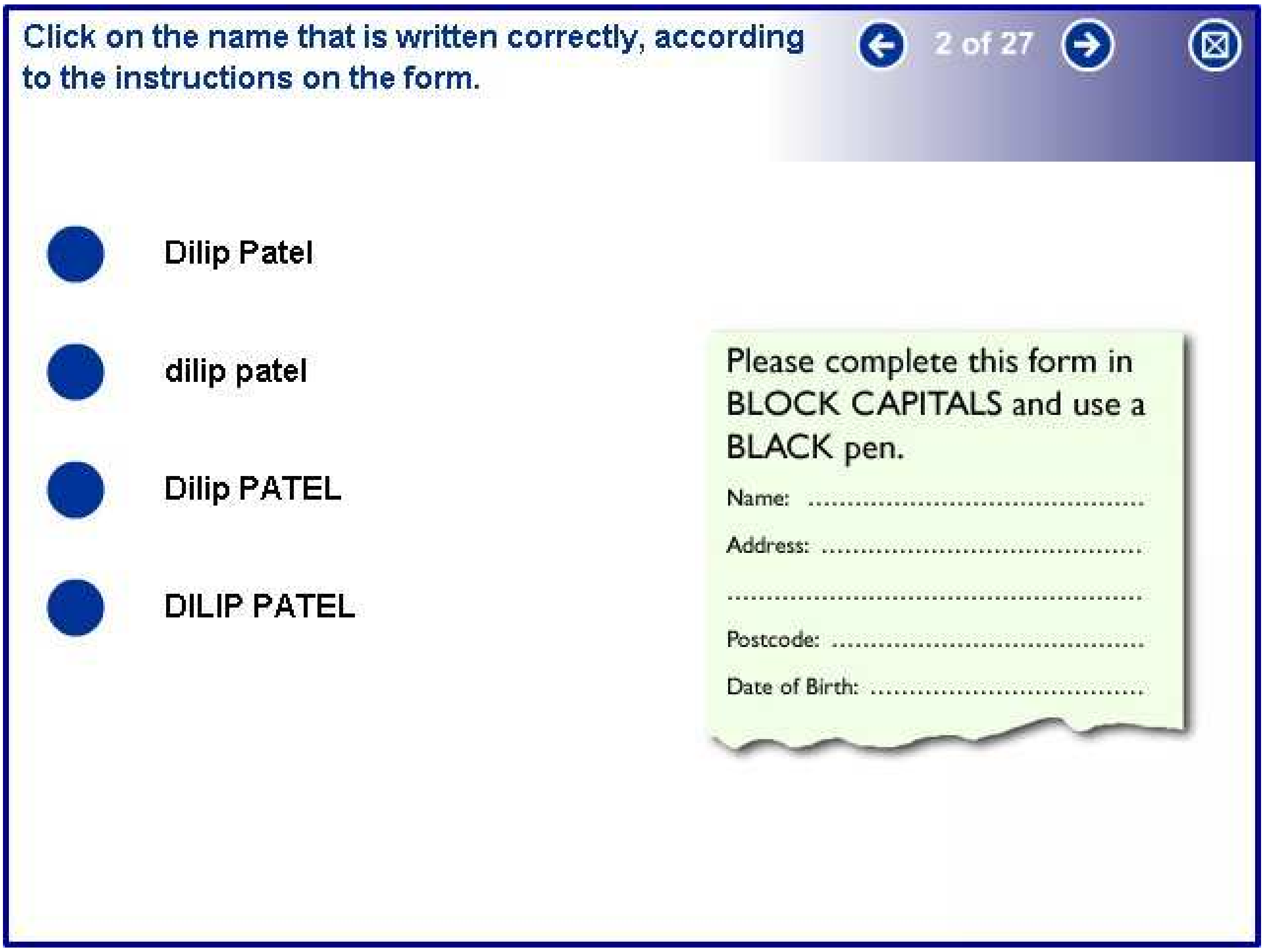
Centre: Aberdeen

Welcome!

To launch a test please enter your details and click on the test you want to take.

Tests

Figure B.1: Start screen



Click on the name that is written correctly, according to the instructions on the form.

● Dilip Patel

● dilip patel

● Dilip PATEL

● DILIP PATEL

Please complete this form in BLOCK CAPITALS and use a BLACK pen.

Name: ....

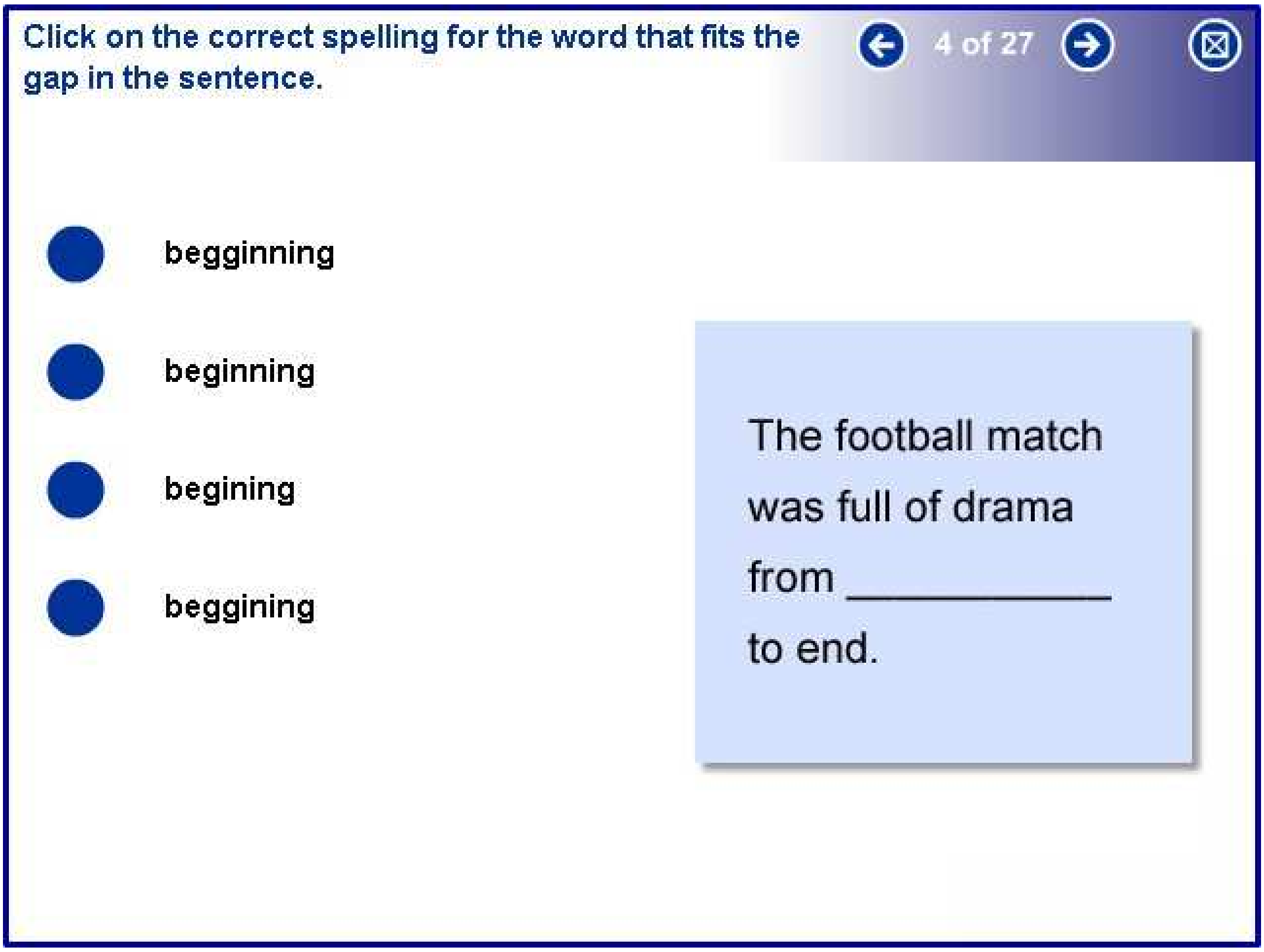
Address: ....

....

Postcode: ....

Date of Birth: ....

Figure B.2: Example question 2



Click on the correct spelling for the word that fits the gap in the sentence.

● begginning

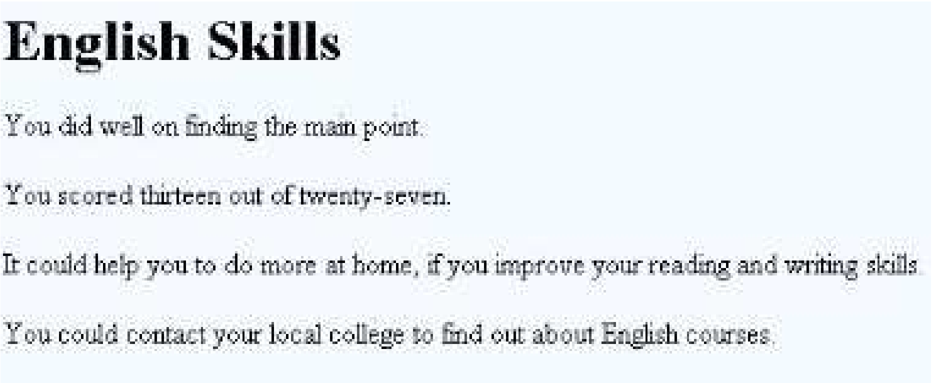
● beginning

● begining

● beggining

The football match was full of drama from ____________ to end.

Figure B.3: Example question 4


English Skills

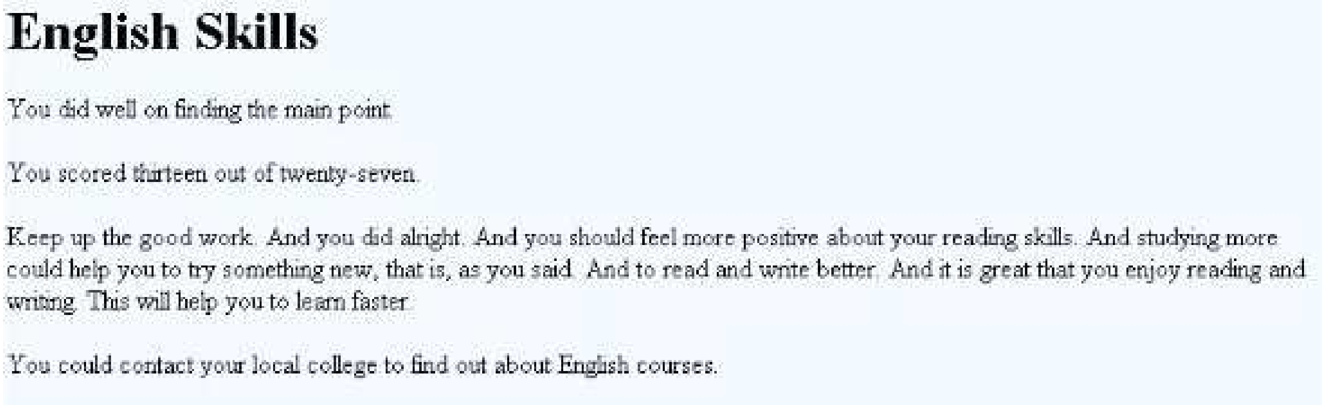
You did well on finding the main point.

You scored thirteen out of twenty-seven.

It could help you to do more at home, if you improve your reading and writing skills.

You could contact your local college to find out about English courses.

Figure B.4: Old report


English Skills

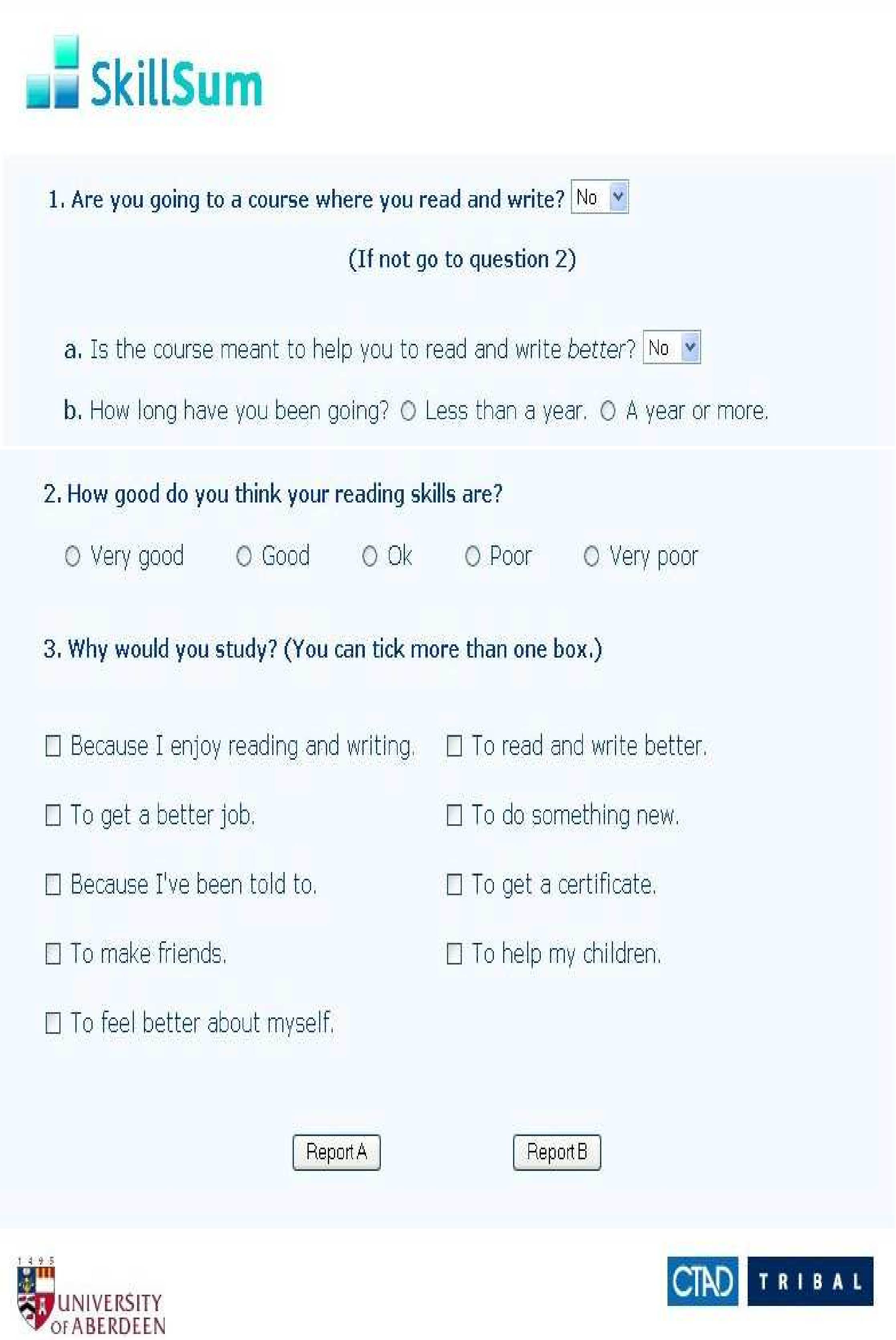
You did well on finding the main point.

You scored thirteen out of twenty-seven.

Keep up the good work. And you did alright. And you should feel more positive about your reading skills. And studying more could help you to try something new, that is, as you said. And to read and write better. And it is great that you enjoy reading and writing. This will help you to learn faster.

You could contact your local college to find out about English courses.

Figure B.5: New report



SkillSum

1. Are you going to a course where you read and write? No

(If not go to question 2)

a. Is the course meant to help you to read and write *better*? No

b. How long have you been going? ○ Less than a year. ○ A year or more.

2. How good do you think your reading skills are?

○ Very good ○ Good ○ Ok ○ Poor ○ Very poor

3. Why would you study? (You can tick more than one box.)

□ Because I enjoy reading and writing.

□ To get a better job.

□ Because I've been told to.

□ To make friends.

□ To feel better about myself.

□ To read and write better.

□ To do something new.

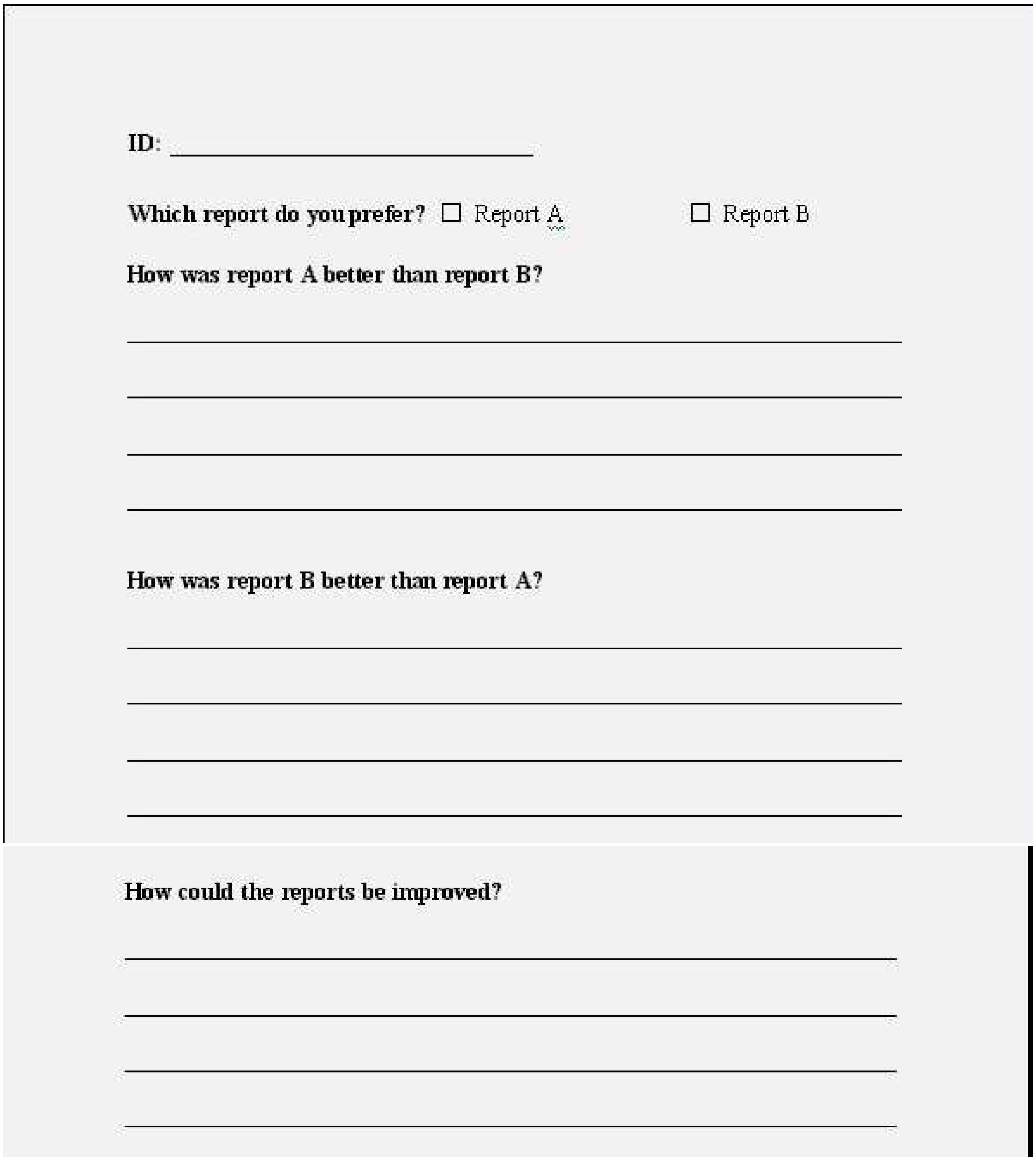
□ To get a certificate.

□ To help my children.

Figure B.6: Online questionnaire

Appendix C

Evaluation



ID: ______________________________

Which report do you prefer? □ Report A □ Report B

How was report A better than report B?

____

____

____

____

How was report B better than report A?

____

____

____

____

How could the reports be improved?

____

____

____

____

Figure C.1: Evalutation sheet