



# **A Web-based Personalised Textfinder for Language Learners**

*Jasmine Bennöhr*

Master of Science  
Artificial Intelligence  
School of Informatics  
University of Edinburgh  
2009

# **Abstract**

This thesis describes the theoretical foundations for, and the implementation of, an application designed to help language learners - or teachers on their behalf - to find texts geared to individual ability and interests. For the purpose of providing learners with texts of suitable difficulty, a new formula is developed which measures both the ability of a learner and the readability of texts. A search-engine enables the efficient supply of texts from an XML database according to a learner's query. The database was created specifically for this application by downloading texts from a diverse range of online newspapers. An experimental evaluation of the application is carried out to identify its strengths and weaknesses with the hope of eliciting improvements in future work in this field.

# Acknowledgements

It has been a year of hard work and I would not have made it this far if it was not for the support of various people:

I want to thank my supervisor Helen Pain for being open to and foster my idea of a text-finder. Further I thank her for her invaluable feedback and support throughout the course of the project. I would also like to express my thanks to my co-supervisor Miles Osborne in particular for his guidance on questions concerning text retrieval.

Many thanks to the teacher Thomas Rau from the Graf-Rasso-Gymnasium, Bavaria, Germany for providing the backbone of the evaluation by planning and carrying out lessons at short notice. I am thankful to his pupils from grade 9d, too, for providing texts for the evaluation, using the text-finder and filling out the questionnaire.

I also want to show my boyfriend, Sönke Häsel, my gratitude and let him know how much I appreciate his suggestions for the statistical part of this dissertation and that he was always prepared to discuss my ideas. Further the many hours of his conscientious proof-reading have to be acknowledged. Finally, I consider myself very lucky that he keeps me company and that I can rely on him at all times.

I am very grateful to my grandfather, Jürgen Bennöhr, who was always ready to test new versions of the text-finder and who kindly scanned the *New Headway* texts, which the new formula is based on. At the age of 83 the natural use of new media deserves very special respect! I would like to thank my grandmother, Maria Bennöhr, for many cheerful conversations on the telephone, when I needed a break from work.

Special thanks also to my flatmates Pei-Ting Yi, Ying-Chih Liao and Ya-Ching Yang who provided me with a friendly, cheerful and considerate environment throughout the last year, which I will miss very much.

I am most grateful to my parents Yvonne and Hartmut Bennöhr who raised me with love and care. They would always support my plans and listen to what I have to say. They taught me to learn, to cope with problems, to keep going and they never lost faith in me. Only thanks to their unselfish effort over all those years and their always being there for me, could I get this education and a positive view on life.

Thanks to all those named for supporting me at dead ends and giving me encouragement when the many technical difficulties appeared insurmountable.

# Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

*(Jasmine Bennöhr)*

To my dear parents.

# Table of Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	1
1.2	Outline . . . . .	3
1.3	Target Group . . . . .	4
1.4	Short Overview of the Application . . . . .	5
1.5	Related Work . . . . .	6
1.5.1	REAP . . . . .	6
1.5.2	Uitenbogerds' Framework . . . . .	8
<b>2</b>	<b>User Modelling</b>	<b>10</b>
2.1	Potential Features . . . . .	10
2.1.1	Reading Ability . . . . .	10
2.1.2	Learners' Interests . . . . .	12
2.1.3	Topic Familiarity . . . . .	13
2.2	Explicit versus Implicit Evidence . . . . .	13
2.2.1	Explicit Information Accumulation . . . . .	13
2.2.2	Implicit Information Accumulation . . . . .	14
2.2.3	Comparison . . . . .	15
2.3	Relating User Ability to Text Difficulty . . . . .	16
2.3.1	Reading versus Writing . . . . .	17
<b>3</b>	<b>Text Readability</b>	<b>19</b>
3.1	Introduction . . . . .	19
3.2	Related Work . . . . .	20
3.2.1	Overview . . . . .	20
3.2.2	Readability Formulae . . . . .	20
3.2.3	Conclusion . . . . .	23

3.3	Applicability to Adult Language Learners . . . . .	24
<b>4</b>	<b>A New (Read)Ability Formula</b>	<b>26</b>
4.1	Selecting the Training Material . . . . .	26
4.2	Preparing the Training Set . . . . .	27
4.3	Variable Selection . . . . .	28
4.3.1	Dependent Variable . . . . .	28
4.3.2	Independent Variables . . . . .	29
4.3.3	Neglected Variables . . . . .	34
4.3.4	Further Ideas . . . . .	34
4.4	The New Formula . . . . .	35
<b>5</b>	<b>User-specific Text Retrieval</b>	<b>37</b>
5.1	Text Source - Newspaper Corpus . . . . .	37
5.2	Integrating Lucene . . . . .	38
5.2.1	Introduction to Lucene . . . . .	38
5.3	Indexing . . . . .	38
5.4	Searching . . . . .	39
<b>6</b>	<b>Updating the User Model</b>	<b>41</b>
6.1	Ability Score . . . . .	41
6.1.1	Update Rate . . . . .	41
6.1.2	Update by Rating . . . . .	42
6.2	Constant . . . . .	45
<b>7</b>	<b>Description of the Application</b>	<b>46</b>
7.1	Architecture . . . . .	46
7.2	Important Program Descriptions . . . . .	47
7.2.1	General . . . . .	47
7.2.2	Offline . . . . .	48
7.2.3	User involved . . . . .	50
7.3	Difficulties . . . . .	51
7.4	Instructions for Rerunning the Application . . . . .	51
<b>8</b>	<b>Evaluation</b>	<b>57</b>
8.1	Considerations . . . . .	57
8.2	Experimental Setup . . . . .	58

8.3	Data Analysis . . . . .	59
8.3.1	Ability Scores . . . . .	59
8.3.2	Questionnaire - Use of Application . . . . .	60
8.4	Further Work . . . . .	61
<b>9</b>	<b>Conclusion</b>	<b>65</b>
9.1	Summary . . . . .	65
9.2	Strengths . . . . .	65
9.3	Weaknesses and Further Work . . . . .	66
	<b>Bibliography</b>	<b>67</b>

# List of Figures

4.1	Sentence Length . . . . .	31
4.2	Word Length . . . . .	32
4.3	Conjunction easy-diff . . . . .	33
4.4	The new formula . . . . .	35
4.5	Actual versus Predicted (Read)Ability . . . . .	36
6.1	Summary of ratings and corresponding updating formulae depending on the relation between the user's ability score and the difficulty score of the rated text. The letters $a$ and $d$ stand for ability score and difficulty score respectively. . . . .	45
7.1	The diagram of the architecture displays dependencies between programs and files. Also the division into a part of the architecture which does not involve the user (offline) and one where the user interacts with the interface (user involved) can be observed. . . . .	53
7.2	The login page. . . . .	54
7.3	The main page. . . . .	54
7.4	An extract of the page that displays query results. . . . .	55
7.5	A selected text together with the ratings at the bottom of the page. . . . .	56
8.1	The English translation of the German questionnaire that was handed out to the students. . . . .	63
8.2	Teacher versus Formula Ranking . . . . .	64

# Chapter 1

## Introduction

### 1.1 Motivation

“Reading is an important means of increasing foreign language skill and has been shown to increase vocabulary” [Uitenbogerd, 2003, p. 1]. However, reading is evidently most satisfying when we can choose the material ourselves. Foreign language learners are often forced by standard textbooks to read texts whose contents are of no particular interest to them. Language learning contexts are often far removed from the real world; people would not usually read texts similar to those they read for the sake of language learning. This is supported by empirical research by Professor Bausch of the Ruhr-Universität Bochum in Germany. He found that senior high school students are often frustrated with foreign language education. The main reason lies in uninteresting and trivial texts that fail to provide sufficient challenge and stimulus for the students’ intellect and curiosity [Informationsdienst Wissenschaft, 1997].

Just as badly, if learners are studying in a class, usually everyone reads the same texts regardless of the students’ individual interests and differences in reading ability. “It is impractical for a single teacher to seek out unique texts matched to each student’s abilities” [Brown and Eskenazi, 2004]. For students it might prove difficult though to find suitable and interesting texts on their own. While material for learners, and authentic material in particular, used to be difficult to obtain for those not studying a language in or near the target-language country, the internet has become a valuable source of a vast range of reading material. Compared to material that is targeted at learners of a specific level of language ability, such as textbooks or graded readers<sup>1</sup>,

---

<sup>1</sup>Graded Readers (sometimes called Readers or Basal readers) are books written specifically for language learners to develop their reading ability. They are made easy to read by simplifying the vocabulary

the internet offers a wider range of topics and a large amount of freely available, up-to-date and authentic information. However, “it has become increasingly difficult for users to find information on the WWW that satisfies their individual needs since information resources on the WWW continue to grow” [Sugiyama et al., 2004, p. 675]. That applies even more to people whose ability in the language the texts are written in is limited. A general problem facing a learner looking for texts is that only upon reading it will the learner know whether or not a specific text is suitable in terms of its topic and level of difficulty. The learners might be disappointed when they frequently have to find out that their ability is insufficient for a chosen text and they have to start a new search. They may even be discouraged altogether from using the internet for the purpose of increasing their foreign language skills.

It therefore seems worthwhile to automate the process of retrieving texts by the use of a system that quickly supplies the user with texts that are tailored to his individual needs. This would enable the learner to concentrate on learning the language rather than carrying out a time-consuming and potentially frustrating search process.

In a similar manner language teachers would benefit from such a program. They could easily provide each student with texts of suitable difficulty and preferred topics if the program was provided with data from which the student’s level of ability can be inferred. So instead of challenging students too much or too little by giving all students the same text to read regardless of their ability, students can be challenged according to individual ability in order to enhance their motivation and hopefully accelerate their learning curve. In addition to that students could choose a topic, which is also likely to enhance their motivation and thus result in faster progress.

With the arrival of computers with internet connections in classrooms and households the way for such a program is paved<sup>2</sup>. It could help to more easily and efficiently select appropriate texts for a variety of learners, which is important in the face of growing time-pressure on people of working age and an increasing demand for individualised learning [Brammerts, Calvert and Kleppin, 2003].

In summary, this project introduces an application that has the aim of providing learners with motivating texts of an appropriate difficulty in order to enable faster and

---

and grammar so the learner can easily understand the story. Graded Readers are not children’s books (although some are written for teenagers and children), but in general they are books for adult language learners. Each Graded Reader is written at a specific difficulty level by using vocabulary and grammar limited to that level [Waring, 2000].

<sup>2</sup>The student survey mentioned in the *Evaluation* (cf. section 8.2), as well as a brief survey of high school websites, show that most schools in Germany are equipped with at least 20 computers with internet connection and that most pupils have access to the internet at home.

more enjoyable learning. Information is provided on the theoretical background of the application and how it retrieves texts, taking into account the specific needs and interests of the user. We also hope to make a contribution to eliciting improvements in the field of user-specific text retrieval by both evaluating the application from the user's perspective and evaluating the technical components.

## 1.2 Outline

The opening sections specify the target group and give a short overview of the application. Following that is an account and discussion of related work.

Chapters 2, 3, 4, 5 and 6 are devoted to the core of the applications' components and are supported by ideas from the relevant literature. Thus chapter 2, *User Modelling*, identifies the facets of a user that have to be taken into account in order to retrieve the most suitable texts. A discussion of ways to obtain that necessary knowledge about a user in section 2.2 *Explicit versus Implicit Evidence* leads on to a description of our solution to collecting evidence about the user's properties and a discussion of their pros and cons. Since text difficulty is central to the application a detailed review of the literature in this area is presented in chapter 3, *Text Readability*. The chapter concludes with a discussion of the specific applicability of the variables used in standard formulae to adult foreign language learners, and it will thus form the basis for designing a new formula for text difficulty.

The approach of creating the new formula by means of regression analysis is depicted in chapter 4, *A New (Read)Ability Formula*. The chapter begins with the selection of training material and its preparation. The dependent and independent variables of the formula are then listed and a rationale is provided for the use of each. Following that, all neglected variables are discussed and some further ideas are presented.

Chapter 5, *User-specific Text Retrieval*, reports on the text source and how the actual retrieval of documents is performed. It starts with the description of the creation of a diverse newspaper corpus from which documents will be retrieved. The chapter then introduces *Plucene* the information retrieval library used for the retrieval component. It provides an overview of indexing. A section about searching follows that.

Chapter 6, *Updating the User Model*, shows ways of updating the user model.

Chapter 7 presents the application itself. The architecture is described first and then all programs are presented in more detail. Some implementation decisions are justified and difficulties mentioned. Also instructions for rerunning the application are given

here.

The evaluation is the subject of chapter 8. The starting point is a description of the necessary preparation including the recruitment of subjects for the user-centred evaluation and the specification of materials for the technical evaluation respectively. An overview of the experimental setup is given and the different experiments that were carried out are described. Following the data analysis the results are reported. The chapter concludes with a summary of the results and the identification of some major problems, as well as areas of further work.

Chapter 9 summarises the work and comprises an account of strengths and weaknesses as well as suggestions for further work.

### **1.3 Target Group**

Authentic reading material directly targeted at foreign language learners is sparse and thus it is much harder for language learners to find appropriate material. Foreign language learners might therefore gain even more from the proposed application than native speakers would. The latter will often know from people in their environment whether a text from a given source suits their reading ability, or the target audience for a text may even be stated explicitly, for example in advertising. Interestingly, most methods reported in the literature for matching learners to texts are primarily designed for native speaker (mostly American) school-children (cf. chapter Text Readability). However, these methods cannot easily be applied to foreign language learners since these two groups differ with respect to what constitutes a difficult text for them. The approach proposed here aims to close this gap by developing a measure of readability specifically for foreign language learners.

We restrict our analysis to adult learners mainly because different didactic assumptions would apply to child learners and more adults than children learn autonomously. For instance for adults there is no need to consider too many pedagogical peculiarities such as the restriction of content for certain ages. Moreover, from their experience of life adults usually have a more detailed idea of what and how they want to learn (technical jargon for instance). They therefore need less guidance and less external incentives than children. From youths and adults we can usually expect some more familiarity with internet applications and computer literacy or at least the ability and willingness to read a manual.

At least a basic level of proficiency in the language is required in order to un-

derstand newspaper articles. Thus the application probably is not suitable for pure beginners.

Although the application proposed here will enable the selection of individual texts for each student, there are some implications of individualised learning that might pose problems. A teacher may find it infeasible to read all texts before he gives them to his students. In that case he cannot prepare questions for each text, which would allow a deeper processing of the text and would check the student's comprehension. Also he needs to make sure that the materials come from a trustworthy source. Another potential problem is an increasing gap between the best and worst students. A less serious problem will be that common reading-out aloud phases and discussions will have to be arranged in a different way.

In light of these potential problems such an application is currently most suitable for additional reading, project work, when classes are small, and above all for autonomous learners.

## **1.4 Short Overview of the Application**

The first step in implementing the application was to create a database of reading material by automatically downloading articles from a diverse range of online newspapers. Each of the articles is assigned a difficulty score. An efficient full text search of the database will provide the learner with articles that contain the query terms. He will be shown a choice of texts ordered by difficulty score along with their titles, newspaper source, publishing dates and snippets containing the search term. If he clicks on a title the corresponding article is extracted from the database and displayed on a separate page. The user is then asked to rate the perceived difficulty of a text after reading it.

The application can estimate what an adequate level of difficulty is because the learner will have been asked to upload a text written by himself. This text is then analysed in terms of difficulty in the same way as the newspaper articles are, so that it can be matched with texts from the archive. The application updates its estimate of the learners ability by the user's ratings, relative to his ability.

Text difficulty is calculated using a formula we developed specifically for adult second language learners. In line with existing research, we use average word length and the average number of words per sentence as indicators of difficulty, but go beyond existing work by introducing the relative number of easy and difficult conjunctions/adverbs as a variable.

## 1.5 Related Work

This section discusses systems and ideas that are similar to the text-finder on the high-level. More detailed related work can be found throughout later chapters, e.g. in section 3.2 work in the area of Text Readability is discussed.

### 1.5.1 REAP

[Brown and Eskenazi, 2004] designed REAP, a system which is similar to ours in that they want to “assign each student individualised readings” using “the large amount of authentic materials on the Web” [Brown and Eskenazi, 2004, p.1]. Whereas their approach is limited to lexical knowledge set by the curriculum, our approach incorporates several factors that represent language ability and is designed to model continuous ability values rather than being based on pre-defined ability levels. Further our approach emphasises the free choice of text topic on the part of the students. [Brown and Eskenazi, 2004] on the other hand primarily base their system on the lexical knowledge of a student.

The authors specify a passive and an active user model, both of which are represented by word histograms. The passive model comprises words seen by the student while reading texts using the system. The active model stores a histogram of the words the user has demonstrated knowledge of. Updating the passive model is done primarily by taking into account the vocabulary from each new text the user has read. The active model is updated whenever a user shows knowledge of a word in the vocabulary quiz following reading a text. Unfortunately the authors are not clear about what they define as a word.

Retrieval is performed by comparing the language model of the user to the language model of texts and ranking them according to their similarity. In addition to that the authors propose to rerank the subset of appropriate documents according to the curriculum model, i. e. the words to be learned. As alternatives they also suggest to prioritise high frequency or even low frequency words from <sup>3</sup> the curriculum or focussing on words that are included in the passive model but not in the active one. They also suggest building additional topic models for topics a teacher wants his pupils to

---

<sup>3</sup>Prioritising low frequency words is to avoid the bottleneck of being left with just infrequent words at the end of a semester for instance, for which texts of the appropriate difficulty might not easily be found. Even if texts of an appropriate difficulty can be retrieved it is unlikely that they contain many or all of the infrequent words, but there will not be enough time to read a text for each word left to be learned.

learn about and to re-rank texts according to the topic model instead of the curriculum model. This necessitates, however, that the teacher provides “a few documents known to fit the topic” [Brown and Eskenazi, 2004, p. 4].

Although it seems likely that, once the words a user knows are established, their approach could yield very accurate results, it is doubtful that they can obtain that knowledge, particularly for students who do not start to use the system at beginners level. Obtaining data for the passive model might more or less work even for advanced students as the model will grow much faster than the student’s ability. This might not be feasible for the active model, however, since vocabulary quizzes are not suited to testing all the words occurring in the texts the learner read. That a word cannot be counted as known when the student demonstrates knowledge about it only once, makes it even more difficult<sup>4</sup>. Thus the active model may give quite a poor indication of the true active vocabulary ability of the student.

That the system works well is most easily conceivable when a learner starts to learn a language with the system and uses it as his main tool. However “REAP is designed to be used as an additional resource in teacher-led classes” [Brown and Eskenazi, 2004] not the main one, so other methods for determining language ability will have to be used.

Another risk the authors do not address is that not all documents on the web are trustworthy and suitable for reading, but they do not restrict their search to selected sites. REAP, being lexically based, can process arbitrary web pages with less mistakes. It is not affected by noise surrounding texts as much as models based on variables such as sentence length, which require smooth plain text with correct punctuation. Yet the content of many web pages will hardly be appropriate for vocabulary acquisition and the word histograms may be misleading, not to mention pages that have been translated to English by machine translation or pages with dubious content. For example a web page just containing a list of words will not be desired just as little as muddled link lists to various other pages, but those might match the user model better than other web pages that are not as diverse, but contain words to be learned in a decent context. The previous critique would become obsolete if the authors incorporated grammar difficulty and document cohesiveness into their approach, which they intend to do. [Collins-Thompson and Callan, 2004] also give an overview of the REAP Project, in which they propose to create a database of at least 20 million pages for grades 1 to 8,

---

<sup>4</sup>Brown and Eskenazi suggest a word may be counted as learned when a learner has shown knowledge of it three times - clearly an arbitrary number.

with off-line annotation and indexing, rather than to search the internet directly.

“Vocabulary acquisition is the primary factor we use in matching texts to a student’s abilities.” [Brown and Eskenazi, 2004]. For that they use vocabulary defined for a curriculum and partitioned into levels. They mention that their system is designed for L2 learners as well as L1 learners. The difference would be a different list of vocabulary, which they do not even explicitly mention and a difference in partition into levels, which they (over)emphasise:

In L1 learning, we can model US grades 1 to 12, but in L2 learning, there are generally five levels that are defined by the Defense Language Institute and others [...]. This would give us a coarser grain of level definition and much more vocabulary to master per level. It would be more desirable to divide the present five levels into ‘semesters’, or half-levels, for finer vocabulary control [Brown and Eskenazi, 2004].

Unfortunately the authors are not clear about where they obtain their vocabulary lists or whether they use some universal list. This raises problems in general and for autonomous learners in particular, as there is no universal curriculum so that the vocabulary and also the partitioning of the vocabulary into levels may not coincide with what students need or want to learn or have already learned. Learners or teachers usually cannot or do not want to type in a list by themselves nor have they access to an electronic list. The definition of an exact list of vocabulary for each grade/level directs learners too much, especially considering the goal of distributing individualised reading material.

On a more positive note, [Brown and Eskenazi, 2004] and [Collins-Thompson and Callan, 2004] also mention another purpose of the REAP system, which is to assist researchers in testing instructional hypotheses, which is very desirable. For example they suggest that “the effect of 10% vocabulary stretch, which has been impractical to test in the past” [Collins-Thompson and Callan, 2004, p. 2] can be investigated using their approach.

[Collins-Thompson and Callan, 2004] plan three year-long studies with both adults and children. It will be interesting to see if and how the results support REAP.

## 1.5.2 Uitenbogerds’ Framework

[Uitenbogerd, 2003] proposes “to develop a means of automatically providing reading material based on a user’s current skill-level, where the source of the reading material is the vast quantity of text available on the Internet. This application could be

used to provide further supplementary reading of current material in the user's target language.”

She wants learners to select texts from a ranked list of web pages. As potential criteria for sorting the documents she mentions user-specific vocabulary, vocabulary complexity<sup>5</sup> and general readability. As methods of determining reading ability she suggests estimating a user's vocabulary, reading sample texts, establishing the preferred level of readability by rating several texts, or completing cloze-tests (cf. Uitenboger, 2003, p. 425).

She also suggests two improvements for her proposal:

A further enhancement of the application could be the use of relevance-feedback to fine-tune the assessment of readability. Similarly, the use of collaborative filtering techniques could allow users to benefit from other users' experience of the available texts. [Uitenboger, 2003, p. 427]

Uitenboger's approach is particularly notable for the many interesting options only some of which are mentioned here - of what to include in such an application, but probably not all of them can exist in the same system simultaneously. Moreover, she does not go into detail and does not discuss the advantages and disadvantages of these options. Neither does she mention any potential problems of her application, so that it is difficult to judge the overall merit of the proposal.

---

<sup>5</sup>In Uitenboger's framework, vocabulary complexity refers to the frequencies of words contained in a text rather than a list of specific words.

# Chapter 2

## User Modelling

### 2.1 Potential Features

In order to allow text retrieval that is tailored to the individual learner, what kind of knowledge about the user is most useful has to be established. Next we need to examine what information is available, how it can be represented in a user model and how this information can be related to the readability of texts.

There are three areas in which knowledge about the user is necessary in order to supply him with the most appropriate texts. Most importantly, his reading ability has to be ascertained, but knowledge about his interests and topic familiarity would also be helpful.

There will be many facets determining each of these three features, but only some can be measured and even fewer can be taken into account when implementing a user model. In the next three subsections we investigate the determinants of reading ability and give a very brief discussion of topic familiarity and user interests. The chapter concludes by detailing what information is retrieved for the purpose of this application, and how that is achieved.

#### 2.1.1 Reading Ability

According to [Ehlers, 2003] the process of reading is composed of several levels. The visual analysis of script is followed by an assignment of graphemes to phonemes. Other components are the recognition of words and understanding their meaning. Ehlers further mentions the syntactic and semantic analysis of phrasal entities and sentences. The final level is that of text analysis, during which a reader connects sentences

and bigger units of texts such as paragraphs to reduce the text to its essential parts.

On all of these levels a reader can be more or less advanced. We neglect the visual analysis of script here as we assume for adult language learners of English that they have already mastered all letters of the alphabet. Moreover, we are in no position to speculate whether there are any letters that are easier to recognise than others, and how this might vary among individuals<sup>1</sup>.

The assignment of graphemes to phonemes is most important when reading out aloud, but obviously there are easy and more difficult cases of matching graphemes to phonemes. Some graphemes are ambiguous, that is they are pronounced differently in different contexts, for instance the letter *k* in *knowledge* or *keen* as opposed to the letter *m*, and are thus probably easier to process when a learner has well-developed reading skills. Although [Ehlers, 2003] differentiates between the assignment of graphemes to phonemes and word recognition, she points out that there is an interaction between these two levels, the exact process of which is controversial. The example of *thought* versus *sour* clearly illustrates that the two levels often overlap. We only know how to pronounce *ou* once we have recognised the whole word in which it occurs. In the case of *read* even a syntactic analysis has to take place in order to disambiguate the grapheme *ea* to either the phoneme [i:] if present tense or [e] if past tense. It seems reasonable to assume that less advanced learners find it more difficult to perform the disambiguation and thus need longer than advanced learners.

Recognition of words can only happen when a learner has seen that word before, so this step relates to the level of his vocabulary knowledge. To understand the meaning of a word, even more depends on the vocabulary mastery but again strongly interacts with other levels, such as syntactic and semantic analysis. Here it must be said that it is not trivial to decide whether or not a word is known as learning words is a gradual process. A learner may suspect that a word has a good or a bad connotation, but not know the exact meaning. He might know the word in one context, but not in all those possibly occurring. He might even know a word in many contexts, but is not able to use it. A learner might not use a word properly, but the meaning nevertheless can be understood. In all those cases we will not light-heartedly say that the word is known. We might want to differentiate between passive and active vocabulary knowledge, but that too does not lead to a clear distinction. Even if a learner uses a word correctly, that might just be a lucky guess. In the next instance, he may use it in a different context

---

<sup>1</sup>The aspect of visual analysis is more important for languages like Chinese where there are several thousand characters that have to be known for reading an average text.

where it is not appropriate. These observations have to be taken into account whenever vocabulary knowledge resurfaces throughout this dissertation.

For syntactic analysis ambiguity again plays an important role in text comprehension. Also readers usually first master easy and then increasingly more difficult grammatical constructions. Besides certain grammatical constructions being inherently difficult we suppose that sentence constructions which do not exist in the learners' mother tongue might pose problems to them in particular.

In the case of semantic analysis, constructions that are harder to comprehend are often those that cannot be explained by the use of the principle of compositionality<sup>2</sup>, such as idioms or in some cases collocations, and are also different from idioms (and collocations) in the reader's mother tongue. Provided all previous levels have been passed, text analysis does not necessitate further language dependent ability but rather challenges the cognitive ability and general knowledge of the reader.

In summary the level of ability of the following aspects determines the reading ability of a learner. The transfer of graphemes to phonemes is probably quicker with advanced learners. Vocabulary knowledge provides the basis for the recognition of words and for understanding word meaning. Mastery of grammar and vocabulary helps with syntactic and semantic analysis. Text analysis is largely language independent, but in order to link what he reads to the wider context, a learner may need specific cultural knowledge.

### **2.1.2 Learners' Interests**

This is quite an important aspect in order to retrieve texts the learner is interested in. We reckon however, that just typing in queries is enough for the user to express his interests for the time being. Other approaches include letting the user specify categories he is interested in, or inferring topics from the texts he uploaded. The latter two approaches will need advanced methods to classify content which are beyond the scope of this dissertation.

---

<sup>2</sup>“The Principle of Compositionality is the principle that the meaning of a complex expression is determined by the meanings of its constituent expressions and the rules used to combine them” [Wikipedia, 2005].

### 2.1.3 Topic Familiarity

It would have been interesting to check if and to what extent topic familiarity can be determined, but previous approaches (cf. [Liu et al., 2004]) do not seem too promising or not convincing and thus due to the narrow time constraints we cannot pay further attention to them.

## 2.2 Explicit versus Implicit Evidence

Means of accumulating the appropriate knowledge about the user have to be considered.

In general there are two main alternative ways of obtaining that knowledge: explicitly or implicitly. Explicit information gathering would be achieved through engaging the user in activities additional to his natural interaction with a system [Kelly and Teevan, 2003]. Implicit accumulation of knowledge is to infer information about the user from his usual behaviour.

Approaches for both alternatives are described and evaluated. The section concludes with a comparison of the two methods.

### 2.2.1 Explicit Information Accumulation

There are several ways of explicitly obtaining information from the user about his level of reading skills. [Uitenboger, 2003] proposes three different methods.

One of them is to let the user “check a list of words in the target language, and mark them as known or unknown” [Uitenboger, 2003, p. 427]. Thus an exact list of known words<sup>3</sup> can be determined. This would be impracticable since in most cases the size of the user’s vocabulary would be too large to have him mark all the words he knows from a list that would have to be considerably larger than his vocabulary (cf. the approach of [Brown and Eskenazi, 2004], who want to infer such a list from the text a user has read). Alternatively, the vocabulary could be estimated on the basis of the frequencies of the words that the user marked as known. This appears to be more realistic, however it is only an approximation. What is more concerning is that vocabulary size does not necessarily mirror reading ability.

---

<sup>3</sup>Here we encounter again the problem of defining what constitutes a known word. Uitenboger’s approach is to require a learner to be able to define a word for it to be counted as known. We could of course go on to ask what constituted a good enough definition.

The second method [Uitenbogerd, 2003] proposes is to show the reader a sample text for which he shall indicate if he prefers to read an easier or a more difficult text. Just to indicate that preference for only one text would probably not be enough to provide texts of the desired difficulty. There would have to be several passes of rating a text to reliably approximate the desired level. The readability of the sample text that matches the desired level of difficulty most closely would then be used to find the most suitable of the subsequently retrieved texts.

The third method is to let the user complete a cloze test. There are also several other tests that could be used for establishing reading ability, for example the c-test which is reported to be more reliable than the cloze test (cf. [Jafarpur, 1995]). It is also conceivable to ask the users about results they obtained in official language tests such as TOEFL or IELTS, but not every learner has taken such a test.

Furthermore a learner can be asked for a self-estimation of where he is on some widely-used scale, such as the scale of the Common European Framework of Reference for Languages [Council of Europe, 2001].

Apart from rating sample texts, the mapping to text difficulty scores is not trivial. How to transfer vocabulary knowledge, scores of the c-test/cloze test and official language tests or the position on a scale to different readability levels would have to be stipulated.

Lastly, quite a different approach would be to either let the learner copy texts he has recently written or read into the application or to let him state a file or files containing texts and then to calculate the readability of those texts. This idea owes to the implicit method of [Budzik and Hammond, 2000, p.44] who found that “user interactions with everyday productivity applications provide rich contextual information that can be leveraged to support just-in-time access to task-relevant information”. He could even be asked to write a text solely for the purpose of calculating the readability but that might be considered too laborious by many learners.

### **2.2.2 Implicit Information Accumulation**

There are many ways to implicitly collect knowledge about the user. Most of the literature deals with the question of how to infer *interest* from user behaviour, where mouse clicks, saving, printing and time spent on a site are among the most common methods (cf. [Kelly and Teevan, 2003]. [Kelly and Cool, 2002] reckon however that it is also possible to infer *topic familiarity* from a user’s search behaviour. An approach for im-

implicitly gaining knowledge about the *reading ability* through user queries is described by [Liu et al., 2004]. Traditional readability measures cannot be used to identify someone's reading ability from a query, for example of the type that might be submitted to a search engine, as those formulae "become unreliable when the size of the text drops below the requirement" [Liu et al., 2004, p. 548] of 100 words.

He claims however to be able to infer the approximate grade levels of schoolchildren with a method that uses support vector machines. Again it has not been proved that this method yields reliable results for adults. In any event it is difficult to imagine that reading ability could be estimated reliably from queries as short as a single word. Moreover this approach becomes useless when a user has looked up search terms in a dictionary beforehand in order to find texts which contain words he wants to learn. By using contextual information, [Budzik and Hammond, 2000] they want to find documents that relate to the topic a user is currently working on. But in addition to that the texts a learner uses are an equally good indicator of the text difficulty that might be appropriate for that learner. To capture changes of the user's ability over time it might be possible to draw conclusions from the words a user looks up in a dictionary. Other approaches which we cannot pursue include measuring the time a user needs to read a text in order to obtain information about how fast he can process words or use eye-tracking.

### 2.2.3 Comparison

Implicitly collecting information is favoured throughout the research community (cf. [Sugiyama et al., 2004], [Dumais et. al., 2003], [Kelly and Teevan, 2003] and [Belkin et al., 2004]). For example in [Dumais et. al., 2003] implicit feedback is described as more useful than explicit feedback because users are rarely willing to provide the latter. "Implicit feedback can be encouraged by clever interface and systems design. Explicit feedback can be obtained when the incentives are high enough and input is easy enough" [Dumais et. al., 2003]. In commercial contexts the users might not have an incentive to provide feedback as they do not directly benefit from it. Another issue might be that users do not want to state personal information for reasons of safety and data protection. These last two disadvantages apply less to a personal system or restricted system as no-one unauthorized will have access to the personal information and the user should see that he directly benefits from providing the information.

On the other hand, "Implicit measures are generally thought to be less accurate

than explicit measures” [Kelly and Teevan, 2003, p. 18]. That is due to the fact that inference is not the same as direct interaction with the user and thus wrong assumptions are more likely to be made. By contrast, to explicitly ask the user takes time and might annoy the user, but the resulting insights can be very informative, even if only very few questions are posed.

Quite a different issue is that it is not possible to start looking for texts using only implicit knowledge from the beginning because a sufficient amount of information can only be accumulated through repeated interaction with the system. By contrast, just a few well-targeted explicit questions would suffice for the system to provide texts that are at least partly adjusted to the actual needs of the user.

Evidently there are advantages and disadvantages to both implicitly collecting information inferred from user behaviour and to posing questions to the user. It might therefore be a good idea to combine the two ways, for example by establishing a relatively detailed user profile in the beginning from explicit answers and then updating the user profile through collecting implicit information. It can be hypothesised that explicit information will be less informative over time, whereas implicit knowledge becomes relatively more useful due to its low cost to the user.

## **2.3 Relating User Ability to Text Difficulty**

Since the central task of this study is to design a program that will match texts and learners according to their difficulty and ability, respectively, it is obviously very important to find a reliable and accurate procedure for making that match.

As far as text difficulty is concerned, the literature, which will be surveyed in chapter 3, appears to have reached the consensus that text difficulty levels can be determined in a satisfactory manner using a readability formula of some description. We follow that consensus in that we, too, devise such a formula, which will be the subject of chapter 4. One of the factors that distinguishes this new formula from existing ones, however, is that an effort is made to incorporate factors that are specific to adult second language learning.

The debate on user modelling, by contrast, provides no consensus suggestions for a quick and easy means of obtaining the necessary information about the user’s level of ability. The approach we adopt is to use exactly the same method to measure both text difficulty and user ability. In other words, if we have a piece of text that accurately reflects the learner’s level of ability, the obvious way to proceed is to analyse it with

the same formula we use to establish text difficulty.

Most of the methods discussed in the context of the implicit versus explicit debate would pose the problem that a user's reading ability score thus obtained could not be directly linked to the difficulty of a text because the two values would be measured on different scales. We would first have to find - presumably by a series of experiments - some sort of (potentially non-linear) mapping function that helps translate ability scores into difficulty scores.

### 2.3.1 Reading versus Writing

The piece of text we want to use to establish the learner's level of ability could come from two sources: We could either analyse a text which the learner has read and found to be of an appropriate level, or we could ask him to provide a piece of text written either recently or specifically for this purpose. Both of these approaches have been briefly discussed above and would be classified as explicit information collection, which makes sense seeing that in this first interaction with the user, we want to retrieve as much information as possible, even if that requires some effort on the user's part.

We will first focus on reading. Along the lines of Uitenbogerd's proposal discussed in section 2.2.1, we could ask the learner to read a number of texts whose difficulty values have been calculated beforehand using the formula. The learner needs to state which one of the texts most closely reflects his level of ability; and the difficulty level of that text then simply becomes his ability score. As was already mentioned, a major drawback of that approach lies in that the learner may have to read quite a large number of texts before he finds one that sufficiently matches his abilities. Here the costs of explicitly collecting information may be too large.

Alternatively, the program could allow the user to upload any text(s)<sup>4</sup> he recently read and considers to be of suitable difficulty. One of the problems with uploading a text the user read, as opposed to wrote, is that he may feel quite comfortable reading texts of greatly different difficulty levels, maybe depending on his mood. That may make it difficult to pin down his ability. Also, such a text may not always be readily available in electronic format. In that case the learner would have to produce a text himself - a possibility we turn to next.

If we assume that reading and writing ability are highly correlated, a sufficiently

---

<sup>4</sup>The more texts he uploads, the more accurately his ability can be established of course.

long text written recently by the user should reveal a lot of information about the user's abilities. The user cannot write a text that is beyond his command of the target language and he is unlikely to write one that is significantly below it. So we can expect to obtain more accurate information than we would get from the reading approach.

The objection to this is of course that writing ability does not equal reading ability. In almost every case, a text written by the learner will be less difficult (by whatever measure) than one he considers to be of the right difficulty for reading. For we write with our active vocabulary whereas we read with our passive vocabulary, which is usually much larger. Similar considerations apply to other areas of language ability.

The program should obviously take this gap into account. The easiest way to do so would be to add a constant to the ability score obtained from a written text to arrive at the appropriate difficulty level<sup>5</sup>. The optimal size of this constant would have to be determined empirically, which is unfortunately beyond the scope of this project. Instead, the current version of our application allows the user to upload a text which he either read or wrote and then indiscriminately adds an arbitrary number of 10 to the ability score. Later versions would ideally ask the user to state whether he read or wrote the text, and apply the constant only where necessary.

---

<sup>5</sup>Multiplying the "written" ability score by a constant may be better than adding one. One candidate for that constant might be the typical ratio of passive to active vocabulary size.

# Chapter 3

## Text Readability

### 3.1 Introduction

One of the main aims of the text-finder is to come up with texts that are suitable for a learner, i.e. neither ask too much of him nor challenge him too little. The texts should be of a difficulty such that the reader's skills are optimally enhanced. This level of difficulty will of course differ across readers and even differ for a person at different points in time. Some clues as to how this can be achieved can be found in the literature. [Nagy, 1988] claims that readers can learn best from texts containing up to 15% of unknown words and [Brown and Eskenazi, 2004] propose to check "the effect of a 10% vocabulary stretch". [Ghadirian, 2002] writes:

A reader who is familiar with 80% of the tokens in a text, however, is still not able to adequately comprehend the text. Studies by Liu & Nation (1985) and Laufer (1989) point toward 95% as the amount of coverage required in order for a reader to adequately understand a text and guess new words from context.

<sup>1</sup> This suggests that the amount of unknown vocabulary should be somewhere in the range of 5 to 15%.<sup>2</sup>

In addition to containing some new words in order to enhance the learner's abilities, a text should be of interest to the learner and not be too difficult in terms of grammar, vocabulary and content - all of which are user-specific requirements.

Once the ability of the reader is known a text of matching difficulty has to be found. There has been much research on quantitatively determining the difficulty of a

---

<sup>1</sup>Presumably this refers to reading without a dictionary.

<sup>2</sup>Here again we have to deal with the problem of defining unknown vocabulary. Also for the range of unknown vocabulary it must be considered that certain words are easier to guess and learn than others.

text, some of which is presented in the following section.

## 3.2 Related Work

### 3.2.1 Overview

A large number of readability metrics for determining the difficulty level of a text have been devised for a range of different purposes such as determining the educational grade level of science textbooks or as an indication for authors for writing comprehensible technical manuals, to name just two very different ones. Most traditional readability formulae involve computing the sentence length and the number of syllables per word. A more recent approach combines sentence length and a unigram language model to determine difficulty, also taking into account concept difficulty [Si and Callan, 2001]. A rather different approach uses coherence as a measure of text difficulty rather than quantitative characteristics [McNamara and the CSEP lab, 2003]. The next section presents some of these metrics.

### 3.2.2 Readability Formulae

We will first turn to formulae that rely heavily on syllables and then move on to discuss some alternatives to these. Four of the most widely used formulae relying on syllables, and in three cases also on the average number of words per sentence, are the Flesch formula for reading ease, the SMOG readability metric, the Flesch-Kincaid formula and the FOG metric (cf. [Uitenbogerd, 2003, p. 425], [Liu et al., 2004, p. 548], [Si and Callan, 2001, p. 574]).

**The Flesch formula for reading ease (RE)** as described in [Uitenbogerd, 2003]:

$$RE = 206.835 - 0.846 * \frac{Syllables}{100Words} - 1.015 * \frac{Words}{Sentence}$$

Reading ease depends linearly on the average number of syllables per 100 words and the average number of words per sentence.

**SMOG** as given in [Si and Callan, 2001]:

$$GL = 3.0 + Sqrt\left(\frac{PolysyllableWords}{30Sentences}\right)$$

The grade level (GL) in the SMOG metric is a function of the square root of the number of the polysyllable words in 30 sentences.

If a document is exactly 30 sentences long, trivially, the number of polysyllable words is counted. If a document exceeds 30 sentences, the number of polysyllable words in the first 10 sentences, in the middle 10 sentences and in the last 10 sentences of the document are added up. If, however, there are fewer than 30 sentences in a document a conversion table and specific rules have to be used.

The SMOG metric gives higher readability values for a given text than the FOG and the Flesch-Kincaid readability metrics [Si and Callan, 2001].

**Flesch-Kincaid** as described in [Si and Callan, 2001]:

$$GL = -15.59 + 0.39 * \frac{Words}{Sentence} + 11.80 * \frac{Syllables}{Word}$$

In the Flesch-Kincaid formula the grade level depends on the average number of words per sentence and the average number of syllables per word. It is a U.S. Department of Defense standard and is used more frequently than the FOG and the SMOG metric [Si and Callan, 2001].

**FOG** as stated in [Si and Callan, 2001]:

$$GL = 3.0680 + 0.877 * \frac{Words}{Sentence} + 0.984 * MonosyllableWords$$

The two variables determining the grade level of the FOG formula are the average sentence length and the percentage of monosyllables. This metric is considered to give reasonably good results from older primary school level onwards [Si and Callan, 2001]. It has to be noted that the use of syllables in this formula differs from the usage in all the other formulae in that here the grade level increases with the number of monosyllables. In the other cases a high number of syllables or polysyllable words indicates a high level of difficulty, which is more intuitive.

Syllables feature in all of the above formulae. In the case of SMOG the number of polysyllables is even the only variable that is taken into account for computing the grade level. However, that a higher number of monosyllables, shorter words as it were, leads to a higher grade level in FOG contradicts the general view of the other formulae, i.e. that the higher the number of syllables the higher the difficulty. It also defies common sense.

There are also other approaches that do not take syllables into account. Experiments provide evidence that these perform more consistently than those which use syllables. One reason for that might be the contradicting usage of the syllable variable.

Another formula, this time developed by **Dale and Chall** in 1948, is presented in

[Uitenboger, 2003]:

$$GL = 3.6365 + 0.1579 * Wordlist + 0.0496 * \frac{Words}{Sentence}$$

This formula is based on the average sentence length (words per sentence) and the percentage of words in the text that are also on a vocabulary list containing 3000 pre-defined common words. It was shown to yield better results in experiments than the Flesch score for reading ease [Uitenboger, 2003].

Another much more recent approach has been proposed by Si and Callan. Experiments have shown that it is more consistent than the Flesch-Kincaid readability formula. Their approach is to not only measure the difficulty of a text by one or more useful surface linguistic features, which turned out to be sentence length, but they also claim to introduce concept difficulty into their computation by including a language model based on unigrams.

### 3.2.2.1 Discussion of Syllable and Sentence Length as Indicators of Difficulty

Evidently the two most commonly used features for computing readability are the number of syllables and the sentence length. The next section discusses their merits.

The reason for the popularity of the number of syllables in such formulae is probably that “in general word frequency and word length are inversely related” [Uitenboger, 2003, p. 428]. But it is surprising that the number of syllables should be a more reliable predictor than other features, such as the frequency of a word for example, as some infrequent “monosyllable words, such as *quark*, represent concepts that are not easy to understand” [Si and Callan, 2001, p. 574].

In the literature, too, to the usefulness of syllable-related indicators of difficulty is not undisputed. Si and Callan, for example, argue that the percentage of polysyllable words, e.g. as used in SMOG, is not a reliable indicator, because in their corpus “web pages written for grades 3-5 had more polysyllable words than web pages written for grades 6-8” [Si and Callan, 2001, p. 575]. This indicates that the number of polysyllables does not necessarily increase with difficulty.

Uitenboger found in an experiment that word length was rather an unsatisfactory indicator of readability [Uitenboger, 2003, p. 429]. She also mentions that the number of syllables per word “may differ markedly for different languages or definitions of syllable [Uitenboger, 2003, p. 428]”.

Sentence length has been found to be a more reliable feature than syllables. Si and Callan, for instance, obtained results from a sample corpus of 91 science web

pages that showed that sentence length is a good indicator for differentiating texts for the three grade level categories kindergarten-2, 3-5, and 6-8 because the mean values of sentence length in the three groups increase monotonically [Si and Callan, 2001, p. 575]. Uitenbogerd finds a significant correlation at the 95% confidence level of 0.68 and above between the perceived difficulty of a text and the number of words per sentence only. Other factors she investigated, such as vocabulary knowledge, were not significantly correlated with the perceived difficulty of a text [Uitenbogerd, 2003, p. 430]. However, these results must be taken with care as she only studied four subjects of varying ability including herself. Moreover all subjects judged a different set of books. Besides, it seems fair to say that this strand of research in general is littered with ridiculously small sample sizes and therefore shaky results.

Obviously, the list of variables discussed so far is not exhaustive and above all lacks certain factors which we would notice when judging a text, but which are difficult to measure numerically. Such factors may include the difficulty of grammar, the difficulty of contained idioms and phrases and many others.

### 3.2.3 Conclusion

Sentence length is more consistent than the number of syllables as an indicator of text difficulty. Combining several variables into a formula usually yields better results. The introduction of statistical language models improves the performance as more features are introduced. This raises the question of why most formulae only consist of two variables or even only one. Several other factors are probably important in determining the difficulty of a text and should be taken into account, but these cannot, or only with great difficulty, be assessed automatically. Separately assessing language and concept difficulty seems to be possible (with restrictions) but further research is needed here. A study by [Uitenbogerd, 2003] on reading material for learners of French verifies that sentence length is the factor that has the strongest correlation with the perceived difficulty of a text. This suggests that sentence length can be applied to foreign language material in general<sup>3</sup>.

---

<sup>3</sup>This should at least hold for languages cognate with French.

### 3.3 Applicability to Adult Language Learners

Since most of the formulae presented here are designed to determine the native language grade level of schoolchildren it has to be asked if this score can be mapped by some function to the ability level of adult learners of a foreign language, and if so, which grade corresponds to what level of language ability.

With regards to adult language learners the syllable as an indicator of difficulty has to be seen in a different light. Whereas native speaker children may have problems with words composed of many syllables, sophisticated adult language learners particularly from an indoeuropean language family background might find exactly those words very easy if they know them, their counterparts or the meanings of individual syllables from their native language or from other foreign languages they learned before. By contrast they may find (short) words with no similarity to other languages much harder to remember.

Another issue is that in the case of children, language ability, cognitive ability, and world knowledge develop roughly in parallel. The cognitive ability of adults who learn a foreign language, however, is usually fully developed. Thus despite having the capability to understand difficult concepts or complex causal relationships they may not understand a particular text in a foreign language for lack of language ability. So for them text readability consists of two potentially extremely unequal parts. The first is topic difficulty, which reflects the complexity of the concepts and contents. The other is language difficulty, which reflects the linguistic demands, such as grammar and vocabulary. For adult beginners the ability to cope with high topic difficulty will diverge dramatically from what they can cope with in terms of language difficulty. This difference will gradually decline as the language ability grows while the concept ability remains more or less the same.

So in order to provide adults with suitable texts any assessment of text readability would ideally have to incorporate this peculiarity of the gap between topic and reading difficulty which declines as the learners improve their language ability.

Furthermore this gap raises the fundamental question of whether there “is suitable reading material for [adult] learners on the web” [Uitenboger, 2003] at all. Unfortunately, [Uitenboger, 2003] does not provide an answer to that important issue in her paper. The available range of language difficulty levels for reading material of the same topic is probably relatively small. The tendency that the more elaborate a topic, the more complex the language, seems to pose a problem to finding suitable texts for

adult learners.

Assuming there are texts that are relatively easy in terms of language, but describe a complex topic such that both the information needs of an adult learner and his level of language skills are met, a possibility has to be found to separately assess the readability of the text in terms of both language and topic readability. The high correlation of language and topic complexity might be reflected in the readability formulae, but Si and Callan argue that commonly used features in readability metrics “ignore concept difficulty” [Si and Callan, 2001]. Therefore it might be possible to compute language readability by the use of traditional formulae and topic readability with the help of Si and Callan’s approach, which claims to incorporate information about the difficulty of the document content.

Having said all that it has become clear that the variables have to be tested with respect to their applicability to adult foreign language learners. This will be a by-product of designing a new formula to explain text difficulty and assess user ability, the process of which will be described in the following section.

# Chapter 4

## A New (Read)Ability Formula

### 4.1 Selecting the Training Material

A formula can only be as good as the gold standard data used for designing that formula, so caution is required when selecting the data. Texts especially written for adult language learners with a broad range in difficulty provide a very good basis for designing a new formula that will serve a two-fold purpose outlined in section 2.3: to assess text difficulty as well as learner ability.

Those texts would ideally reflect the special abilities, interests and needs of the target group. Textbooks for adult learners should at least partly reflect the fact that adults tend to have a higher concept ability than children, but are at a stage of low but rapidly growing language ability.

In order to determine empirically the variables that correlate with text readability and difficulty, a series of text books or graded readers appear to be most suitable. These should have as many volumes as possible so as to represent a wide range of abilities with a large enough sample. A series of textbooks has the advantage of offering a large number of shorter texts, each with a constant level of difficulty. Thus we obtain a large sample with a slow and steady progression from very easy to difficult texts. Graded readers have the advantage of containing longer texts and hence yielding more reliable results. However, graded readers are usually divided into no more than six vocabulary levels with many books on each level. In terms of regression analysis this means that there will be a maximum of six values of the independent variable, which is less than ideal to produce accurate results. Even less favourable is the fact that books on the same vocabulary level need not be of exactly the same difficulty; some will be rather close to the difficulty of books on the preceding level and some will be closer

to the difficulty of books in the subsequent level. Reasons for this will be the range of vocabulary, grammar use and differing style due to various topics. Another source of inaccuracies would be if the level of difficulty progressed significantly within a level - which makes perfect sense from a learner's perspective, but will blur the statistical results. Because of these reasons and the textbook reading texts being more similar to our target texts (newspaper articles), a textbook series is the preferred option for our study.

Unfortunately it was neither possible to obtain textbooks nor graded readers in electronic form from a publishing house. Despite that, the *New Headway* series of English textbooks for adult learners was decided upon to use in this study. With six volumes taking the learner from beginner to advanced level <sup>1</sup> this textbook series is comparable to school book series in number and range of difficulty, whereas most textbooks series for adults usually only offer up to three volumes. Moreover, like newspapers, it is relatively up-to-date and appears to be of high quality.

## 4.2 Preparing the Training Set

As the texts were not available electronically, all main reading texts were scanned, <sup>2</sup>, manually corrected and proofread, giving a dataset of 97 texts (observations).

The 97 texts are simply numbered in the order in which they occur in the books so as to preserve the progression in difficulty. Thus the first text in volume *Beginner* is assigned the value 1; 97 is the last text in *Advanced*.

Table 4.1: *Headway Statistics*

Volume	number of texts
Beginner	12
Elementary	12
Pre-Intermediate	12
Intermediate	19
Upper-Intermediate	15
Advanced	25
Overall	97

<sup>1</sup>The titles of the volumes in ascending order are Beginner, Elementary, Pre-Intermediate, Intermediate, Upper-Intermediate, Advanced [Soars and Soars, 2003]

<sup>2</sup>Texts written in white a dark background were omitted due to bad scanning quality.

Punctuation marks were added to titles and subheadings in the case of whole sentences, otherwise the whole phrase was deleted. This is to ensure that the variable *sentence length*, which plays an important role in determining difficulty (compare sections Readability Formulae/Conclusion and A New (Read)Ability Formula), is not distorted.

Eventually the texts were fed through a program to obtain values for the variables that are needed to calculate the formula.

## 4.3 Variable Selection

Following standard practice in the literature, the new formula was arrived at by regression analysis, which tries to determine what it is that makes some texts more difficult than others. We only included variables in the actual formula when we could find a rationale for its use and when it turned out to be statistically significant. The  $R^2$  for the training material would have been much higher had we used more variables. These strict criteria for inclusion of a variable were applied deliberately in order to produce a model of general validity rather than constructing a high  $R^2$  by overfitting the training data. This section presents and discusses the dependent and independent variables that were used to produce the final version of the formula. We also talk briefly about a number of variables that did not meet the criterion of significance. The section concludes with some suggestions of additional variables that might be worth investigating.

### 4.3.1 Dependent Variable

Ideally, we would simply regress text difficulty (dependent variable) on as many plausible candidates for determinants of difficulty (independent variables) as we wish. However, problems already start with the question of how to define and measure text difficulty. Different learners will find that a difficult text is characterised by different things. A universally accepted measure of text difficulty is probably no less elusive than a universally accepted measure of beauty. Inevitably, a proxy variable will have to be used - a variable that is highly correlated with whatever we think of as the 'true' text difficulty, but which can be easily and unambiguously measured. Whereas many similar studies use the grade level as a proxy for text difficulty, in this study the dependent variable *Difficulty* is made up simply of the numbers of the texts, i.e. a smooth progression from 1 to 97. The rationale behind this proxy for text difficulty is straightforward. The *New Headway* course was presumably compiled in such a way that each

text should, broadly speaking and with some exceptions, be more difficult than its immediate predecessor. Whatever standards the authors used to judge difficulty should also be appropriate for the present purpose since the aim is alike - to learn a foreign language by reading increasingly difficult texts.

The many advantages of its simplicity aside, this proxy for text difficulty has at least one major drawback. Its assumed value rises smoothly from 1 to 97 by constant increments. There is, however, no reason to believe that the 'true' level of difficulty of the texts does, too. In addition to natural fluctuations, a smooth increase in difficulty is hindered by two more facts. Firstly, a few of the New Headway texts are written in unusual language, e. g. archaic English. Secondly, some of the texts were unsuitable for automatic processing and therefore omitted, resulting in gaps in the progression.

Although the overall difficulty rises, some texts are easier than their predecessor. This problem will increase with increasing language difficulty as the speed of progression is probably highest at the beginning. The progression or learning curve will likely flatten out at higher skill levels.

Thus when the texts are not ordered strictly by their "real difficulty" the accuracy of the resulting formula will suffer. In this particular area there is clearly room for further research and improvement.

Despite this shortcoming, the method of measuring text difficulty employed here is likely to represent an improvement over many other approaches to the problem used elsewhere in the literature.

## 4.3.2 Independent Variables

### 4.3.2.1 Sentence Length

As sentence length was found to be the most reliable indicator of text difficulty we also use it in our study.

Loosely speaking, *Sentence Length* is simply the average number of words per sentence, that is the number of sentences per text divided by the number of words per text. A sentence in our definition is a string of words that ends on one of the punctuation marks dot (.), semicolon (;), question mark (?) or exclamation mark (!), followed by whitespace. This simplistic definition of sentence ensures that decimal points in numbers such as in amounts of money (\$111.45) do not count as punctuation marks, but nevertheless it has its inadequacies. It takes into account neither ordinal numbers (3. ) nor abbreviations (see p. 123, Mr. Blair).

Our definition of word (token) includes all sequences of characters that are separated by whitespace. In addition any commas, colons, double quotes and quotes are stripped off the strings. Most contracted forms such as *I'd* or *doesn't* are replaced by the two underlying words (cf. list of substitutions). By contrast, possessive forms ending on “s” on the other hand count as one item/token (linguistic explanation). It also has to be noted that no distinction is made between upper case and lower case representatives of the same sequence of characters.<sup>3</sup>

One of the main shortcomings of the variable as it is used here is that enumerations, for example of names, cause easy sentences to be quite long, and therefore to be rated as rather too difficult. “[...] including Orlando Bloom, Drew Barrymore, Halle Berry, Scarlet Johansson, and Kate Winslet”. Plans for alleviating that problem already exist but are not yet incorporated in the current version of the formula.

The graph below provides an illustration of the partial correlation between Sentence Length and our concept of text difficulty. A log-linear form was chosen for this variable for two reasons. The first one is empirical: The plot shows that the true relationship can be very well approximated by the logarithmic curve. But there is also a theoretical foundation. Even when the texts become very, very difficult, sentence length cannot increase indefinitely; there is a natural limit to what kinds of sentences are used in good English. Hence the relationship can indeed be expected to flatten out. The very high  $R^2$  of above 0.5 already shows that this variable plays a major (in fact, the most important) role in determining text difficulty.

#### 4.3.2.2 Word Length

The variable word length is the average number of characters per word. Its values are obtained simply by dividing the total number of characters in a text by the number of words. This variable is similar to the widely-used number of syllables, but it is easier to implement as it does not require the use of a lexicon. Thus neither a sparseness problem would arise when a token is not found in the lexicon nor an inaccurate estimate of syllables is obtained. We would of course expect to find longer words predominantly in more advanced texts.

Below we again provide a scatterplot of Word Length against Difficulty. Again the correlation is quite strong, although not as high as in the previous case. The same theoretical case for a log-linear functional form would probably apply here, but such

---

<sup>3</sup>None of this applies to languages such as Arabic, Chinese or Japanese where “words” are not separated by whitespace.

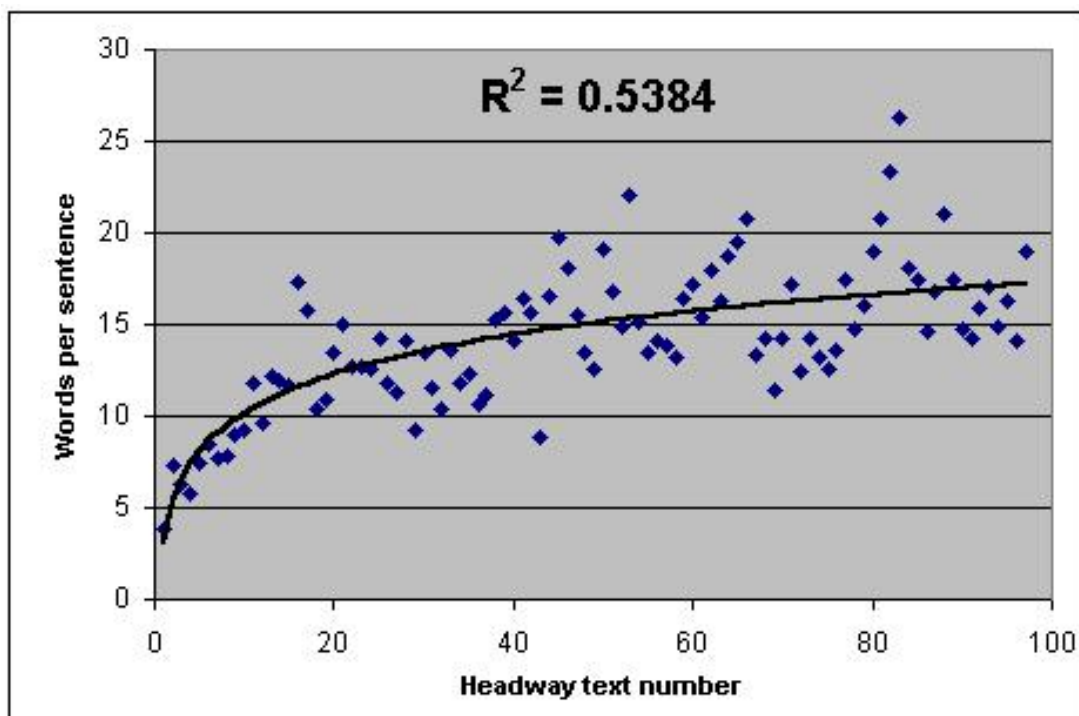


Figure 4.1: Sentence Length

a non-linear relationship is not borne out by the data, so we opt for simple linear regression.

#### 4.3.2.3 Conjunctions and Adverbs

The third independent variable to be included in the model is a - perhaps rather strange-looking - index, which is elsewhere also referred to as “Conjunction easy-diff”:

$$\text{Conjunctions} = \frac{\text{Easyconj} - \text{Difficultconj}}{\text{tokens}}$$

Easyconj stands for the total number of occurrences in a text of four words: *and*, *but*, *or* and *because*<sup>4</sup>. Difficultconj correspondingly refers to the total occurrences of a list of about 60 conjunctions and adverbs that are deemed to be difficult, i.e. tend to be used mainly by more advanced learners. Examples include words such as *furthermore*, *however*, *though*, *hence* and *merely*. The intention was that these are words for which easier and more frequently used substitutes exist.

The rationale for this variable is that the confident use of a variety of appropriate conjunctions and adverbs is by assumption usually coupled with high language

<sup>4</sup>With the introduction of this variable the formula is no longer universally applicable to any language, since the list of conjunctions is obviously language specific.

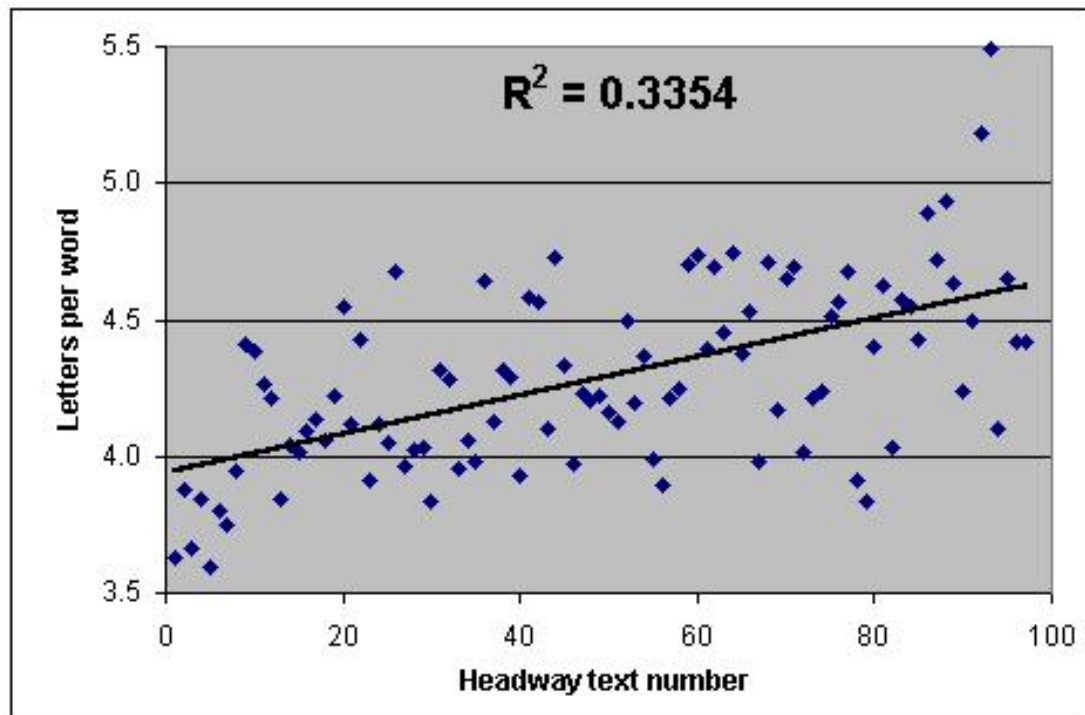


Figure 4.2: Word Length

ability. And although conjunctions are very specific words (data-driven) they usually occur/feature in every text, so there is no problem of sparseness.

If a learner only knows very few conjunctions or has problems paraphrasing, he will use the same conjunctions over and over again. Our formula takes that into account. If easy conjunctions are used frequently but no difficult conjunction is used the resulting value is high and positive. If only difficult conjunctions are used the value is high and negative. The use of a difficult conjunction compensates for the use of an easy conjunction so that an even mix of easy and difficult conjunctions will result in a value of zero as will the rare case of no conjunctions at all.

A potential problem is that a learner might have learned “fancy” conjunctions on purpose in order to polish his English. However, that is a costly signal to fake. The analogy would be that of a golf player who only practises his swing so that he will look good on the tee. Such cases are certainly rare and we may assume that all areas of language ability develop roughly in parallel in the majority of cases.

The judgement of what are easy and difficult conjunctions was made subjectively and should be supported by relevant research. As supportive evidence we could however mention that three of the easy conjunctions (and, but, or) are the three most frequently used ones according to the word frequency list of the British National Corpus

(BNC). Other very easy ones were left out as they are ambiguous as to their part of speech (as, so) or do not have a more difficult corresponding version that is frequently used.

The empirical results confirm our assumption that the use of conjunctions is a very revealing indicator of language ability. The graph below shows a reasonably clear negative correlation between the conjunctions variable and text difficulty.

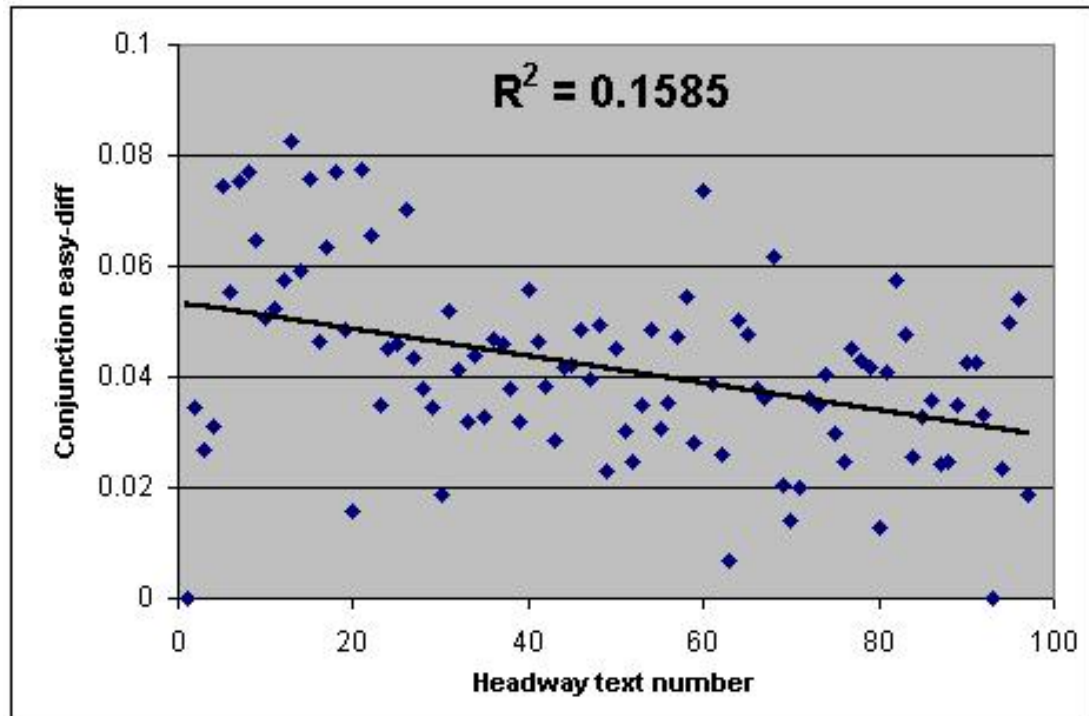


Figure 4.3: Conjunction easy-diff

In summary, the three variables discussed above appear to be strong candidates for inclusion in the formula, with Sentence Length taking the lead. In each case, a sound theoretical background, either from the literature or from our own reasoning, combines with a sufficiently high partial correlation. Since one of the aims was to investigate whether the standard variables used in the literature also apply to the context of adult foreign language learners, we can confirm that this is certainly the case for Sentence Length and Word Length.

### 4.3.3 Neglected Variables

#### 4.3.3.1 Wordlists

Dale and Chall quite successfully used a list of 3000 common words. This seems like an attractive and reasonably straightforward variable to include, however the implementation proved to be not quite that easy.

A first tentative attempt to include ad hoc word lists of various sizes showed that none of them were significantly correlated with text difficulty. The sign of the correlation even changed as we varied the size of the lists. It seems that in general lists of the most *frequent* words are not well suited to our purpose. That is especially true when these lists are compiled from large corpora that include publications outwith the usual reach of language learners.

A more promising approach would be to use lists of words that are typically part of a learner's curriculum. However, such lists<sup>5</sup> are generally not freely available in electronic form.

A drawback of wordlists is of course that they are language-specific so that any formula based on them cannot be readily applied to other languages.

#### 4.3.3.2 Tokens/Types

Another variable that was at one stage included in the formula but later dropped is the ratio of tokens to types in a text. There should be a positive correlation with difficulty as we expect that in beginners texts words are repeated more often than in advanced learners texts. Again this variable proved to be insignificant. A more precise specification in future work may however improve the results.

### 4.3.4 Further Ideas

#### 4.3.4.1 First and Second Person Possessive and Personal Pronouns

We observed that first and second person possessive and personal pronouns<sup>6</sup> are used more extensively in easy texts than in difficult ones. Another way to look at it is to divide the New Headway texts into two categories: The first group often deals with topics such as geography or history and contains no, or hardly any direct speech. Texts in the second group revolve around personal fortunes and perspectives, and contain a

---

<sup>5</sup>Such as the German ZDaF Mindestwortschatzliste.

<sup>6</sup>I, we, you, my, mine, me, our, your, ours, yours

lot of direct speech. Common sense tells us that the latter type is usually easier to read and also contains more personal and possessive pronouns. This variable is conceptually very appealing but unfortunately could not be included due to time constraints.

#### 4.3.4.2 Adverbs and Attributive Adjectives

The more advanced a learner is the more collocational and ornate his constructions and vocabulary will tend to be. Another good indicator of language ability thus might be the use of adjectives before nouns, adverbs before verbs etc. To assign part of speech tags with a relatively low error rate would be possible for texts from the textbooks. Tagging learner's texts would probably yield a higher error rate though, because there will be unusual constructions due to grammatical errors.

## 4.4 The New Formula

The new formula is obtained through regression analysis, using the three independent variables discussed in detail above in addition to our measure of text difficulty. The calculation was performed by MS Excel with the *AnalyseIt!* add-in. This is the result:

$$(Read)Ability = 137.6 + 47.3 * \log \frac{words}{sentence} + 19.2 * \frac{characters}{word} - 447.3 * ConjEasyDiff$$

Figure 4.4: The new formula

The table below reproduces the coefficient estimates together with the probabilities (p-values) that the true coefficients are, in fact, equal to zero. Since the p-values are all way below the standard critical value of 5%, we can reject with great confidence the null hypothesis that the coefficients are zero. All three coefficients are therefore significant and of the expected sign.

Table 4.2: Regression Results

Term	Coefficient	p-value
Intercept	-137.57	0.000
ln(w/s)	47.28	0.000
l/w	19.16	0.004
Conj easy-diff	-457.34	0.000

With an  $R^2$  of over 0.6, the model also has a very good overall fit. More specifically, this means that over 60% of the variation in text difficulty can be explained by our three variables. Adding further explanatory variables to the model would certainly increase the fit even further, however most of these variables will be insignificant. The best way to further improve the model, other than by adding more variables, is probably to try to obtain better data. As discussed before, our proxy for text difficulty, despite comparing favourably to much of what appears to have been used in the literature, is probably the limiting factor in this statistical analysis.

As a final illustration, the scatterplot below graphs the actual versus predicted difficulty levels. With the exception of a few outliers, the model is able to accurately predict text difficulty, at least when compared to the original New Headway texts.

A point worthy of future investigation appears to be that the scatterplot shows a slightly non-linear relationship. That seems to suggest that maybe the functional form of the model could be improved.

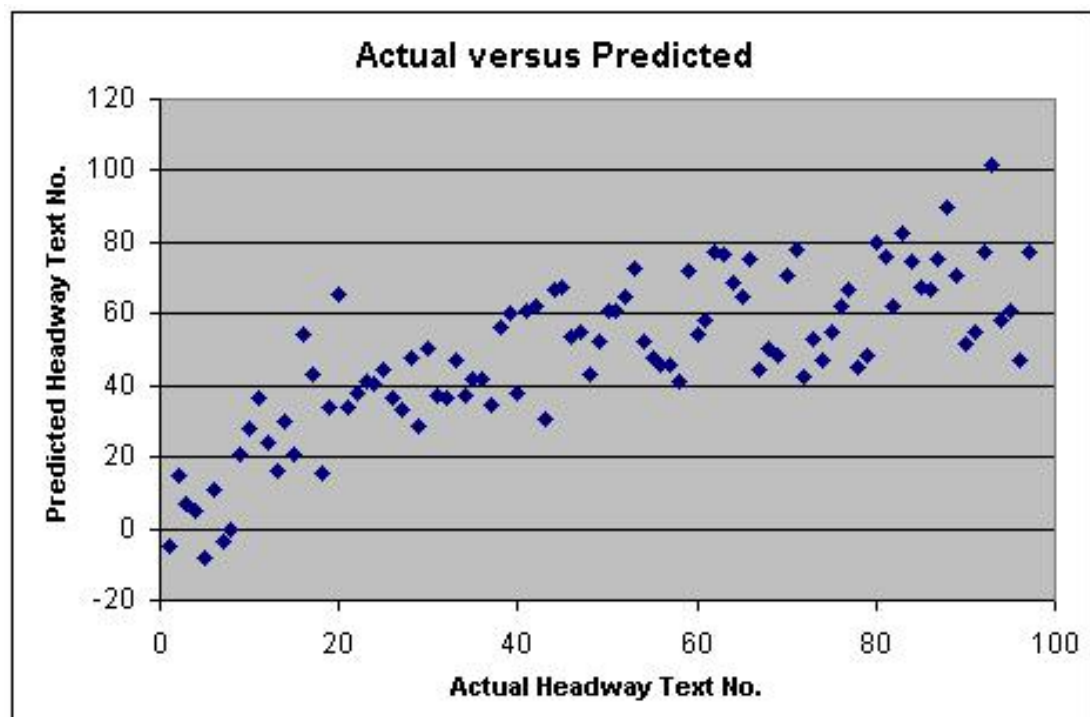


Figure 4.5: Actual versus Predicted (Read)Ability

# Chapter 5

## User-specific Text Retrieval

### 5.1 Text Source - Newspaper Corpus

Ideally the user would be able to search the whole internet to guarantee a variety of topics and difficulty levels. For the purpose of the project however, it will not be appropriate to do an online search over the web as in that case sensibly evaluating the results would not be possible. Moreover, given the time constraints, problems that internet-inherent properties would pose such as the constant change of data and links will be left aside. Instead a corpus is created by the use of a perl script that automatically downloads articles from newspapers that are presumed to be diverse in difficulty and style. Articles from BBC news, CBBC news, The Guardian, The Financial Times and The Sun websites were retrieved by the script between the beginning of March and the 12th of August 2005. The intention of this course of action is to make available texts on the same topic, but of varying degrees of difficulty as could be retrieved when searching over the internet, but yet being provided with knowledge about the composition and size of the text collection.

A risk here was that topics among newspapers do not overlap enough. In that case a less preferable alternative would have been to pose a query to Google in order to find web pages of the relevant topic and calculate the difficulty of a certain number of pages, which would pose problems to the evaluation. Topics in fact overlap enough so that no alternative solution is needed.

The preprocessing of the html source code in order to use it for searching is described in Chapter 7.

## 5.2 Integrating Lucene

### 5.2.1 Introduction to Lucene

The open source Java software Lucene initially written in 1997 by Doug Cutting and in 2000 put up on SourceForge, and in 2001 migrated to Apache is used as a search engine. Now several programmers are working on it and it has been translated into several other programming, such as Perl Plucene which we use. “It powers search in diverse applications like discussion groups at Fortune 100 companies, commercial bug trackers, email search supplied by Microsoft, and a web search engine that scales to billions of pages” [Gospodnetic, Hatcher, 2005, p. XVII].

## 5.3 Indexing

“An index is a critical data structure because it allows fast searching over large volumes of data. Different index structures might be used, but the most popular one is the inverted file” [Baeza-Yates and Ribeiro-Neto, 1999, p. 9].

The whole process of indexing includes defining the text information to be indexed and the way how this information is analysed before it is actually written to the index. For simplicity, we choose to always build an index from scratch whereas for an up-to-date web application which takes newly published newspaper articles into account we would have to incrementally add new documents to the index (and remove old information).

Often texts to be indexed will be available from separate files. We, however, extract the information to be indexed from an XML database and pass it directly to the analyser. There are many different alternatives for analysing. The fundamental principle of the analyser is to specify the nature of the tokens to be indexed, that is essentially the amount of (necessary) information to retain and the amount of (useless) information to eliminate from a document. The definition of what is useful or useless differs with respect to the domain/circumstances. [Baeza-Yates and Ribeiro-Neto, 1999] call the analysing operations text operations or text transformations and say that its aim is to “reduce the complexity of the document representation and allow moving the logical view from that of a full text to that of a set of index terms” [Baeza-Yates and Ribeiro-Neto, 1999, p. 5]

An overview of text transformation operations is presented below.

Eliminating stop words is among the most common text representation options, “frequently used words that do not help distinguish one document from the other” [Gospodnetic, Hatcher, 2005, p. 20], i.e generally function words such as the, a, ... We can extend the list of stop words to other words that are frequent in certain domains that are to be searched for example.

We do not consider it necessary to exclude stopwords from the index as a learner has different information needs from other users who primarily look for information content. Learner’s information needs additionally might include looking for texts containing common words in order to get more familiar with their use.

Another wide-spread option, which we will use is to offer a case-insensitive search.

It is also popular to use stemming which treat all words of a common grammatical root as if they are the same. Thus when searching for singular also the plural occurrences or other word classes would be returned.

Numbers are often skipped while indexing as there are very many of them and they are only seldom searched for.

Sometimes an identification of noun groups is performed, whereby adjectives, adverbs and verbs are eliminated [Baeza-Yates and Ribeiro-Neto, 1999]. This is absolutely inappropriate for our purpose.

Another idea is to introduce orthographic variation for spelling differences e.g. British American or even anticipate common spelling mistakes. We do not incorporate this into our application, but if we did we would notify the user when it is recognised that he has mistyped a word for pedagogical reasons.

## 5.4 Searching

The knowledge about the user and the possibilities of assessing the properties of texts have then to appropriately be incorporated into the actual retrieval process. There are two different possibilities of incorporating the context of the searcher into the retrieval process. Either the query posed by the user is altered prior to the actual search or the initial query is submitted for search and the results are then reranked according to the knowledge about the searcher [Belkin et al., 2004].

For the purpose of the system it is imaginable that some knowledge about the user will be incorporated into the query through query expansion and some might be taken into account when reranking the retrieved texts. Information about interests, context and some about language level could be used to expand the original query. The re-

rieved text could then be investigated in greater detail in terms of topic and knowledge difficulty and reranked on the basis of weighing the factors in an appropriate way that has yet to be determined.

# Chapter 6

## Updating the User Model

### 6.1 Ability Score

In this section two alternatives for updating the ability score are presented. One of them is entirely based on explicit user actions while the other can be regarded as a combination of explicit and implicit information. The first method that results in an update of the ability is to upload a text for ability assessment on the part of the user. This should be a text that was currently written by the user himself. The second method is to update the ability score according to how the user rates a text from the text finder, where the difficulty score of that text is also taken into account.

The first method is more accurate than the second one because the rating is only a subjective estimate, and is done on a rough scale with just five levels. Also inaccuracies that arise from calculating the difficulty of the newspaper articles are likely to be multiplied in the rating method. Moreover certain constellations of user ability and text difficulty may result in wrong inferences about the true ability of the user (see below).

Both methods are used in our application, i.e. the user can change his ability score either by continually rating new texts or by uploading a text he wrote himself. However, for the reasons discussed above, the latter method is programmed to result in a much greater change in the ability score.

#### 6.1.1 Update Rate

Our approach for updating the ability score is presented below in general terms, as well as in terms of the two different methods. We first need to stipulate how fast the ability

score should be updated, that is how much influence a new value has on the overall ability. This will be referred to as the update rate. The update rate can be in the range of 0% modification of the old value, i.e. no update at all, and 100% modification of the ability score - i.e. the score jumps immediately to the new value. Both limiting values may seem sensible to use at first, but the first modification to the ability score should indeed be 100%, namely when the learner uses the program for the first time. Thereafter, the update rate should decline steadily for at least two reasons. Firstly, the ability score will become an ever better reflection of the learner's true ability. Once the ability score has levelled out at the true value, we would not want it to fluctuate wildly, maybe in response to texts of uncharacteristic difficulty or measurement errors. Secondly, we expect learning progress to be slower at more advanced levels, so that too should be reflected by a lower update rate.

### 6.1.2 Update by Rating

Since updating the ability score by uploading a new text is a relatively straightforward process, we focus instead on the method of rating. Depending on the constellation of the ability score versus the difficulty value, a given rating of a text by the user will have different effects. For each of the ratings *too easy*, *somewhat easy*, *just right*, *somewhat difficult*, *too difficult* we specify two distinct formulae.

When a text is rated as *too easy* the following assumptions apply:

We first turn to the case where the difficulty score is higher than the ability score<sup>1</sup>. If the "adjusted"<sup>2</sup> difficulty score is higher than the ability score, we expect the reader to rate the text as *difficult*<sup>3</sup>.

If yet he rates it as *too easy*, his ability score should be increased. The increase should be bigger the greater the difference between ability and adjusted difficulty score. Our formula incorporates that constraint. A new ability score is obtained by adding the weighted difference between the adjusted difficulty score and the ability score to the ability score.

$$\text{newabilityscore} = \text{ability} + 0.6 * (\text{adjusteddifficulty} - \text{ability}) \quad (6.1)$$

<sup>1</sup>The texts should usually be selected in this way in order to challenge the reader. Cf. i+1-hypothesis

<sup>2</sup>This is the difficulty score minus the constant that makes up for the gap between reading and writing ability.

<sup>3</sup>From hereon *difficult* includes both ratings *somewhat difficult* and *too difficult*. Accordingly *easy* stands for *somewhat easy* and *too easy*.

When on the other hand the ability score is much higher than the difficulty score the student should find the text easy. If the ability score is still higher but both scores are almost equal we would expect the student to give the rating “just right”. Thus if the student rates the text as *too easy* whilst the ability score is much higher than the difficulty score, we cannot increase the score. We can however draw the conclusion that a student’s true ability exceeds his ability score when a text’s difficulty score is only very slightly less than the ability score when he rates it as *too easy*. Under these circumstances the score should increase. The closer the two scores are to equality the greater the increase should be.

$$newabilityscore = ability + 0.01 * \frac{1}{(adjusteddifficulty - ability)} \quad (6.2)$$

In the case of the rating *somewhat easy* the formula is the same in principle but the update rate should be lower as previously, as the rating *somewhat easy* signifies a weaker statement between the ability score and the difficulty score than the rating *too easy*.

So we have two formulae that correspond to 6.1 and 6.2 but use different weightings:

$$newabilityscore = ability + 0.4 * (adjusteddifficulty - ability) \quad (6.3)$$

$$newabilityscore = ability + 0.001 * \frac{1}{(adjusteddifficulty - ability)} \quad (6.4)$$

A reader should rate a text as *just right* when his score is approximately equal to the difficulty score. In that case there should be no change to his ability score. There should be an increase if his score is well below the difficulty score and this increase should be the larger the more obvious the difference between the ability score and the difficulty score, i. e. the lower the ability score compared to the difficulty score. If his ability score is higher than the difficulty score, the decrease in his ability score should be greater the bigger the difference. As the ratio is the same in both cases only one formula is necessary to capture the update, for which the difference in difficulty score minus ability score can be positive, negative or zero and thus increase, decrease, or not change the ability score at all:

$$newabilityscore = ability + 0.6 * (adjusteddifficulty - ability) \quad (6.5)$$

If the learner's score is much lower than the difficulty score he should find the text difficult. If he rates it *somewhat difficult* indeed his score should stay the same. If the learner's score is almost the same as the difficulty score but still lower, a slight decrease in his ability score should ensue.

$$newabilityscore = ability - 0.001 * \frac{1}{(adjusteddifficulty - ability)} \quad (6.6)$$

Conversely, if the learner's score is higher than the difficulty score and the rating is *somewhat difficult*, he should never find it difficult. Thus the higher the learner's score compared to the difficulty, the greater a decrease in ability score should result.

$$newabilityscore = ability - 0.4 * (adjusteddifficulty - ability) \quad (6.7)$$

The ratings *somewhat difficult* and *too difficult* relate to one another in the same way as *somewhat easy* and *too easy* in that the latter rating each time is more extreme than the former. Therefore the rating *too difficult* can lend itself to equations 6.6 and 6.7, but with a stronger weighting.

$$newabilityscore = ability - 0.01 * \frac{1}{(adjusteddifficulty - ability)} \quad (6.8)$$

$$newabilityscore = ability - 0.6 * (adjusteddifficulty - ability) \quad (6.9)$$

The development of the ability score should be displayed to a user as a curve plus the appropriate regression line. Whenever the user has uploaded a text he should be given the possibility to delete certain scores. This is essential in the case of mistakenly uploaded non plain-text documents such as word files or even pictures.

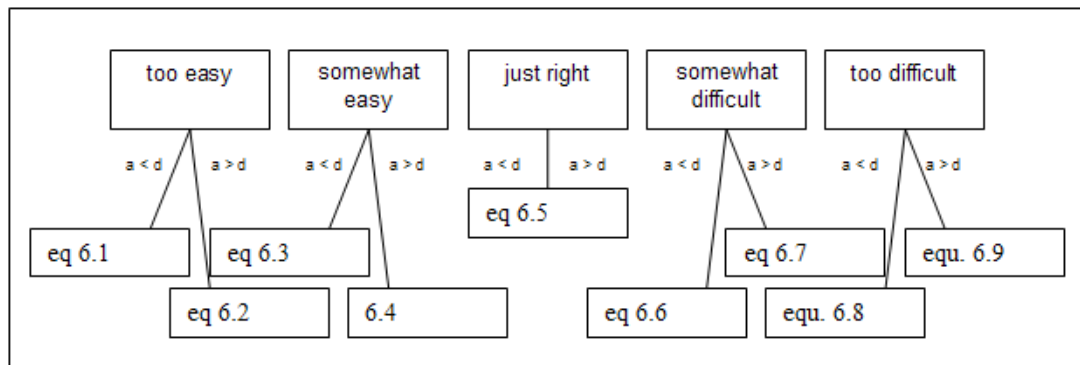


Figure 6.1: Summary of ratings and corresponding updating formulae depending on the relation between the user's ability score and the difficulty score of the rated text. The letters  $a$  and  $d$  stand for ability score and difficulty score respectively.

## 6.2 Constant

As one learner will want to learn faster than another we can have different weighting of update-speed. Or we could adjust the difference by adding a personal constant to the ability score, that is bigger constants for learners who want to learn faster than others. But this is a concern for future work.

# Chapter 7

## Description of the Application

We decided on a web-application as it is easily accessible from everywhere and is always up-to-date.

In this chapter first an overview of the architecture of the application is given. Following that all programs and important files are presented in more detail. Difficulties during the implementation are discussed and the solutions presented.

### 7.1 Architecture

This section gives an overview of the architecture of the system. We start with the presentation of the main architecture that is the textfinder's core. We then give an account of how we analysed the *New Headway* texts. Finally we briefly come to talk about the electronic questionnaire.

Our system is made up of standard perl scripts and html pages from which perl cgi scripts can be invoked. The perl scripts access and modify existing files in xml/plain text format or write new files to disk.

The structure of the architecture can be observed in figure 7.1. It can be divided into two parts. One part, marked as *Offline* is remote to the user, that is the user will not directly interact with that part. All programs are run and files prepared off-line before the user uses the system. For the part *User involved* it depends on the user which programs are evoked. He also influences by his actions in what way data about himself is modified and what files are created.

Our starting point is the perl script *build\_newcorpus.pl*, which extracts newspaper articles coded in html from newspaper websites. These are written into distinct corpora files (*bbc etc.*) in order to allow processing that is adjusted to the peculiarities of the

structure of each newspaper. The distinct files are all preprocessed by the script *preprocnewscorpus17.pl* and written to a single xml structure *corpus1.xml*. This xml file is not in UTF-8 encoding as the extracts contain special characters. Thus we need to convert the special characters into proper UTF-8 by use of the script *conversionUTF8.pl*. For the resulting xml database *corpus\_conv.xml* we specified an xml schema that ensures conformity (in addition to well-formedness) when checking with a “validator” against the xml data. The file *indexer\_plucene.cgi* extracts certain meta data and text content from element nodes which is then written to the index for faster access during the search. The index is in fact not just one file but a directory containing several files.

The starting point for the user is the login page (see figure 7.2). He invokes the script *login.cgi* by submitting his user name and password. This script checks with the file *users.xml* if the details are correct. If login fails, a page points out to the user that his details were not correct<sup>1</sup>, otherwise he arrives at the main page (see figure 7.3). From here either queries can be entered in order to find a text or texts can be uploaded.

If a text is uploaded, the script *upload.cgi* assesses it by our (read)ability formula (cf. 4.4), updates the ability score accordingly for the appropriate user and writes the text to a file labelled with the user id in a separate directory. Submitting a query invokes the program *plucene\_searcher.cgi*. That script makes use of the index files and displays all texts containing the query term as a list of retrieved documents which are ordered by difficulty (see figure 7.4). If a text is chosen by clicking on its link, the script *display.cgi* extracts the text from the xml database and displays it on a web page. This web page also contains buttons to rate the perceived difficulty of the text. Submission of a rating calls the script *rating.cgi*. It changes a user’s ability score depending on how the user rates a given text relative to his original ability score and the actual difficulty level of the text. It also brings the user back to the mainpage.

## 7.2 Important Program Descriptions

### 7.2.1 General

Whenever a script accesses an XML file the perl module `XML::LibXML` is used to build a tree structure which can be read and modified. One way of querying it is by xpath commands. We use a Document Object Model (DOM) representation so that we can create and replace nodes easily, although not all “functions” listed in the

---

<sup>1</sup>This is not explicitly shown in the diagram.

documentation of the module work properly so that we had to find less elegant ways around these problems. It is not clear whether the process of simultaneously accessing the xml file by several users is robust.

## 7.2.2 Offline

### 7.2.2.1 *build\_newcorpus.pl*

The program crawls newspaper sites by following certain links to news articles from one main page or several category pages of each newspaper. From Children's bbc we selected eight topic pages, each containing only a small number of links, from which links are followed. Both for bbc news and the Financial Times only one page is necessary that contains a summary of many links to important newly published articles. From the Sun site we selected the categories news and sports. The Guardian is split into 14 categories from all of which we follow links to articles.

By regular expressions it is specified that no external links or links to summary pages but only links leading to article pages are pursued. The contents of the pages are written into one of the newspaper corpus files *bbc*, *cbbc*, *guardian*, *ft* and *sun*. The files were saved to a different place from time to time, so that now there exist 18 different corpora files (*cbbc1* to *cbbc4*, *bbc1* to *bbc4*, *financial\_times1* to *financial\_times4*, *guardian2* to *guardian4* and *sun2* to *sun4*).

The newspaper's links pages are updated at different intervals ranging from daily in the case of the Guardian to within minutes in the case of bbc news. The script was usually run every day and often several times per day until the 12th of August. Prior to adding a new article to the corpus files, the program checks for each page if it is already contained in the corpus, in order to avoid duplicates. For *cbbc*, *bbc* and the *guardian* the unique identifier for that purpose is the article number contained in the link and contained in the previously saved web page. In the case of the *financial times* and the *sun*, titles are checked against the titles in the corpus files. For all unique identifiers it is indicated on the screen whether or not they are added to the corpora. Due to emptying the files by moving their contents to a different file, duplicates may be introduced as the new files are not checked for duplicates.

### 7.2.2.2 *preprocnewscorpus17.pl*

The structure of the web pages of the newspapers is very different. Thus we analysed each of them carefully and produced distinct regular expressions which extract

the desired data from the pages. A problem here was that even within one and the same newspaper the structure was not always the same, so that a compromise between too general and too specific regular expressions had to be found. Frames, links not contained in the actual text and other superfluous data are not retained.

Both part of the source HTML text and the main article in plain text are saved in an XML datastructure. The HTML representation is saved in the XML file so that it can directly be displayed nicely formatted on the webpage again when the user selects that article. The main article text without any mark-up is used for the calculation of the text difficulty before it is written to the XML database, too. The plain text is preserved for later use in indexing.

In addition to the actual text we separately write the title, the publishing date and the difficulty value of each article to the xml database. Each article is identified by an id and the source of the article is specified as a parent.

#### 7.2.2.3 *conversionUTF8.pl*

We wrote this simple program to convert special characters such as  $\text{\textcircled{O}}$  to numeric character references  $\text{\#169}$  in this case. Character entity references could not be used, they led to an error message. We manually created a list of the special characters and their equivalents such as the more common  $\text{\textcircled{C}}$ ,  $\text{\textsterling}$ ,  $\text{\text{u}}$ , and those which we would not expect to see in English newspaper articles  $\text{\text{a}}$ ,  $\text{\text{I}}$  and  $\text{\text{z}}$  as we initially did not expect that the texts contained such a big variety of special characters.

#### 7.2.2.4 *indexer\_plucene.pl*

This file produces the index that is important for efficient search. For each article id it extracts the text to be indexed from the xml database *corpus\_conv.xml*. For indexing the text stripped of html tags is used. We use Plucene's *SimpleAnalyzer* to analyse the text, that is tokens are separated by whitespace. Numbers, punctuation, etc. are removed. Document objects are then created which contain the article id, the title, the content, the date, the article source and the text difficulty as separate fields so that each could be searched for independently. The difficulty is used as the keyword field. That is because with Plucene, number searching, which we intended to incorporate, is only possible for the keyword field at the moment.

## 7.2.3 User involved

### 7.2.3.1 *login.cgi*

This is the first program invoked by the user when submitting his access data. It checks if the submitted user name and password match the entries for the user id in the users database *users.xml*. If the login data are incorrect it will be pointed out to the user on a separate html page. If correct, the user gets to the main page. His name is displayed together with his last login date and his current login time is written to the *users.xml* file.

The user id will be passed on as a hidden field when a button on the main page is hit that either calls *plucene\_searcher.cgi* or *upload.cgi*. The query term is passed on to *plucene\_searcher.cgi* and the text for uploading is passed on to *upload.cgi*.

### 7.2.3.2 *plucene\_searcher.cgi*

This script makes use of the previously submitted query term. Plucene cannot deal with empty queries. Thus we test whether no query is submitted and if that is the case we display a separate page saying “You have not submitted a query. Therefore no results...”. Otherwise a page listing results is displayed, which can be empty if there are no hits. Here again we use the *SimpleAnalyzer*. We specify that we want to search for the query term in the contents field, which contains the text of the newspaper article. We get a list of results which we order by the difficulty of the articles. For each hit all information previously written to the fields is displayed together with a snippet showing the query term in context. The title actually is displayed as a link, which contains the user id, the id and the difficulty as additional arguments which we also want to pass on to *display.cgi* when a user clicks on the link.

We initially used the snippet routine provided by [Cozens, 2004]. This routine is very slow however and as the runtime is restricted from the side of the service provider only very few hits could be displayed. We therefore created our own version for displaying the keyword in context (KWIC) which is much faster. It displays up to 15 words to either side of the query term, which stands out in bold type. For queries made up of two words or more, snippets are displayed twice when the words stand close together. It should be no problem, however, to correct that in future versions.

### 7.2.3.3 *display.cgi*

This script extracts the html version of the selected text from the corpus database and displays it together with buttons for ratings on a page. Upon hitting one of the rating buttons, the type of rating and the hidden field arguments user id, title, article id, and the difficulty score are passed on to the invoked script *rating.cgi*.

### 7.2.3.4 *rating.cgi*

*Rating.cgi* writes the title of the read text with the corresponding difficulty of the text, the user's rating and the user's current difficulty score to the user database. Based on that the ability score is updated as described in section 6.1. Also after rating the main page is reached again.

### 7.2.3.5 *upload.cgi*

There are two cases to be differentiated when uploading. The user has either specified a file for uploading or written a text to the text area. Depending on which way the user has chosen the process of handling the text is slightly different. In both cases the text is assessed by the (read)ability formula and then written to a separate directory marked with the appropriate user id. A separate directory is used in order to avoid the risk of accidentally overwriting scripts or important data files in the case that a user uploads files with the same name as one of our files. The calculated ability score is then used to update the user's ability score (cf. 6.1). The possible file size for uploading should be restricted.

## 7.3 Difficulties

## 7.4 Instructions for Rerunning the Application

All scripts and files created for this project are ready for inspection at ... on the School of Informatics server. Some of them cannot be run there as certain perl modules are not installed on the informatics computers and could not be installed locally.

The web-based interactive system can be run at [www.josiephil.de/textfinder/login.html](http://www.josiephil.de/textfinder/login.html). To access the site ... and ... are necessary as user name and password. The personal login for marker 1 is user name ... with password ... and for marker 2 the access data are ... together with .... All files and programs that are stored on the server will be available

on the informatics file system, too, but most cannot be run there as the Plucene module and other modules are not set up. <sup>2</sup>

---

<sup>2</sup>Added August 14th, 2006: Personal logins can be requested from [jasmine.bennoehr@li-hamburg.de](mailto:jasmine.bennoehr@li-hamburg.de)

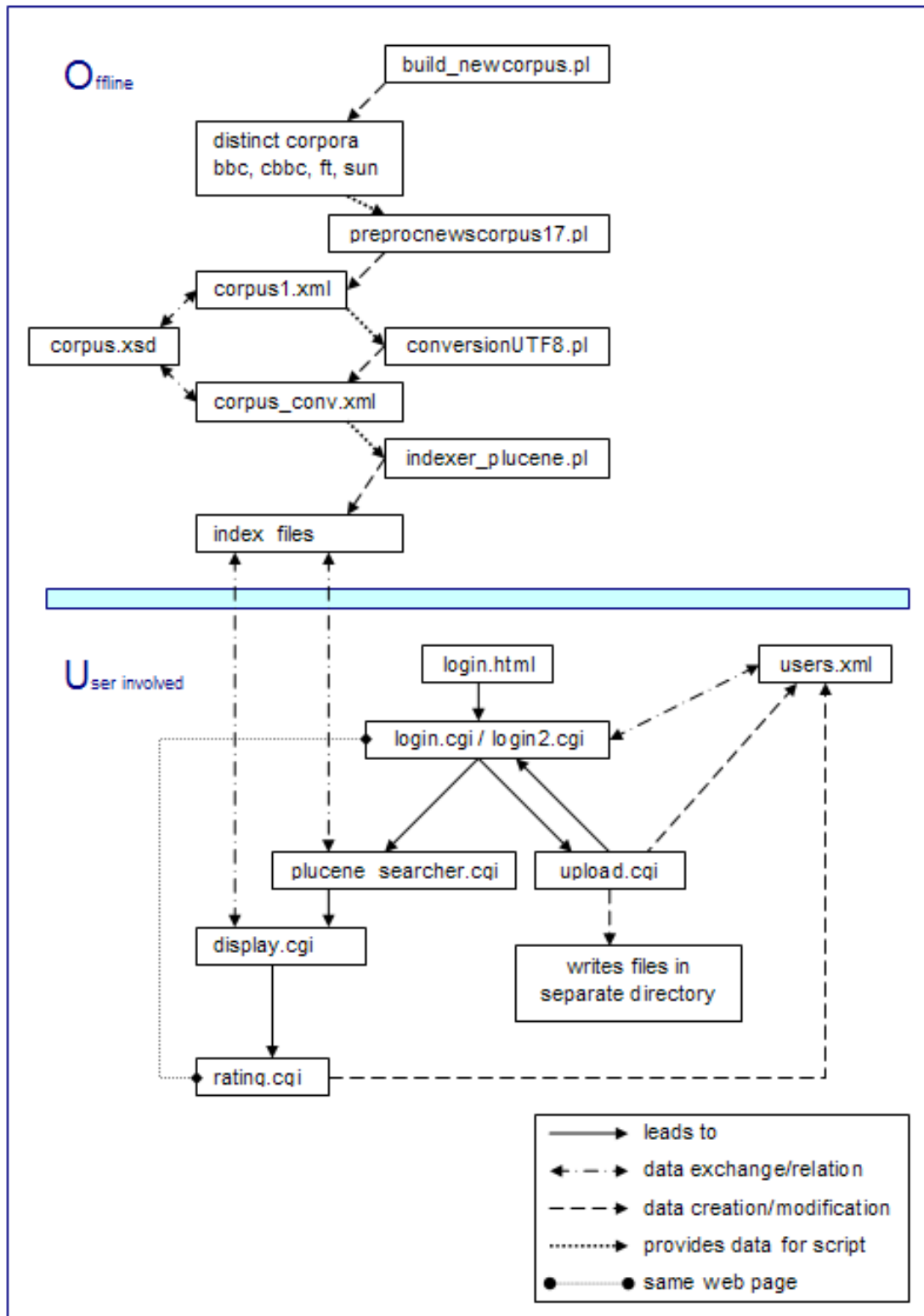


Figure 7.1: The diagram of the architecture displays dependencies between programs and files. Also the division into a part of the architecture which does not involve the user (offline) and one where the user interacts with the interface (user involved) can be observed.

Enter your name and password, please.

Name

Password

Figure 7.2: The login page.

A Web-based Personalised Text Finder for Language Learners [Fragebogen](#)

---

jasmine logged in at: Sat Aug 20 15:25:36 CEST 2005      last login at: Wed Aug 17 10:43:30 CEST 2005      reading score: 70

### Search for texts

enter query:

### Determine reading score

upload currently written file:

file must be plain text, for example .txt extension

enter text:

---

Project for the M.Sc. in Artificial Intelligence at the [School of Informatics](#) at the [University of Edinburgh](#).  
Mai-August 2005, Supervisor [Helen Pain](#)  
last updated at 20th July, 2005

Figure 7.3: The main page.

Your ability score is: 66.91

Feel free to select a text that is close to your ability score. The results are displayed in order of ascending difficulty.

Query: reading

Results:

⋮

**Title:** [Hotseat | Hotseat: Top author Michael Rosen](#)  
**Difficulty:** 46.01  
**Date:** Thursday March 03 2005 20:06 GMT  
**Source:** cbbc  
 ... Do you think children should start **reading** at a young age? Michael: I think that children should be read to from a ...  
 ... young age, and if they start **reading** when they're young too then that's fine. And if they don't they generally catch ...

⋮

**Title:** [Princess Beatrice 'has dyslexia'](#)  
**Difficulty:** 65.75  
**Date:** Wednesday March 23 2005 18:36 GMT  
**Source:** cbbc  
 ... Beatrice's mother, the Duchess Of York, said that her daughter was having help with her **reading** and writing skills. ...  
 ... About one in 10 people are affected by dyslexia. Most of them have difficulty recognising, **reading** and spelling words. The Duchess was speaking as part of her work for a charity ...  
 ... Children, which sends volunteers to city schools to help pupils who are behind in their **reading** and writing. She revealed that she too was ...

**Title:** [Your Reports | Making a film is great fun!](#)  
**Difficulty:** 66.77  
**Date:** Sunday March 27 2005 09:00 GMT  
**Source:** cbbc  
 ... Ben, 11, **Reading** If you're interested in film-making click on the First Light web link on ...

**Title:** [Your Reports | 'Schools should do more to beat the bullies'](#)  
**Difficulty:** 67.51  
**Date:** Tuesday March 22 2005 13:57 GMT  
**Source:** cbbc  
 ... If any of you **reading** this report are being bullied, tell someone before it's too late. Remember bullies can't harm ...

⋮

**Title:** [Sir Rees Davies](#)  
**Difficulty:** 91.90  
**Date:** Thursday May 26, 2005  
**Source:** guardian  
 ... farmers. His parents quickly recognised his intelligence and encouraged him in an early love of **reading**, though he must have been seven or eight before he learned English. From a village ...

Figure 7.4: An extract of the page that displays query results.

## Drivers clamp down on 'parking ticket' adverts

Faisal al Yafai  
Saturday May 14, 2005  
[The Guardian](#)

The small print on the parking ticket said "Don't panic", but for the hundreds of shoppers who returned to their vehicles in a supermarket car park to find the dreaded black-and-yellow penalty notices on their windscreens, reading the small print was not their first reaction.

Disbelief at getting a ticket in a car park slowly turned to relief as shoppers who read on discovered the tickets were fake and simply an advert for a south London estate agent carrying the slogan "If we've managed to get your attention, you should see how good we are at finding tenants and purchasers".

Needless to say, the agency in Bermondsey took a lot of calls, but for all the wrong reasons. After the relief had passed many of the shoppers who found the fake £40 tickets on their cars last weekend became angry.

"We've had phone calls from people actually complaining. People said they don't think it's legal. They've had panic attacks, they've said it's not the way to get attention," said Sarah Jones, manager of estate agents Yeah!. About 30 people called to complain, one even saying they would spread the word not to use her agency.

Such was the anger generated that the agency - for which this was the first attempt at a marketing campaign - took the step of publicly apologising through the local newspaper, the Southwark Press.

Ms Jones said her agency is surrounded by prestigious developments with locked gates and security. Denied the ability to post leaflets through letterboxes, Yeah! had to come up with a more novel way of attracting attention.

She told the Southwark Press: "If a leaflet goes through the door, most people chuck it in the bin. We thought this was going to work. We wanted a hard-hitting campaign to get us noticed, because we are up a side street. We pride ourselves on being radical ... but perhaps this time we went a little too far."

Ms Jones said she was undecided about the wrath of those who took the trouble to call her up. "Perhaps they were overreacting. I just think it's the stigma attached to parking tickets and parking attendants."

The agency now has 8,000 adverts that look like parking tickets sitting unused in its offices. Yesterday it said it had still not decided what its new advertising campaign would involve.

Guardian Unlimited © Guardian Newspapers Limited 2005

Please rate the difficulty of the text.

too easy	somewhat easy	just right	somewhat difficult	too difficult
----------	---------------	------------	--------------------	---------------

Figure 7.5: A selected text together with the ratings at the bottom of the page.

# Chapter 8

## Evaluation

### 8.1 Considerations

One important feature of the system is that the program ascertains a student's ability level and a level of difficulty of a text in order to provide the student with texts of suitable difficulty.

To evaluate the quality of matching a user's ability score against a difficulty score of a text, we use both the rankings of the teacher's judgement of a group of students and the ranking of ability scores from texts written by those students and computed a spearman rank correlation coefficient.

The experiment could have been carried out either just using the students' texts or having the students themselves use the system.

To actually carry out experiments involving users being present is more time-consuming for users and the developer both in terms of preparation and carrying out the experiment. However it may generate valuable insights on whether and under what conditions learners and teachers would use the application. Since the system would be futile if nobody would use it and thus knowledge about its potential use is fundamental, we decided on introducing users (opinions) into the process of evaluation.

In order to gain insight into how much approval such a software meets, the students shall be asked by their teacher to fill in an electronic questionnaire, after taking a good look at the application. Another important outcome is that texts of a desired topic are found. To check if the users think that their information needs are met a relevant question is inserted in the questionnaire.

An electronic questionnaire is the fastest way to obtain the data (opposed to having to sent it by post) and can be directly linked to other information about the user.

However there are other advantages to it, too. The data analysis can be automatised to a certain extent and there are no problems deciphering the handwriting. Moreover the teacher is slightly exonerated as he does not need to print and distribute the questionnaires in addition to the anyhow above-average amount of planning and organization due to the experiment.

Either a group or groups of learners of English from a summer school in Edinburgh or a school with English language classes in Germany were considered as subjects for the experiment.

School classes are presumably preferable to summer school classes because of greater homogeneity in their background (age, native language, knowledge of facts, world knowledge and to some extent even interests) and language ability so that factors that bias the results can be better controlled for. That is also the reason for just choosing the Gymnasium of the three main school types of the German “Gymnasium”, “Realschule”, and “Hauptschule”. Moreover teachers in schools have usually known their pupils for several years and are therefore likely to give better estimates of their ability. That is why we recruited subjects from German schools.

Pupils can only read and rate one or maybe two texts per experimental session. That is only enough to get an impression of how students perceive the difficulty of texts, but not enough to fine-tune the constant that makes up for the difference between reading ability and writing ability. For reliable and convincing results a long-term evaluation is needed, which cannot be part of this study.

It is desirable to carry out many more experiments. It must be kept in mind however that there is only limited time for recruiting subjects and the data analysis and it is not sensible to carry out experiments that will eventually not be analysed or written about.

## 8.2 Experimental Setup

First the recruitment of subjects from a school/schools in Germany had to be organized. A first choice of schools to contact was made by looking through webpages. Especially [www.lernnetz-sh.de](http://www.lernnetz-sh.de) was useful in that respect with a list of projects to make out active schools that seemed likely to be willing to participate. Due to the summer holidays in Germany there was only a narrow time-frame available at the end or the beginning of the school year depending on the federal state that had to be fit into the time-plan of this study and therefore response was low.

The teacher Thomas Rau from the Graf-Rasso-Gymnasium in Fürstenfeldbruck,

Bavaria, Germany kindly agreed to carry out the experiment in his class.

Some other teachers also showed interest and some negotiations were made, but finally no evaluation could be carried out with those.

First the course of the experiment was coordinated with the teacher by email and telephone. The setup of the actual experiment in principle was as follows with me being available by telephone and email throughout the experiment: All students were assigned a personal user name and password from a list of access data that was sent to the teacher beforehand. All students wrote a new text in class. The texts of the students were on the same topic to guarantee a minimum of comparability. They individually uploaded the texts and obtained an ability score. They could then type in search terms, whereupon they were supplied with a list of texts, of which they should at least select one for reading. After reading the text they could rate the text on a scale of 1 to 5 as to whether it was of appropriate difficulty or not (cf. user modelling - update). The students were then asked to fill in an electronic questionnaire (see figure 8.1), with questions as to whether typing in queries helped to find interesting texts, whether the operation of the software was straightforward and if they would consider using it, to name just three examples (cf. section 8.2). The questionnaire was in German, thus no problems with understanding the questions and giving the answers could arise.

The teacher sent ratings and a decent amount of “post-paration” had to be done: additional questions were settled (assignment of texts to students, texts not written by the pupils themselves were “detected”). The students had problems to save documents as *.txt* as they were used to just using *MSword*.

## 8.3 Data Analysis

### 8.3.1 Ability Scores

The students were ranked according to the expert judgement and according to the ability scores. After that a spearman rank correlation coefficient was computed.

In addition to that the ranking between the teacher and another human rater was 0.6 with a t-value of 3.602 which is only slightly better than the values produced by our latest formula made up of three variables.

That the teacher judges the student’s ability and not their texts circumvents that the teacher might be biased by his view on the student when judging a text. The performance of a student, in this case a text written by him, does not necessarily perfectly

Table 8.1: *Ranking Results*

The first column indicates the number of variables in the formula and the version. The second column is the correlation between the teacher ranking and the ranking the formula produced.

formula version	teacher - score	t-value
2.1	0.3478	1.574
2.2	0.2429	1.062
3.1	0.6001	3.182
3.2	0.6160	3.318

reflect his actual ability. Therefore some texts' ranks may deviate from the teacher's ranking due to the student performing better and more likely worse than usually. Thus the correlation might be even higher.

### 8.3.2 Questionnaire - Use of Application

Table 8.2: *Summary of Questionnaire Evaluation*

	yes	no
easy to use	12	3
interesting text	9	6
internet access at home	13	2
general use of electronic media	7	8
use system for autonomous learning	11	4
	paper	indifferent
screen or paper	8	6

The answers to the most important questions of the question are presented. 15 questionnaires were filled in. The results of the evaluation are promising:

12 of 15 learners are happy with the ease of use of the textfinder, one of which even finds it easy to use. 9 users could find an interesting text and 6 could not. With regard to only financial times texts being available at the time of evaluation that is a reasonably good result. As the users are between 15 and 16 years old they are likely to be more satisfied when other newspapers are available too. 7 of the learners use electronic media in general and 8 do not. In view of that it is extremely encouraging

that 11 of 15 people want to use a system as ours for autonomous learning. 8 learners prefer paper to the screen and 6 are indifferent<sup>1</sup>. We expected a clearer preference for reading from paper. That some people are indifferent may open up a more frequent use of the system which allows for more accurate estimates of user ability.

## 8.4 Further Work

We decided on a small number of experiments only because of the time constraints. Evidently, to carry out a cycle of implementation and expanded and alternative experiments would be useful. We give an overview of some such expansions and alternatives in this section.

An additional experiment would be to compare the program's and the teacher's efficiency in selecting appropriate texts for several students from a pre-defined corpus to test if the program could help teacher's lesson preparation. This would yield valuable information as to whether the program or the teacher is better at selecting appropriate texts and which is faster. After reading both texts, the students would have to judge by a number of criteria how well the texts are suited. That should be done by blind testing. When the students are not told whether a text was selected by the teacher or by the program respectively, a personal bias is out of question. Yet, judgements might be influenced by a lot of factors and it is doubtful if objective/reliable results apart from the duration of the selection process are possible.

When it has been established that the program will in fact be used, a large-scale evaluation with hundreds of students' ratings is desirable for improving the system, for example in order to fine-tune weightings (the constant that makes up for the difference).

The same applies to validating the formula on the large-scale. A conceivable approach is to rank students by the scores of assessed writing ability in an official language test and assign another rank to the text written in that test and then calculate the pearson rank correlation coefficient as described earlier. The computer-based IELTS for example is particularly suited to this purpose as students' texts are already available in electronic format. The coarse-grained scale of the IELTS with band scores from 1 to 9<sup>2</sup>, however is quite different from our continuous scale. It would only allow to test

---

<sup>1</sup>one answer is missing

<sup>2</sup>Band scores are reported in half bands for reading and listening and whole bands only for writing and speaking [IELTS, 2004].

our formula with respect to the coarse organisation of the IELTS scale but not if it is good for more fine-grained allocations of ability values. The advantage of standardised tests is that results will be more reliable both due to the standards of assessment but also because there is much more data to base the study on than when carrying out the experiment with only 20 or 30 students of language classes.

Fragebogen: A Web-based Personalised Text Finder for Language Learners	
Many thanks for your participation!	
All the information you provide is on a voluntary and confidential basis. Your answers will only be used for scientific analysis and improving the text-finder.	
• Age:	<input type="text"/>
• What is your native language?	<input type="text"/>
• Do you speak any languages other than German and English?	<input type="text"/>
• How easy did you find the use of the text-finder?	<input type="radio"/> easy <input type="radio"/> ok <input type="radio"/> difficult
• Did the search engine come up with one or several results that you found interesting?	<input type="radio"/> yes <input type="radio"/> no
• Do you learn English outside school using your own study aids?	<input type="radio"/> yes <input type="radio"/> no
• Do you have internet access at home?	<input type="radio"/> yes <input type="radio"/> no
• Do you use electronic media for language learning?	<input type="radio"/> yes <input type="radio"/> no
• If so, which?	<input type="text"/>
• Would you consider using software such as this for autonomous learning?	<input type="radio"/> yes <input type="radio"/> no
• If not, why?	<input type="text"/>
• Do you prefer to read off a computer screen or from paper?	<input type="radio"/> yes, I prefer the screen <input type="radio"/> indifferent <input type="radio"/> prefer paper
• Do you have any suggestions for improving the text-finder?	<input type="text"/>
<input type="submit" value="submit"/>	

Figure 8.1: The English translation of the German questionnaire that was handed out to the students.

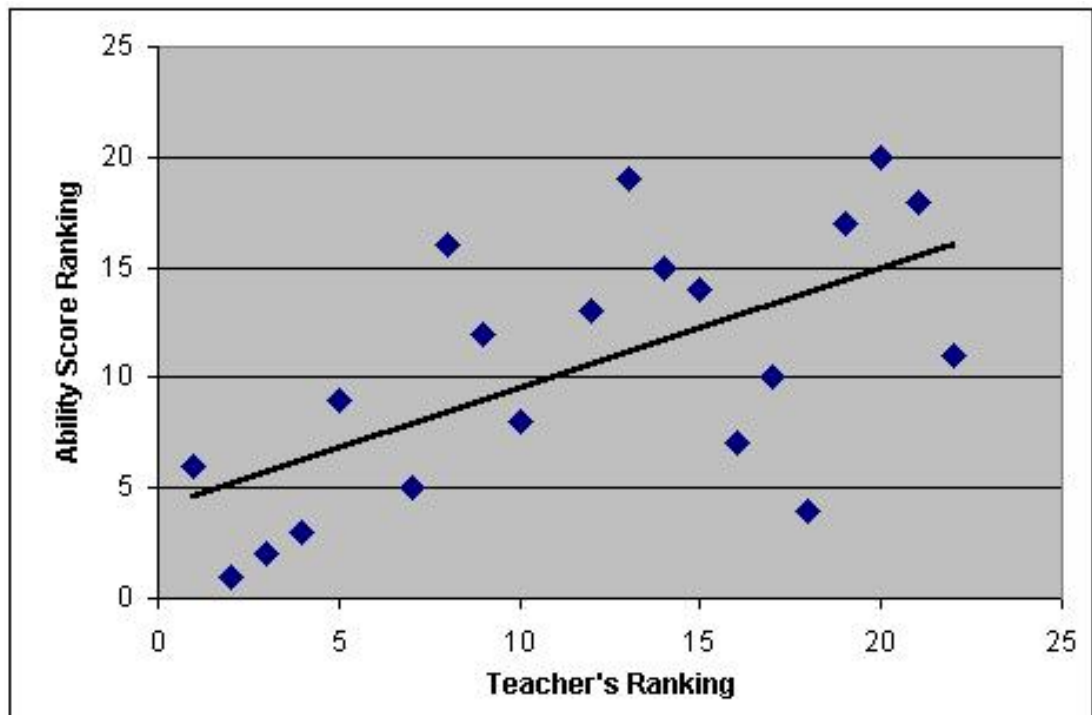


Figure 8.2: Teacher versus Formula Ranking

# Chapter 9

## Conclusion

### 9.1 Summary

We provided the theoretical framework for a system that helps adult foreign language learners find texts adjusted to their individual needs. In order to achieve that goal we specified how to assess students reading skills and how to relate that to finding texts of appropriate difficulty. We specifically designed a formula for calculating both reading ability and readability. We explained how the retrieval of texts from our specially created newspaper corpus works and what the retrieval could be like in later versions. Ways of updating the user profile were presented that ensure that the user's needs will still be satisfied when his ability evolves. An overview of the program's architecture and functioning was given. Finally we carried out an evaluation which provides us with clues as to how good certain components of the text-finder are and if learners find such a system useful.

### 9.2 Strengths

One of the strengths of this study is that we designed a formula directly adjusted to adult foreign language learners. As far as we are aware this has not been done before. It is of importance as adult foreign language learners may have difficulty in finding interesting learning material as authentic material directly targeted to them is sparse - if there is any at all - unlike for native speaker children. In the case of assessing students' ability the formula correlates as high with the teacher's judgement as the judgement of another human, which is a satisfying result. Another positive outcome of this study is that we can offer a working system for language learners that is reasonably easy to use

and provides - even only feeded with a relatively small amount of newspaper articles - interesting texts according to the user survey. It also adjusts the user's ability score as his reading ability evolves.

### 9.3 Weaknesses and Further Work

Although we gained some pleasant results there is certainly room for improvement. We could incorporate more variables into the formula to cover a more exact image of the concepts to assess. Besides that, it would be very interesting to see through further evaluation whether the formula is in fact more appropriate for adult language learners than other formulae not adjusted to them. We also need to evaluate the formula on reading texts. Large-scale long-term studies would be a good way to fine-tune parameters. This also applies to updating the ability score and finding a reasonable constant.

We would also have liked to incorporate more information about the user into the retrieval process. Another interesting area to pursue is topic familiarity.

A general problem with automatically assessing ability, readability, topic familiarity and similar concepts is that we only can account for a small number of features of these constructs but not gain a holistic insight which a human is more likely to have. We will also always face the problem that ever more data is needed to obtain yet better results.

We are interested in using different textbook series as the basis for the formula. We could also imagine to reorder single texts from the New Headway series or other textbooks to get closer to the texts being in pure ascending order, which would improve the accuracy of our formula. Another interesting area for further work will be to test if the same variables are equally good for predicting the readability of texts for language learners in other languages.

It is desirable to extend the text-finder to a bigger learning platform that also offers vocabulary and/or grammar exercises on the basis of the texts. Further instructions on how to read texts extensively or intensively could be provided. Finally collaboration with other users and collaborative filtering would open up new opportunities.

# Bibliography

- [Baeza-Yates and Ribeiro-Neto, 1999] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, New York, NY, 1999.
- [Belkin et al., 2004] N.J. Belkin, D. Kelly, H.-J. Lee, Y.-L. Li, G. Muresan, M.-C. Tang, X.-J. Yuan and X.-M. Zhang. Rutgers' HARD and Web Interactive Track Experiences at TREC 2003. In *D. Harman & E. Voorhees (Eds.), TREC2003, Proceedings of the Eleventh Text Retrieval Conference*. Washington, D.C.: GPO, 2004.
- [Brammerts, Calvert and Kleppin, 2003] H. Brammerts, M. Calvert and K. Kleppin. Lernberatung, in: Bausch, K.-R./Christ, H./Hüllen, W./Krumm, H.-J. (Hrsg): *Handbuch Fremdsprachenunterricht*, Tübingen: Francke, 344-346, 2003.
- [Brown and Eskenazi, 2004] Jonathan Brown and Maxine Eskenazi. Retrieval of Authentic Documents for Reader-Specific Lexical Practice. In *Proceedings INSTIL 2004*, Venice Italy.
- [Brown and Eskenazi, 2004] Jonathan Brown and Maxine Eskenazi. Abstract of Retrieval of Authentic Documents for Reader-Specific Lexical Practice. [http://www.isca-speech.org/archive/icall2004/iic4\\_006.html](http://www.isca-speech.org/archive/icall2004/iic4_006.html).
- [Budzik and Hammond, 2000] Jay Budzik and Kristian Hammond. User interactions with everyday applications as context for just-in-time information access. *International Conference on Intelligent User Interfaces archive, Proceedings of the 5th international conference on Intelligent user interfaces*, New Orleans, Louisiana, United States, ACM Press New York, NY, USA, pp. 44 - 51, 2000.
- [Collins-Thompson and Callan, 2004] Kevyn Collins-Thompson and Jamie Callan. *Information Retrieval for Language Tutoring: An Overview of the REAP Project*, 2004.

- [Council of Europe, 2001] Council of Europe. A Common European Framework of Reference for Languages: Learning, Teaching, Assessment A General Guide for Users. Strasbourg: Council of Europe. (Document DGIV-EDU-LANG 1), 2001.
- [Cozens, 2004] Simon Cozens. Find What You Want with Plucene. <http://www.perl.com/pub/a/2004/02/19/plucene.html>, February 19th, 2004.
- [Doyle, 2001] Matt Doyle. Uploading Files Using CGI and Perl CGI. <http://www.sitepoint.com/article/uploading-files-cgi-perl>, July 27th, 2001.
- [Dumais et. al., 2003] Susan Dumais, Thorsten Joachims, Krishna Bharat and Andreas Weigend. SIGIR 2003 workshop report: implicit measures of user interests and preferences, ACM SIGIR Forum archive, Volume 37, Issue 2, Fall 2003, pp. 50 - 54, ACM Press, New York, NY, USA, 2003.
- [Ehlers, 2003] Swantje Ehlers. Übungen zum Leseverstehen, in: Bausch, K.-R./Christ, H./Hüllen, W./Krumm, H.-J. (Hrsg): Handbuch Fremdsprachenunterricht, Tübingen: Francke, 213-217, 2003.
- [Ghadirian, 2002] Sina Ghadirian. Providing Controlled Exposure to Target Vocabulary through the Screening and Arranging of Texts. *Language Learning & Technology*, January 2002, Volume 6, Num. 1, pp. 147-164, 2002. <http://llt.msu.edu/vol6num1/pdf/ghadirian.pdf>
- [Gospodnetic, Hatcher, 2005] Otis Gospodnetic and Erik Hatcher. *Lucene in Action - A guide to the Java search engine*. Manning Publications, Greenwich, CT, USA, 2005.
- [IELTS, 2004] International English Language Testing System, Score processing, reporting and interpretation. <http://www.ielts.org/teachersandresearchers/scoreprocessingreportingandinterpretation/default.aspx>, 2004.
- [Informationsdienst Wissenschaft, 1997] Informationsdienst Wissenschaft. Neue Wege zur Mehrsprachigkeit. <http://idw-online.de/pages/de/news3999.html>, January 8th, 1997.
- [Jafarpur, 1995] Abdoljavad Jafarpur. Is C-testing superior to cloze? In *Language Testing*, 12, pp. 194-216, 1995.

- [Kelly and Teevan, 2003] Diane Kelly and Jaime Teevan. Implicit feedback for inferring user preference: a bibliography. *ACM SIGIR Forum archive*, Volume 37 , Issue 2, Fall 2003, ACM Press New York, NY, USA, pp. 18-28, 2003.
- [Kelly and Cool, 2002] D. Kelly and C. Cool. The effects of topic familiarity on information search behavior. In *Proceedings of the Second ACM/IEEE Joint Conference on Digital Libraries (JCDL '02)*, USA 74-75, 2002.
- [Liu et al., 2004] Xiaoyoung Liu, W. Bruce Croft, Paul Oh and David Hart. Automatic Recognition of Reading Levels from User Queries. *Annual Proceedings of the 27th annual international conference on Research and development in information retrieval*, Sheffield, United Kingdom. pp. 548-549, 2004.
- [McNamara and the CSEP lab, 2003] D. S. McNamara and the CSEP lab. Coh-Metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension. Annual project report submitted to the Institute of Education Sciences, 2003.
- [MSc Project Guide, 2004-2005] School of Informatics,  
<http://www.inf.ed.ac.uk/teaching/courses/diss/guide.html>, 2004.
- [Nagy, 1988] William E. Nagy. Teaching vocabulary to improve reading comprehension. Newark, DE: International Reading Association, 1988.
- [Si and Callan, 2001] Luo Si and Jamie Callan. A Statistical Model for Scientific Readability. *Proceedings of the tenth international conference on Information and knowledge management*, Atlanta, Georgia, USA. pp. 574-576, 2001.
- [Soars and Soars, 2003] Liz Soars and John Soars. New Headway English Course (six volumes, Beginner to Advanced). Oxford University Press, Oxford, UK, 2003.
- [Sugiyama et al., 2004] Kazunari Sugiyama, Kenji Hatano, Masatoshi Yoshikawa. Adaptive web search based on user profile constructed without any effort from users. *International World Wide Web Conference archive Proceedings of the 13th international conference on World Wide Web*, ACM Press New York, NY, USA, pp. 675-684, 2004.
- [Uitenbogerd, 2003] Alexandra L. Uitenbogerd. Using the Web as a Source of Graded Reading Material for Language Acquisition. *Lecture Notes in Computer Science*, pp. 423-432, Vol. 2783, September 2003.

[Waring, 2000] R. Waring and S. Takahashi. The Oxford University Press guide to the 'why and 'how' of using graded readers. Oxford University Press, Tokyo, 2000.

[Wikipedia, 2005] Wikipedia. Last modified, 19th of August, 2005. [http://en.wikipedia.org/wiki/Principle\\_of\\_compositionality](http://en.wikipedia.org/wiki/Principle_of_compositionality).