A School Text Book Analysis

George R S Weir and Garry Doherty
Department of Computer and Information Sciences
University of Strathclyde
Glasgow G1 1XH

Abstract
This paper reports on a project to digitise and analyse sample reading books recently used within British Primary schools. Through analysis of the textual content of example texts from this corpus, we aimed to illustrate statistical characteristics of these texts and consider implications for the expected rate of progression across texts intended for different school levels. Our project also includes the design and implementation of software tools for assisting with this undertaking. Our approach to textual analysis relies heavily on frequency lists for each of the school text books and their comparison with frequency lists of common usage (words per million) derived from the British National Corpus. Our software toolset eases the task of contrastive analysis as well as providing helpful graphical display of analysis data.

1. Introduction

The 'Ginn Reading 360 Series' is a set of books intended for use in British Primary schools as reading texts. The name 'Reading 360' reflects the all-round approach it offers to reading and language learning and this series aims to provide a comprehensive development from pre-reading to age 13, with an integrated course in language and study skills. The series of books is produced for 13 school levels, becoming progressively harder through each level. The main aims of the different levels are:

Levels 0, 1 and 2 lay the foundation for children's formal introduction to reading and language. They also seek to establish important ideas that can be shared and that reading is fun. Levels 3 and 4 provide children with the opportunity to move gently into a more structured approach to reading and other reading and language skills. Consequently, these books are written with particular attention to story content and narrative quality. Levels 5 and 6 include books with significant informative material, modern writing of high adventure and good fun, as well as a broad selection from traditional literature. These texts aim to progress children onto books of a high literary quality. Levels 7 and 8 provide challenging and varied material aiming to encourage children to read more widely and to understand more about the books they have read.

After completion of levels 7 and 8, the objectives of reading change. While the main purpose of levels 1-8 of 'Reading 360' is providing a solid foundation, levels 9-13 allow children to further develop their reading and language skills by reading books which are illustrated in a great variety of media and styles. In levels 9-13, the concept of 'reading age' places a greater emphasis on children's development and 'interest age' rather than the child's actual age. In terms of text difficulty, the books of level 9-13 aim to cover a wider span than levels 1-8.

2. Approach

Since our time and resources were limited, we opted to digitise and analyse a selection of the available Reading 360 books. Ten texts were selected for this purpose and we chose only to deal with texts from levels 1 -10. Each level included approximately 10 books, with considerable variety within levels on many factors, including text and illustration layout, sentence length, content and theme, text length, and illustrations. Consequently, we undertook to select a book of average difficulty for each level. Following their digitisation and conversion to text editable form, this content provided the target corpora for our remaining analytical work.

Analysing the textual content of each book had three distinct aspects. Firstly, we performed a frequency analysis. Using a reference frequency list derived from the British National Corpus (Leech et al, 2001), we were able to use the commonality of each word as a metric, thereby creating frequency lists for each text as a basis for cross text comparison. Word frequency afforded a gauge of common occurrence for textual content, with words of higher frequency considered more likely to be known and words with lower frequencies less likely to be known to the readers. Secondly, we employed traditional fog indices as a coarse guide to the readability of the sample texts. Thirdly, for each book, we created a manifest to provides a summary of all books in the entire 'Ginn Reading 360 Series' and documents various aspects of each of the books physical dimensions such as number of words per page and the printable area per page.

As noted earlier, our aim was to perform analyses and comparisons across the selected books, with the added facility provided by software tools developed for performing comparisons and displaying results. Figure 1 illustrates the main components in this work.

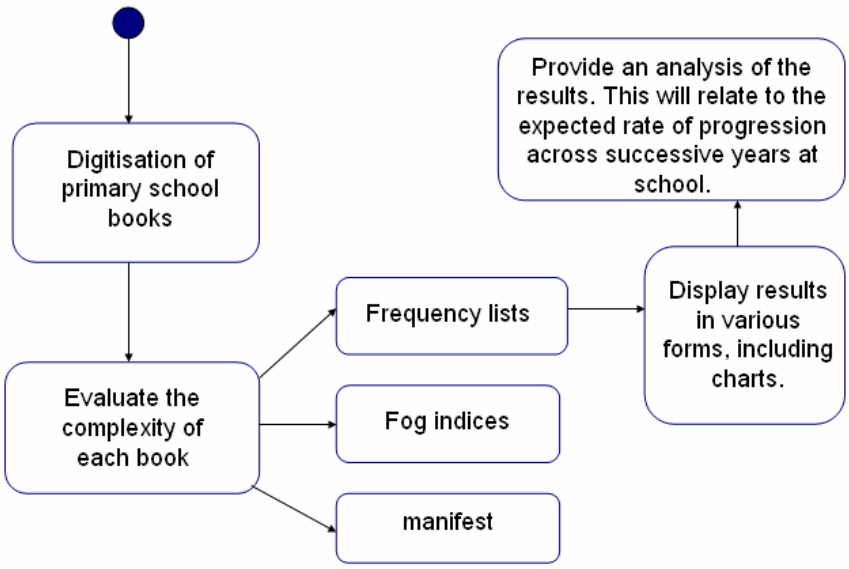

Figure 1: Main Project Components

3. Software tools

An integral feature of our software facility is the display of statistical information derived from the texts under analysis. For instance, frequency tables are available in a variety of chart styles. In addition, the system assists in gauging the results in relation to expected

rate of progression across successive years at school. This is based on a direct comparison of characteristics of the sampled texts. Standard factors, such as fog indices, are included as output characteristics in the software system.

In addition to the data from our textual analysis, we also compare data from the manually constructed manifest for each sample text. This manifest considers aspects such as the available textual area in each book. When considered with the total number of words, this as a basis for deriving a 'text density' factor for each text.

In terms of the frequency lists provided by the system, two separate lists are created for each analysed text. One list gives the results in word alphabetical order, and the other displays the results in the order in which words appear in the source text. The reason for this is that alphabetical ordering of results makes it easier to locate particular words. Displaying the results in an ordered fashion also simplifies the view of where the more complex words lie, e.g., do more common words appear as the book progresses? Are they randomly spread throughout the book? The frequency lists are also used to calculate the rate at which new words are introduced in a particular text (see Table 1).

Rate of New Words for level 2	1.26
Rate of New Words for level 6	1.50
Rate of New Words for level 10	2.16

Table 1: Sample data from the system.

Also included is a count of the number of times words of a particular length appear (between 1 and 13 characters). This is useful as a further means of comparison, in terms of word density.

Words in a text are checked against the BNC reference list to provide a gauge of common usage (words per million). As a result words which have a higher ranking within the BNC (for example words such as '*is*', '*the'* and *'a'*) means they appear more often in standard written and spoken English text. In terms of the primary school textbooks, words which appear more frequently and are generally more familiar words and will tend to have a higher BNC frequency compared to words which appear less frequently.

Word frequency analyses are displayed both alphabetically and in the order in which words appear within the sample texts. This is illustrated in the sample screenshot below (Figure 2).

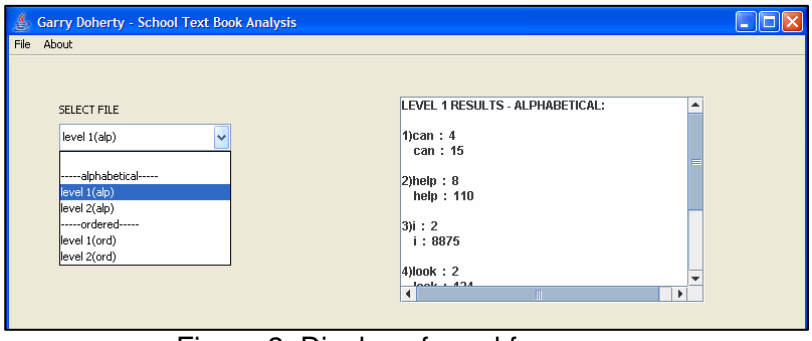

Figure 2: Display of word frequency

After using a number of techniques for analysing the corpora, results are displayed using charts. This allows for easy display and comparison in relation to the other books in the corpus. The software user also has the facility to modify the contents of the reference frequency list. As well as viewing this frequency list, users can search and add words. Another feature of the system is that users are able to view and edit the files used to store the digitised version of each of the primary school textbooks.

Displays of analysis results afford a variety of different types of chart, including line, bar and pie. There is also some flexibility, so that, in some contexts, users can change the charts for better display of the desired features.

Since we aim to ease the task of comparing primary school textbooks, displaying the results in a graphical form is an excellent way of showing the various results collated in different configurations.

On selecting the 'view frequency list' function, the data for the selected book will be displayed in a Table (see Table 2). Three frequency lists are created, based on the text content of the selected file: an alphabetical list, ordered list and British National Corpus list. In addition, a summary is generated of the word density results.

Word Frequency (1)	Word Frequency (2)	BNC Word Frequency	Word Density
LEVEL 1 RESULTS: ALPHABET.	LEVEL 1 RESULTS: ORDERED	FOR LEVEL 1 FILE	
1) can : 4	look : 2	can : 2672	TOTAL NUMBER of words of length '1' is : 2
2) help : 8	i : 2	help : 110	TOTAL NUMBER of words of length '2' is : 2
3) i : 2	can : 4	i : 8875	TOTAL NUMBER of words of length '3' is : 6
4) look : 2	help : 8	look : 124	TOTAL NUMBER of words of length '4' is : 10
5) we : 2	we : 2	we : 3578	TOTAL NUMBER of words of length '5' is : 0
6) you : 2	you : 2	you : 6954	TOTAL NUMBER of words of length '6' is : 0

Table 2: Frequency list from the level 1 textbook

A range of charts is available for displaying the results of file textual content comparison. For example, selecting 'Average Word Length' from the drop down list will display the results gained from performing that method of analysis (Figure 3).

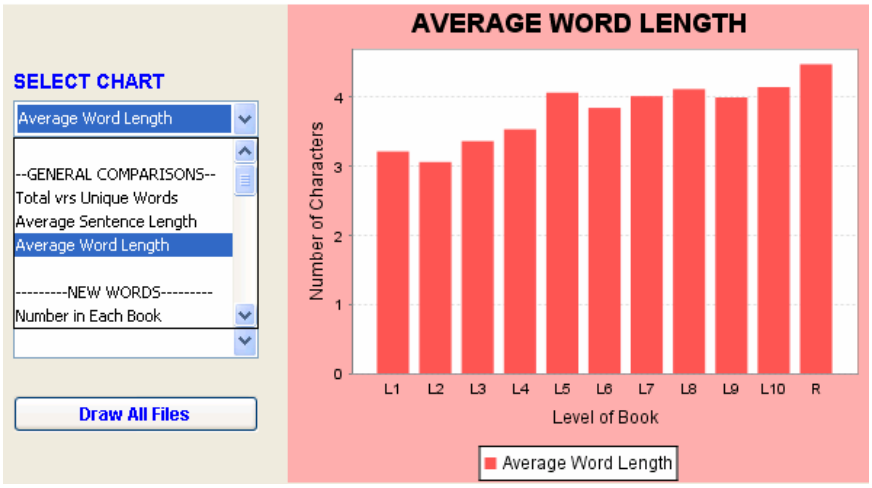

Figure 3: Average Word Length chart

3.1 Readability comparisons

Several attempts have been made to provide a measure of readability of a particular text and in doing so attempt to quantify comprehensibility. This can never be absolute, and only a guide, since people differ in their experiences and skill in reading. In linguistics, a Fog Index is a test designed to estimate the readability of a sample of English text. We chose to adopt this method as another means of comparing the different levels of primary school textbook.

Although there are many varieties of fog indices, we applied four:

- Flesch Reading Ease Score
- Flesch-Kincaid Grade Level
- Coleman-Liau Grade Level
- Gunning fog index

For each of the different 10 levels of book, each of the above fog indices has been calculated and these values provided as data to the software comparison facility.

In addition to detailing the rate at which new words are introduced, charts also identify in which quarter of levels 8, 9 and 10 the majority of the words are introduced. (For example, Figure 4 shows the breakdown of level 9 words.)

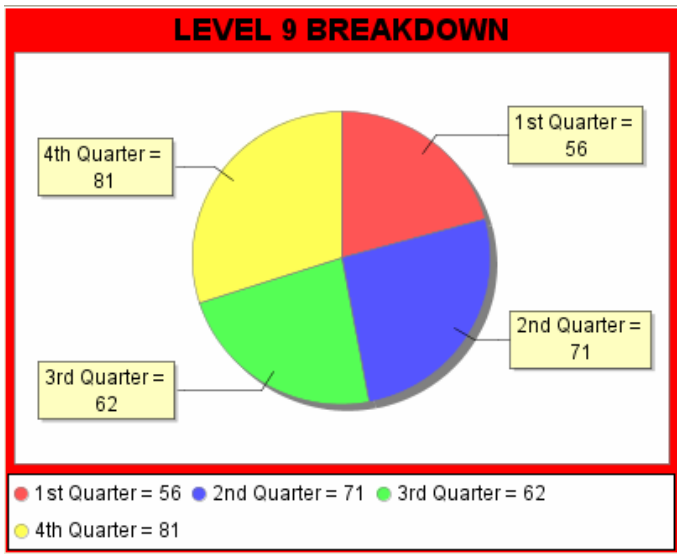

Figure 4: Display of word distribution across quarters of a book

3.2 Text density

The user can view several different charts related to text denmsity, including:

- Text Density per Page.
- Text Density versus Print Area per Page.
- Average text density for all books in each level.

These charts are created from data taken from the manifest for each book. Providing this facility allows the user to analyse the text density of each textbook and relate this to the printable area per page and/or book and observe how this increases or otherwise across the selection of textbooks. Including both digitised and non-digitised books, the average text density for all books in each level is also computed and compared against the text density of the digitised book.

3.3 Word Density

Another area of comparison represented in chart form is word density. This examines how many times the length of words of lengths 1 -13 appear in each file. However unlike the other charts which showed all the files, the user has the ability to specify which files they wish to compare, up to a maximum of 4, as well as comparing all 10 files. Figure 5 illustrates output from a comparison of word density of the level 2, level 5, level 6 and level 8 files. A line chart was selected because it allows the user to easily identify and compare each of the different word lengths.

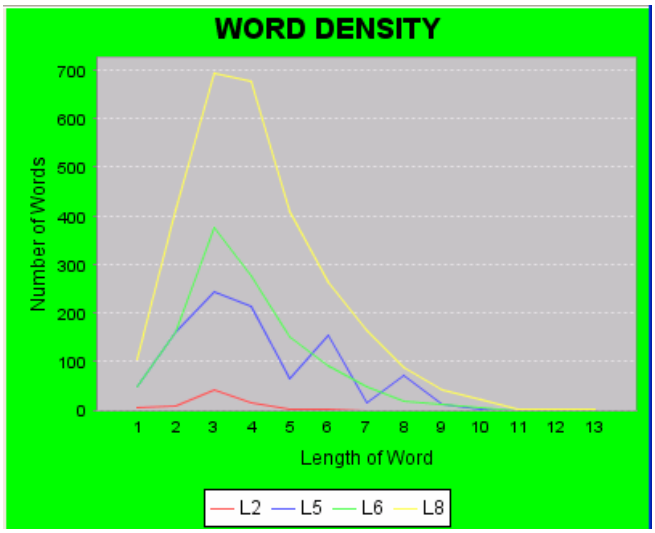

Figure 5: Word Density chart

3.4 Fog indices

Results from the fog index analyses can be displayed as charts and the user may specify which measures to compare for books of their choice.

3.5 Extra Book Analysis

As a further check on system performance, an 'extra' book that was not part of the 'Ginn Reading 360 Series' was digitised and included as part of the data analysed by the software facilities. Charts were created to display some of the data calculated from analysing the book, entitled 'Max's Amazing Summer'. The charts primarily focussed upon the variation of the number of words per page, which includes calculating the minimum, maximum and mean number of words per page and mapping how each page compares to the mean.

The alternative system allows users to perform comparisons on their selected files, rather than the 10 pre-determined primary school textbooks. As with the main system the results of the comparisons are displayed in the form of charts

4. Analysis results

From the outset, one might assume that our different forms of textual analysis would reveal a linear progression as the school level of book increases In order to give further contrast to the reading levels of the digitised school books, another book out-with the 'Ginn Reading 360 Series' was also digitised. At the outset, this book was considered to be of a reading level that matched the intended age range fits of the 'Ginn Reading 360 Series'.

Based on total versus unique words, average sentence length and average word length this analysis does generally progress in a linear fashion. Such that vocabulary and sentence length are two significant factors in contributing to the difficulty of a particular text, may explain this trend of progression. In terms of the random text, for both average sentence length and average word length it is rated among the upper levels of books. Based on the 5-

14 curriculum guidelines for Scottish Education the book is the rated as level C, which is the same as levels 9 and 10, so this result matched what was expected.

A linear progression is evident for the number of new words introduced. The main contributing factor to this is that the number of words in each primary school book in the main increases as the level of the book increases. Hence this would lead to the number of new words increasing in a linear fashion. The rate at which new words are introduced also followed a similar trend. A reason why later level books, such as levels 8, 9 and 10 may have a higher ratio is that such books may contain a higher number of unfamiliar words. This theory is supported because these files have more new words being introduced in the final quarter of the book, than in any of the other quarters. Hence, suggesting the individual books progress in difficulty as they progress.

Each of the different fog indices, differ greatly in terms of their measure of readability for each particular text. In that respect the trend for each of the individual levels is non-linear. Furthermore the data gathered from the 'Flesch Reading Grade Level' index also indicated a non-linear pattern over all levels of textbook. The other three fog indices progressed in a linear manner, when the different levels were compared. A possible reason for this is that some measures rely upon number of characters per word, whilst others rely on syllables per word.

Based upon the printable area per page and number of words per page, text density was calculated. Throughout the levels of book, the printable area per book varies and is non-linear. However, as stated before, the higher level of books in general have more words, hence a greater number of words per page. This has a direct influence on making the text density decrease in linear fashion, as the level of primary school book increases.

Common factor through-out much of this section has been the influence of the total number of words per book relating to a linear progression as level of books increases. In terms of word density, this is once more the case. Due to the fact that this measure of textual analysis is only related to the words in each level, progression in a linear model was expected.

In summary, the various measures of corpus analysis identified that the expected rate of progression across successive years at primary school could be mapped in general, for the majority of analytical techniques progressed in a linear fashion. The fact that a linear trend occurred over most the measures, suggests the various measures were successful in there analysis of the books. As not only did they rate they books in the expected order but this was a consistent trend over all the metrics.

References

Leech, G., Rayson, P. and Wilson, A., *Word Frequencies in Written and Spoken English: based on the British National Corpus*, 2001, Longman, London.