Insights and Innovations:
Revelations from the computerisation of text analysis

Rick Duley

North Perth

Western Australia 6006

rickduley@gmail.com

KEYWORDS:

Lexical variation, readability, reading grade, computerisation

ABSTRACT

Readability scales and lexical variation have served for decades to provide insight into the use of language. However, they have been restricted in their usefulness by the labour-intensive nature of the analysis involved. This paper reports on a computerisation of that analysis including the results of experiments conducted during the development of the software. It presents empirical evidence related to long-held theories on readability formulae and on the characteristics of lexical variation revealed by computerised analysis. It concludes by presenting empirically-based techniques for the use of lexical variation as a comparative statistic.

This paper does not presume to comment of the work of the linguistics greats. Although research into that work was fundamental to the construction of the software this paper does not compare or contrast the formulae devised to predict the characteristics of longer lexical samples. This paper presents the facts — what *does* happen to language when sample sizes range into many thousands of words.

INTRODUCTION

Readability scales and lexical variation have served for decades to provide insight into the use of language. However, they have been restricted in their usefulness by the labour-intensive nature of the analysis involved. This labour intensity leads empirical researchers in Linguistics to take small lexical samples from small groups often with limited language skills.

- Wells worked with 32 children between 15 and 42 months old whose parents did not speak English as their first language[1].
- Duin and Graves worked with 80 12year-olds from middle-class backgrounds studying Language Arts in the United States mid-west who were categorised into *'ability groups'* of about nine students (Duin & Graves, 1987, p.216).
- Lauren worked with groups of pupils aged between 10 and 14, usually numbering less than 20, learning Swedish while their native language was Finnish (Laurén, 2002, p.65).
- Some of the work of Durán *et al.* was based on data from *'32 pre-intermediate level learners of English'* (Durán, Malvern, Richards, & Chipere, 2004, p.236).

Consequently, little information is available to support analysis of the lengthier prose output. Manual analysis of longer works for Lexical Variation or using classical formulae

Acknowledgements: The author is greatly indebted to many people for their assistance in the researching and preparation of this paper. They include:

Professor Geoffrey Roy and Ms Jocelyn Armarego of MU-SES for their support in the research;

Marian Kemp of the Teaching and Learning Department of MU for her assistance in the statistical interpretation of the data;

Martin Dowie of BAE Systems (UK) for his development of Ada.Directories.Searches without which *Analyse* would not work.

This paper and its findings are a credit you. -rd

for Readability would be impractical because of the work time involved. It was, therefore, appropriate to develop Software Application support. *Analyse* was conceived as a batch system which processes a large number of files without interaction with a human operator. Pre-editing the large number of files (e.g. to remove tables and lists) would constitute a correspondingly large amount of work, so it was important that the program handled the anomalies that might arise from a lack of pre-editing.

With this in mind, alongside the idea that selected systems had to be suitable for automation[2], a web search for readability assessments was undertaken. Results of that search overwhelmed the concept of selection criteria by returning a vast majority of 'hits' on three assessment formulae: Flesch Reading Ease (FRE), Flesch-Kincaid Grade Level (FKGL), and the Gunning Fog Grade Level (GFGL). Quite clearly, these are the most generally accepted of the wide range of readability tests available and they were included in the project. This is not to suggest that their popularity makes them the correct or only choice; it is simple recognition of the fact that they are de facto standards permitting comparisons to be made. FORCAST, on the other hand, was initially selected because it was created specifically for assessing technical manuals (and seemed, therefore, to be appropriate). Also it provides some counterpoint to the others in that it does not require full sentences for assessment (this means it should not be affected by the presence of tables or numbered/bulleted lists as the others are supposed to be)[3].

Lexical Variation provides an insight into the author's command of the language as distinct from the readability of the text. It compares the number of individual words (*Types*) in the document to the total number of words (*Tokens*). Introducing a new word into the document requires (1) that the author knows the word (has a broad vocabulary) and (2) that the author has the word on recall (has an easy familiarity with the vocabulary). As a raw datum, however, this Type/Token Ratio (TTR) presents problems in interpretation dealt with later in this paper.

ANALYSE AND OTHER SOFTWARE APPLICATIONS

To be of use as a producer of formative assessment data, *Analyse* must produce results that are comparable to those from readily available software that students themselves may use as a guide. Eleven literary works were collected in text format as a test suite and some readily available software packages (MSWord2000 (Spelling Checker), MSWord2000 (Grammar Checker), Grammar Expert Plus 1.5[4] and Text-Stat 3.0[5]) were used (manually) to generate basic data and readability statistics for the suite. Results for word count, sentence count, and syllable count were compared with those generated by *Analyse* (see Figure 1). Some variations were expected. What was less predictable was the effect pre-editing would have (or not have) on the raw data produced and, therefore, on the statistics produced.

Software packages must produce varying results when analysing the same document simply because the collection of the base data is deceptively difficult.

> *"... even if the formulas were identical, the programmers still had to give the program a way to identify and count words, syllables, and sentences. ... Three different ways of estimating syllables will lead to three different grade-level estimates for formulas that rely on a syllable count."* (Hochhauser, 1999, p.23)

In practice, each package goes about the analysis in a slightly different way and the consequent variation in the base data produces variations in the results. For example, in the Sentence Count graph in Figure 1, TextStat 3.0 consistently counted many more sentences than the other packages yet it uses the same sentence terminator set as does *Analyse* (the set [.!?]). This merely highlights the fact that the calculation of readability statistics is not mathematically precise. Not only is there disagreement about the basic data collected but also there are some hundreds of formulae available to assess the readability of a document and their results are not consistent. Mailloux *et al.* (1995) cite one researcher who:

> *"... assessed readability of six* [medical] *self-care instruction pamphlets using seven different formulas and found that some individual scores deviated from the average score by as much as 41%."* (Mailloux, Johnson, Fisher, & Pettibone, 1995, p.222)

Little wonder that commentators recommend:

> *"It is important to keep in mind that the* [Reading Grade Level] *is, at best, a rough measure of the document's readability. The mathematical process for calculating RGL may give the impression of a greater degree of certainty than is warranted."* (Merriman, Ades, & Seffrin, 2002, p.132)

Furthermore, in tracking the academic development of students — the purpose for which *Analyse* was created — it is more important to have a consistent basis on which to compare student to student, cohort to cohort, and the change in one student from year to year than it is to have some supposed clinical accuracy. Therefore, the basic data performance of *Analyse* (as seen in Figure 1) was deemed acceptable.

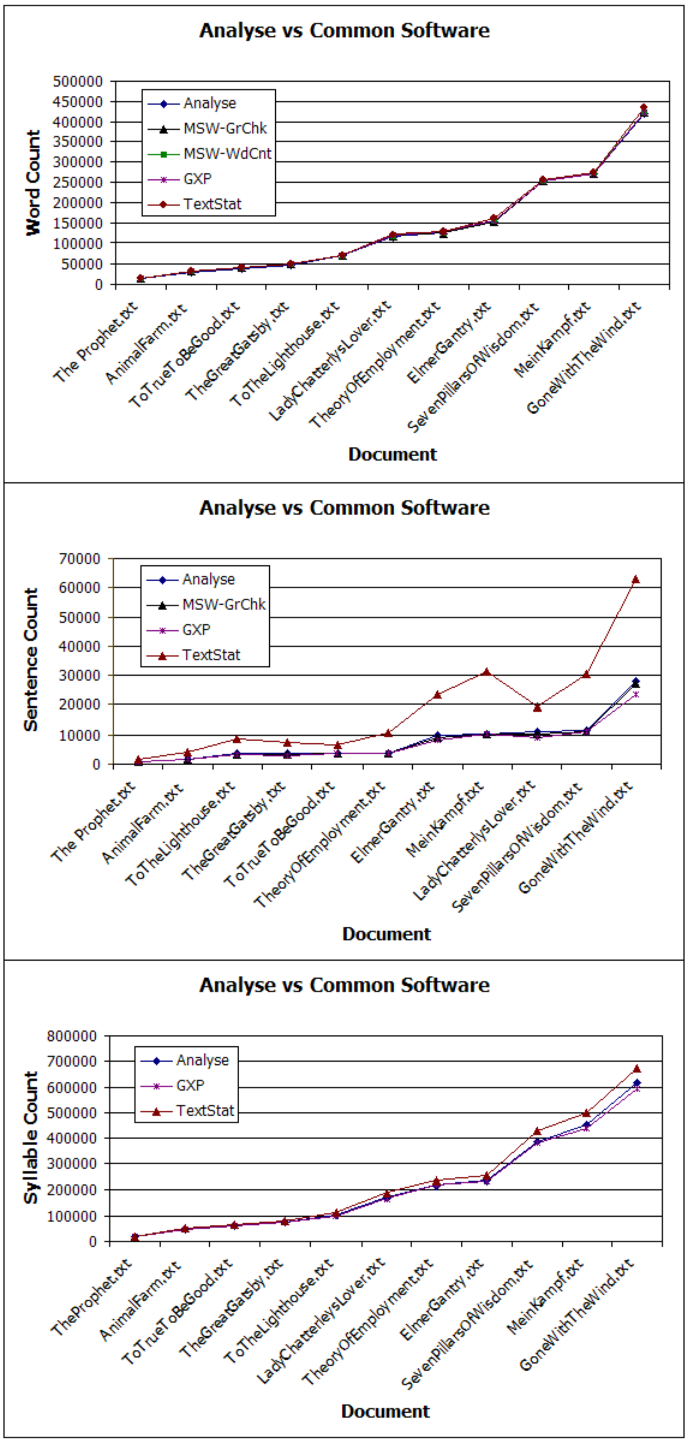

Figure 1: Three Basic Data

FOUR FACETS OF DATA COLLECTION

Three basic data are required to calculate FRE, FKGL, and GFGL. These are the number of words, the number of sentences, and the number of syllables. None of these is as easy for software to distinguish as might at first appear — these is even variation between packages in a count of the number of text lines read from file, something one would expect to be elementary. Again, what defines the end of a sentence? *Analyse* operates on the set [.!?] but packages vary. Even before sentences can be counted, there is the issue of pre-editing to consider. In giving instructions on the use of his formulae Rudolf Flesch, the creator of FRE and FKGL, advised users to:

> *"Skip titles, headings, subheads, section and paragraph numbers, captions, date lines and signature lines. Count the words in your piece of writing. Count as single words contractions, hyphenated words, abbreviations, figures, symbols and their*

> *combinations, e.g., wouldn't, full-length, TV, 17, &, $15, 7%. Count the syllables in your piece of writing. Count the syllables in the words as they are pronounced. Count abbreviations, figures, symbols and their combinations as one-syllable words. If a word has two accepted pronunciations, use the one with fewer syllables."* (Flesch, n.d.)

Johnson gave advice on the counting of numbers (Johnson, 1998, p.4), Klare discusses the problems posed by tables and bulleted lists (Klare, 2000, p.153), and the current ubiquity of e-mail and the Internet present the modern analyst with the issue of URLs and e-mail addresses.

Analyse was to be batch-operated, i.e. would run in the background progressively analysing files without operator intervention. Manually pre-editing large numbers of text files promised as much tedium as analysing them so the question became, *"How little pre-editing is necessary?"* Several experiments were carried out to decide pre-editing issues.

Allowing for Numbers

Mathematical readability presents an entirely different set of problems from the matter of natural language readability — for a start, mathematical prose is a specialised field.

> *"Mathematics, just as any other subject, has its own very specific language in which every word is rigorously defined. For example, a common word, like 'between', when used in geometry obtains a very precise meaning: we say a point C is* ***between*** *points A and B only if all three points are on the same line and AC+CB=AB. Often the words are defined in terms of formulas; this is the nature of mathematics. But at the same time, all formulas have verbal meanings that are analogous to the translation from one language to another and work as a glossary. For example, the well-known formula* $a^2+b^2=c^2$ *has an equivalent translation that can be read as "the sum of the squares on the legs of the right triangle is equal to the square of the hypotenuse".* (Flesher, 2003, p.38)

Secondly, the reader must be able to recognise and interpret a large superset of character symbols and these symbols may not be arranged in simple linear format e.g.

$$\sum_{j=1}^{n} x_j \int(x_j)$$

To compound the problem, the meaning of the symbols may be modified by other symbols and even those modifying symbols may be modified e.g. consider the Central Limit Theorem of the distribution of independent random variables (Spiegel, Schiller, & Srinavasan, 2001, p.56):

$$\lim_{n\to\infty} P\left(a \le \frac{s_n - n\mu}{\sigma\sqrt{n}} \le b \right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-u^2/2} du$$

Computerised evaluation of the number of words and syllables represented by an equation such as this would be difficult. Frequently, and to a much larger degree than in the reading of text, the reader must understand the solution and mentally approximate the answer to be able to interpret the mathematical sentence.

This all leads to the observation that numeric sentences are readable in a sense not applicable to text. To complicate matters further, the concept of a *'numeric word'* in a document also involves dates, percentages, and money. This leaves the question, *"How many natural words, and how many syllables, should be counted for each numeric word?"*

Experiment One:

Analyse was modified to copy numeric words to file. It was run on a range of files resulting in the collection of some 70,200 numeric words. From these, another program was used to randomly select ten samples of 50 numeric words each from that file and these samples were manually parsed for the relevant number of natural words and syllables. Consistent results of collating these figures indicated allowing one natural word and three syllables for each numeric word encountered. *Analyse* was again modified, this time into one form which made the indicated allowances and another which did not. Both forms were run on a test suite of approximately 150 files including literary, professional, and technical writing. This gave results for FRE, FKGL, and GFGL (all heavily dependent on word and syllable counts) where (1) numeric words were counted and allowed for (adj), and (2) where numeric words were ignored (stet). Performing an independent t-Test (assuming equal variances) on these readability scores gave the results shown in Table 1 — the data with adjustments (adj) and the data with no adjustments (stet).

Table 1: Allowance for Numeric Words

t-Test Allowance for Numeric Words		
	Mean	**P(T<=t) two-tail**
FRE-adj	21.8622	0.248263
FRE-stet	19.7232	
FKGL-adj	14.3485	0.568265
FKGL-stet	14.5195	
GFGL-adj	18.8442	0.562582
GFGL-stet	19.0320	

Finding One:

While it appears valid to allow one word and three syllables for the appearance of a numeric word in a document, the P-values in Table 1 indicate that making such allowance makes no significant difference to the resultant readability statistics. *Analyse*, therefore, simply replaces numerals (and numeric operators) with spaces thereby removing them from the readability calculations.

Allowing for Abbreviations

Use of abbreviations has the potential to skew readability statistics severely, firstly because they may represent more than one word, and many syllables (e.g. the abbreviation *'U.S.A.'* represents four words and nine[6] syllables)[7]. There is, however, a second, rather less obvious problem in that abbreviations regularly end with or contain full stops (*'Mr'* and *'Mrs'* do not). Here the problem is not only that of counting words and syllables but also that of counting sentences. Taking the full stop as a sentence delimiter results in the sentence *'This item was made in the U.S.A.'* (one sentence, seven words if the acronym is taken as one word) becomes three sentences with an average of three words each. Also, instead of having seven words and 10 syllables for an average of 1.4 syllables per word there are now nine words and nine syllables for an average of one syllable per word (see Table 2). Fully expanding the abbreviation gives 11 words and 16 syllables at an average of 1.5 syllables per word. It gets complicated. *Analyse* was originally designed to fully expand, and make allowances for, common abbreviations but the question arose, *"What difference does it make to the readability statistics for a document if abbreviations are completely ignored?"*

Table 2: Parsing an Abbreviation

Parsing Sentence with Abbreviation		
Sentence	**Words**	**Syllables**
This item was made in the U.	7	8
S.	1	0
A.	1	1

Experiment Two:

Table 3: Allowance for Abbreviations

t-Test Allowance for Abbreviations		
	Mean	**P(T<=t) two-tail**
FRE-adj	58.6627	0.982251
FRE-stet	58.7168	
FKGL-adj	9.7775	0.985337
FKGL-stet	9.7658	
GFGL-adj	12.6639	0.998103
GFGL-stet	12.6621	

Forty-five well-known literary works totalling over five million words, downloaded from the Internet, were converted to text files. *Analyse* was compiled in two forms, both of which replaced abbreviations with blank characters to protect the sentence breaks and one of which made appropriate adjustments to word and syllable counts while the other did not. Performing an independent t-Test (assuming equal variances) on the readability results for the three common readability formulae (all heavily dependent on word and syllable counts) gave the results shown in Table 3 — where *'stet'* indicates the unadjusted data and *'adj'* the adjusted data.

Finding Two:

Despite the ominous potential for abbreviations to skew readability statistics, the P-values in Table 3 reveal that making allowances for abbreviations in the word and syllable counts makes no significant difference to the resultant readability statistics. *Analyse*, therefore, simply replaces abbreviations with spaces thereby removing them from the readability calculations.

Pre-editing — Dealing with Artificial Text Structures

Table 4: t-Test Results for Pre-editing

t-Test — Pre-editing		
	Mean	**P(T<=t) two-tail**
FRE-stet	58.06	0.93
FRE-edit	58.57	
FKGL-stet	9.81	0.98
FKGL-edit	9.76	
GFGL-stet	12.70	0.99
GFGL-edit	12.65	

Debate rages over the effect on readability statistics of lists, tables, indexes, and other artificial text structures. This paragraph from Klare (2000) gives an insight into the problem:

> *"There is no doubt that readability formulas have sometimes been misapplied or the resulting scores misinterpreted. Redish*[8] *provides an example herself when she says that 'readability formulas will say you have long sentences' when you use bulleted lists. ... Schriver*[9] *argues that you 'doctor the text' when you remove such lists and might as well therefore not use a formula. But bulleted and numbered lists that do not use complete sentences constitute only a minor part of most text; there is no reason therefore why they cannot be omitted in formula applications while evaluating the bulk of the text."* (Klare, 2000, p.153)

Such manual pre-editing of the number of files to be subjected to *Analyse* would be a thankless task and the fact that it would have to be repeated every semester makes the idea most unattractive. There is also the factor of consistency, i.e. would a human editor get all the lists etc. every time? Further to this, algorithmic identification and deletion of artificial text structures presents profound difficulties so the question was, *"Could lists etc. be left in the text without adversely affecting the results produced?"*

Experiment Three:

Two small selections of files, 11 literary texts and 12 technical, were copied into two sets of folders — the selections were small because pre-editing is a tedious and time-consuming process. One set was pre-edited to remove headers, footers, tables of contents, indexes, reference lists, captions, tables, mathematical calculations, and numbered or bulleted lists etc. *Analyse* was run on both sets and the resultant data plotted in Figure 2. Readability scores for the full (stet) and pre-edited (edit) versions were t-Tested and the results are shown in Table 4.

Finding Three:

Counter-intuitively, the P-values in Table 4 show that pre-editing the selected texts made no significant difference to the readability statistics produced. This effect is assumed to be related to the size of the documents involved — in large documents the number of words contained in tables etc. may be a minute percentage of the total word count. Thus, the elimination or otherwise of tables etc. has little effect. However, in smaller documents, or in cases of excessive use of tables, differences will be noted. One student document produced FRE=1.8, FKGL=24.23, and GFGL=27.57. Investigation revealed the document to be composed almost entirely of one table.

Table 5: Regression Results for Pre-editing

Regression Test Pre-editing	
	P-value for X-axis
FRE-stet	0.65
FRE-edit	0.66
FKGL-stet	0.65
FKGL-edit	0.70
GFGL-stet	0.64
GFGL-edit	0.72

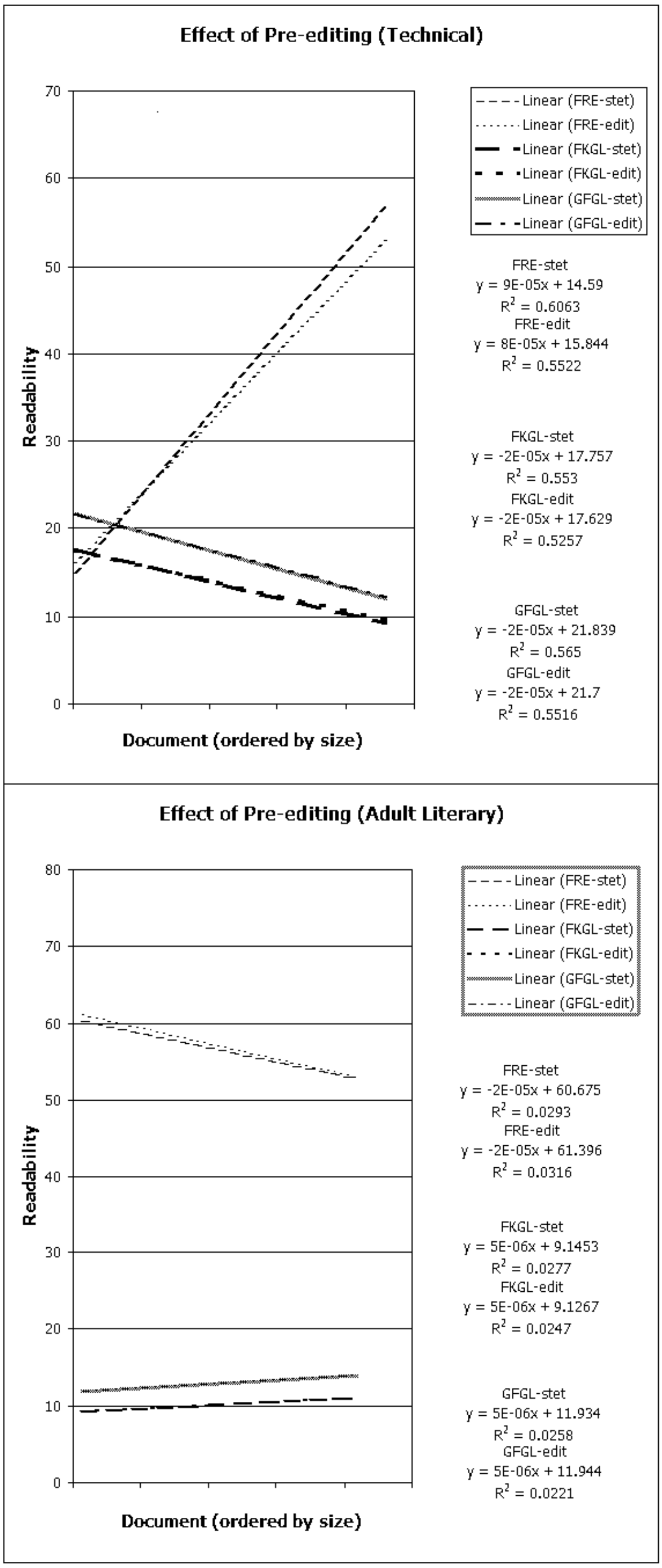

Figure 2: Effect of Pre-editing

Graphical Misconceptions

While the statistical strength of this finding was surprising, the graphical evidence was more so raising questions about a relationship between the readability scores and the length of the documents where none was expected to exist. Both graphs in Figure 2 reveal line slope indicating such dependence. t-Tests were conducted on the slope of the regression lines (Regression Tests) and the results are shown in Table 5. These clearly indicate that the line slopes in Figure 2 are statistically insignificant, but how could this be reconciled to the clearly visible slopes in the graphs? In fact, the line slopes are a deception; the reason for this is hidden in the scale of the X and Y-axes.

While the ranges of the Y-axes are 0 to 80 for the unedited literary samples and 0 to 70 for the unedited technical samples, the ranges of the X-axes are 0 to 418800 and 0 to 559524 respectively. Quite simply, if the same physical scales were used for both axes in each graph the lines would, to all intents and purposes, be horizontal. It is, therefore, assumed that the actual slope shown is a function of the text samples selected and not a characteristic of a relationship between readability scores and word count. This assumption is supported by the presence of both positive and negative slopes.

Finding Four:

Computerisation of readability analysis has enabled the study of text samples of far greater length than was erstwhile possible. This has revealed minor inaccuracies in the constants used in the formulae resulting in a bias that might, on occasion, indicate a relationship between readability score and word count where one does not, in fact, exist. This bias is trivial considering the overall imprecision of the formulae.

INVESTIGATING LEXICAL VARIATION

Lexical Variation provides an insight into the author's use of vocabulary.

> *"Lexical diversity is a term used among interpersonal communication scholars to describe the range of a speaker's vocabulary. As an example, a speaker who only uses the term* 'approach' *throughout a speech is not considered as lexically diverse as one who uses* 'approach', strategy', 'plan', *and* 'program'.*"* (Nelson, 2002)

Lexical Variation (also referred to as Lexical Diversity) is calculated by creating a lexicon of all the words in the document, counting multiple appearances of those words, then establishing the ratio between the number of unique words and the total word count for the document. As with most things in the study of language, lexical variation is one aspect of a discourse and should not be dealt with in isolation. As Nelson (2002) goes on to point out, one can overdo the use of a thesaurus:

> *"... it's possible (and often easy) to take the range of vocabulary too far and appear pretentious, ..."*

This is especially true in professional discourse where terminology is regularly defined with some precision within the discipline. Here the simplistic use of natural language synonyms is inappropriate; the issue is not merely knowledge of the word and its synonyms but also knowledge of the way they should be used in a given context. This, in turn, becomes a matter of choosing from the appropriate words we know. Still, LV is seen as *'a good criterion for the linguistic quality of an essay'*[10].

Types, Tokens and Type/Token Ratios

In a study of the words in a discourse it is useful to determine the ratio between the number of different words (types) in a total number of words (tokens) (Full Report of Research Activities and Results, n.d.). This simple ratio, known as the *'Lexical Variation'*[11] of the document, illustrates both the extent of the author's vocabulary and the facility with which words are chosen from that range — linguistic competence.

> *"Vocabulary is essential for language acquisition and language use. Its extent and quality with regard to variation and the use of content words and form words*[12] *are part of the linguistic competence, which in different studies has proved to be covariant, for instance, with grammatical competence."* (Laurén, 2002)

Type/Token Ratios (TTR) can vary between 1.0 and 0.0. Take, for example, the sentence *"Jesus wept."* (John 11:35 KJV) This has two words in total, two unique words (two tokens, *'N'*), two types, *'U'*), and so has a TTR of

$$TTR = \frac{U}{N} = \frac{2}{2} = 1$$

Equation 1: Upper Limit of TTR

When every word in a sentence is the same, for example the *'Hallelujah Chorus'* from Handel's *'Messiah'*, then we have the general case of the minimum TTR — for any number of types where the number of tokens *'N'* approaches infinity then

$$TTR = \frac{U}{N} \rightarrow 0$$

Equation 2: Lower Limit of TTR

In practice, any TTR will lie between the two extremes but values are not linear with increase in discourse length. In a short discourse, many of the words will only appear once. However, as the length of the discourse increases then the vocabulary of the writer/speaker will eventually be fully utilised and words must be reused.

> *"Adding an extra word to a language sample always increases the token count ... but will only increase the type count ... if the word has not been used before."* (Durán et al., 2004, p.221)

In practice, re-use of common words regularly begins within two sentences. As a consequence the value of the TTR rapidly drops away from 1.0 as discourse length increases but it will never, in normal discourse, reach 0.0 (see Figure 3 which clearly shows the relationship between TTR and the number of words in the document. Some researchers have devised mathematical models to describe this behaviour. *Analyse* can reveal what actually happens.). For TTR to be a useful means of comparison between documents of differing lengths — and, therefore, for comparison student to student and cohort to cohort — some means is required to cancel out the effect of this dependency.

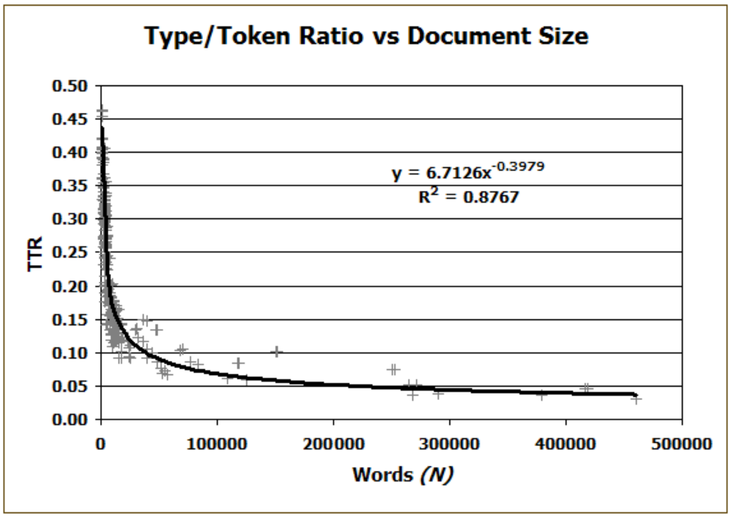

Figure 3: TTR vs Document Size

Computerising TTRs

From the outset, this study differed from most previous published work in this field in that:

- The Token Count (N) is higher — there are more words in each document than the samples of a few tens of words previously studied (written assignments ranging up to ≈5000 words);
- The Type Count (U) is higher — although some of the students have English as a second language (ESL) or as a foreign language (EFL), MU-SES entry requirements mean their English linguistic skills are at a higher level than has often been the case for people taking part in earlier studies;
- The number of text samples is larger.

It is the last of these factors that, in the first place, made computerisation the only feasible option — manual analysis of that amount of text is impractical. It was this computerisation, however, which provided a special insight into the TTR/N relationship.

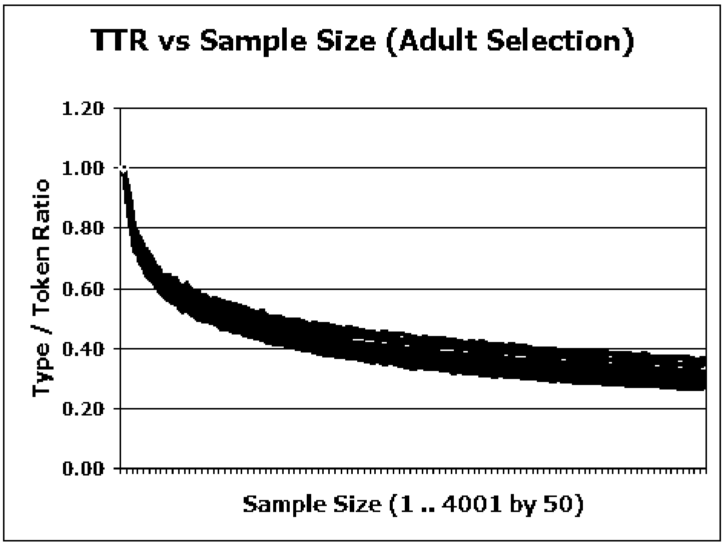

Figure 4: TTR/N for Adult Literature

The TTR/N Curve

Computerising TTR analysis made it possible to derive accurate TTR values for documents of extended length, e.g. in Figure 3, the extreme right-hand coordinate represents TTR for a technical manual for a database management system consisting of 460489 words. Counting that many words manually is a task at which most researchers would baulk — leave alone calculating TTR at the same time! Figure 3, therefore, shows the characteristic TTR/N curve across a more extended range than was possible without computerisation. Figure 3 affirms that:

- There is a dependency between TTR and N;
- A TTR derived for a complete discourse is, therefore, invalid as an indicator of comparison between discourses of differing lengths.

This second affirmation raises the question of the effect of defining a sample size for an evaluation of TTR. Durán et al. cite a test in which this was done:

> *"... *vocd* begins with 35 tokens and plots the first point on the transcript's curve by undertaking 100 trials of randomly sampling 35 tokens from throughout the text without replacement and calculating their average TTR. The number of tokens is then increased to 36 and the calculation repeated, and so on up to 50 tokens. In all, therefore, 16 points are plotted from N=35 to N=50."* (Durán et al., 2004, p.225)

Using data from *Analyse* derived on a similar basis, TTR/N curve coordinates were gathered for examples of Adult Classical Literature. Random samples were taken across the complete texts starting at 1 token and incrementing by 50 tokens at a time to 4001 tokens. Ten samples of each size were taken and an average TTR for each sample size was calculated. These coordinates were plotted as Figure 4 revealing the same general form of the TTR/N curve as in Figure 3. This left the question of whether or not a variation in the position of the band of curves would be discernable in

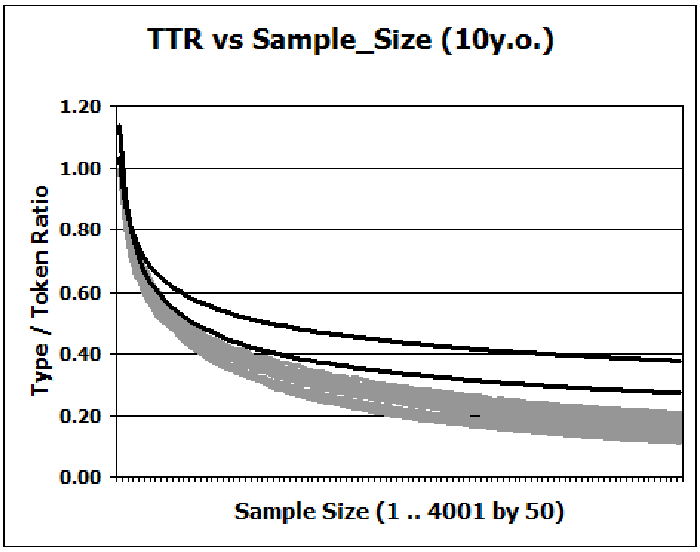

Figure 5: TTR/N for 10y.o.

data from a different selection of texts.

Variation in TTR/N

Analyse was executed on texts written for 10-year-olds to read by themselves. These data were plotted and fit lines for the upper and lower boundaries of the band of curves in Figure 4 superimposed to create Figure 5.

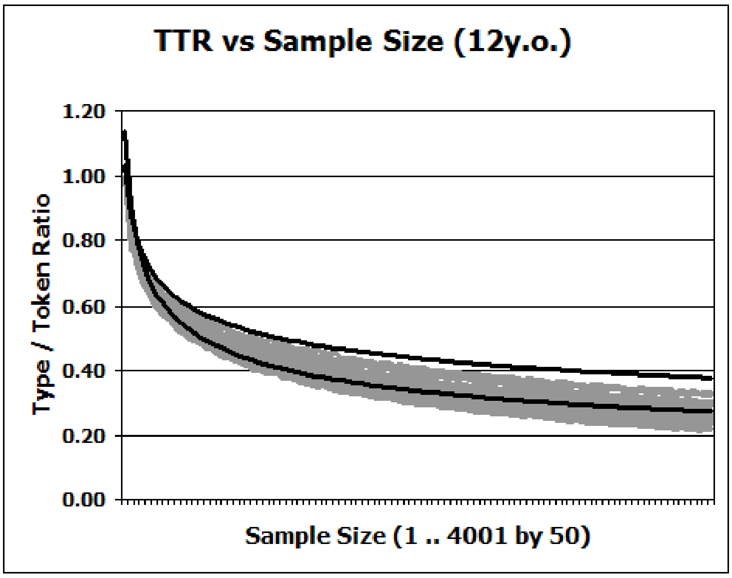

Figure 6: TTR/N for 12y.o.

It was to be expected that 10y.o. children would be comfortable with reading material having a lower TTR than adults — after all, generally their vocabularies would be smaller. As can be seen, the consistency and the visible degree of variation of these results from those of the adult selection support the conclusion that a discernable variation exists. Now the question was, if there is a difference in TTR/N between the reading material enjoyed by adults and reading material enjoyed by children, would that variation show for an intermediate age group.

Repeating the 10y.o. experiment with literature for 12y.o. readers resulted in Figure 6. Here a distinct shift in the span of curves towards the position of the span of curves for the adult literature selection can be seen. Different reading age groups appeared to prefer reading

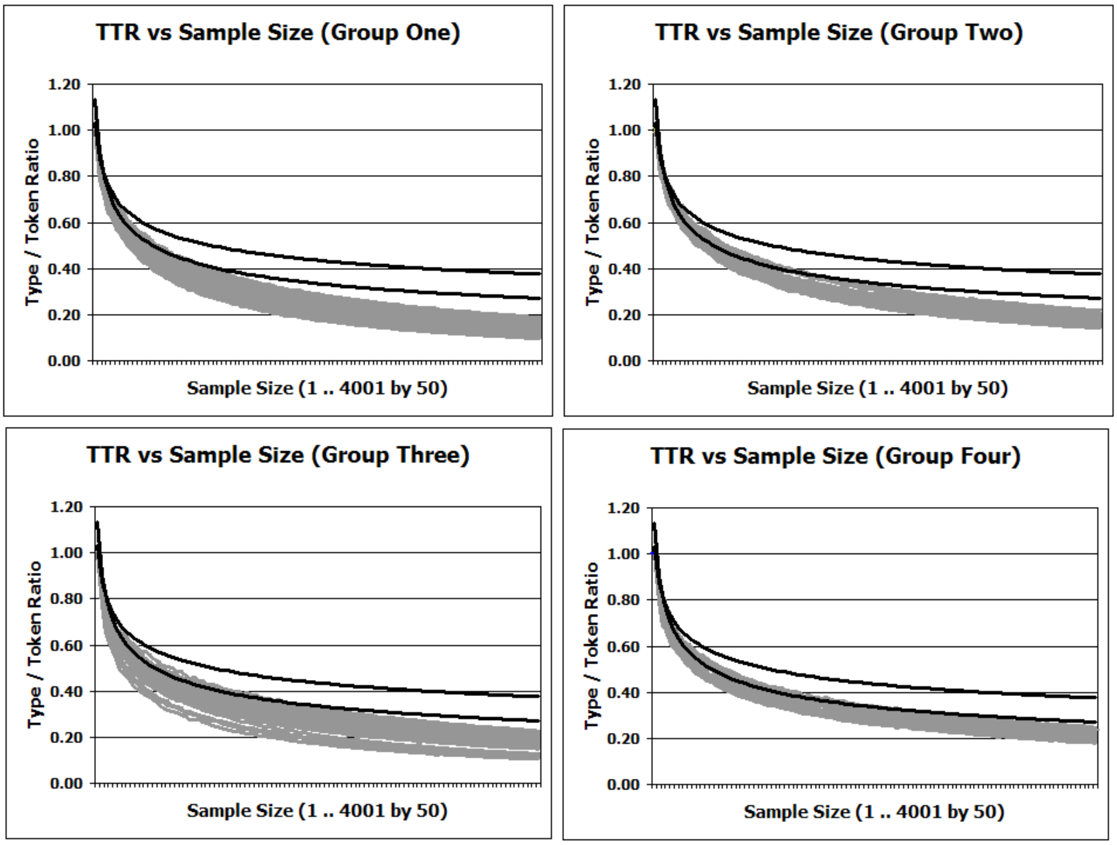

Figure 7: TTR/N for Journal Articles

material with different TTR/N traces. This raised the question, does this hold for other groups.

Articles were collected from four well-known computer industry journals. One might read Groups Three and Four if their focus happened to be one's particular field of interest (or if one had a particularly difficult and erudite problem to solve for the current project); the other two journals one might browse for current developments while sitting in an airliner.

Analyse sampled these document groups in the same way as the 10y.o. and 12y.o. texts and the results are shown in Figure 7. Evaluating these four graphs reveals:

- That each journal group produces a distinct, unique trace lending further support to the assumption made by Durán *et al.* (2004);
- That all of the groups have traces centred lower on the graph than the trace for 12y.o. reading material, noticeably lower than the adult literature and, in fact, as far down as the 10y.o. traces. This surprising finding was put down to the voluntary restriction of the adult vocabulary undertaken by the professional author writing for a professional audience. Professional-speak, jargon, appears to take a toll.

Tracing Linguistic Development

For tracing the change in TTR/N, two methods suggested themselves — (1) mathematical comparison of the fit lines for the TTR/N curves or (2) selecting a common sample size.

Working to the Line

One method of obtaining a single number to represent the relative placement for the curve generated for a particular document would be to calculate the area beneath the curve. For example, if this were done for the two trendline curves used in previous illustrations curves the two numbers would represent an upper and a lower boundary against which a similar figure for an individual document could be compared. There are a variety of other techniques.

One factor working against the use of a method of this nature is the derivation of the coordinate data to produce the curve. Admittedly, student prose artefacts, which *Analyse* was designed to review, are unlikely to vast size but the search is for a realistic and appropriate standard. Secondly, this is a matter where mathematical precision and values precise to *'n'* decimal places overstate the case (that is, are inappropriate). Given that any readability analysis can, at best, only produce approximate values, the question remained, *"Is there a simpler way which produces a sufficiently accurate result?"*

Working with Samples

In tracing any trend there is an obvious need for a standard measurement. Because of the manner in which the trendlines used as a standard were derived, i.e. random sampling, the answer is, "*Yes; the process of comparison can be simplified by random selection of samples."* Each of the samples collected from the document by *Analyse* was selected on a random basis from the full text[13]. Therefore, each sample represents (as well as is possible given the size of the document and the size of the sample) the distribution of words in the document. Therefore, the TTR for any given sample size **n** for any given document may be compared with the range for the same range **n** in any given set of standard boundary trendlines — with two inhibiting factors in the choice of **n**:

- Where **n** is ludicrously small the sample cannot be taken as representative;
- Where **n** exceeds the size of the document the sample it will, while being representative, produce misleading information.

Selecting a Standard Sample Size

In seeking the greatest possible difference between the upper and lower standard trendlines, to provide adequate granularity in comparisons with individual documents, it would be natural to move to the right-hand side of the TTR/N graphs used so far because the lines appear further apart. However, because of the general nature of the curves presented, a vertical section of the curve span close to the Y-axis can provide as much range as one further to the right. Figure 8 presents a simple, graphical illustration. Here the lines at Y=0.00, 0.20, 0.40, 0.60, and 0.80 make equal intercepts on the diagonal. Therefore, vertical lines through these intercepts meet the X-axis at equal intervals. The intersections of these verticals with the upper and lower trendlines are compared in the inset showing minimal difference in range and thus permitting the selection of a smaller sample size..

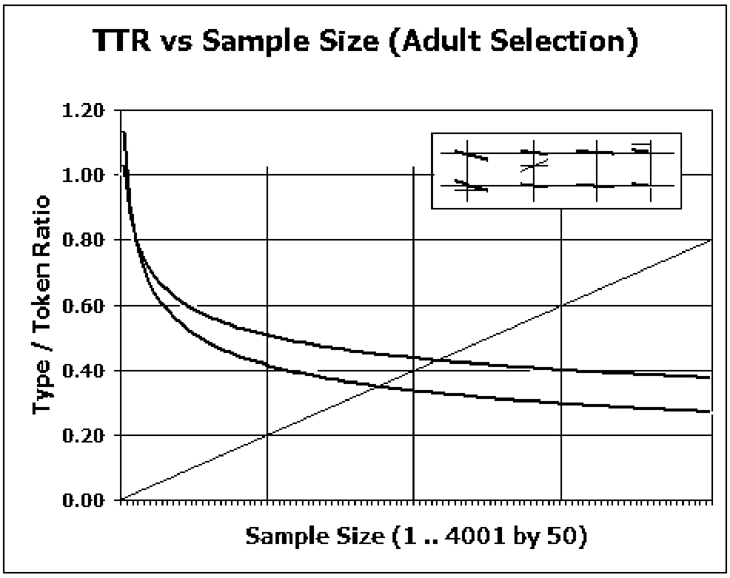

Figure 8: Selecting a Standard Sample Size

Remembering that the purpose of this research is to investigate the work of undergraduate students, sample size must relate to the artefact size expected. Major student prose artefacts at Murdoch University School of Engineering Science (MU-SES) are expected to reach ≈4-5000 words but in some courses, notably programming courses, a majority of the prose artefacts produced by freshman students at MU-SES contain as few as 1000 words. Therefore, to be able to review as many artefacts as possible the sample must be relatively small. Because of this, coupled with the fact that smaller sample sizes result in faster computer processing, *Analyse* takes (by default) 100 random samples of 1000 words each from each document analysed.

Effect of Averaging Lexical Variation

Analyse was set to take the mean of the results from these 100 samples to produce one number representing the TTR/n of the document at n=1000 while producing a second number representing TTR/N where N=(total words in document). Results of an *Analyse* run on a sample of 45 literary documents of over 2000 words are plotted in Figure 9. For the line TTR/n a t-Test gave p=0.204, indicating (p>0.05) no significant slope. Again, the apparent slope is attributable to the problem with scale referred to above. Comparison of the line TTR/n with the line TTR/N shows that the relationship of LV to document size has been eliminated.

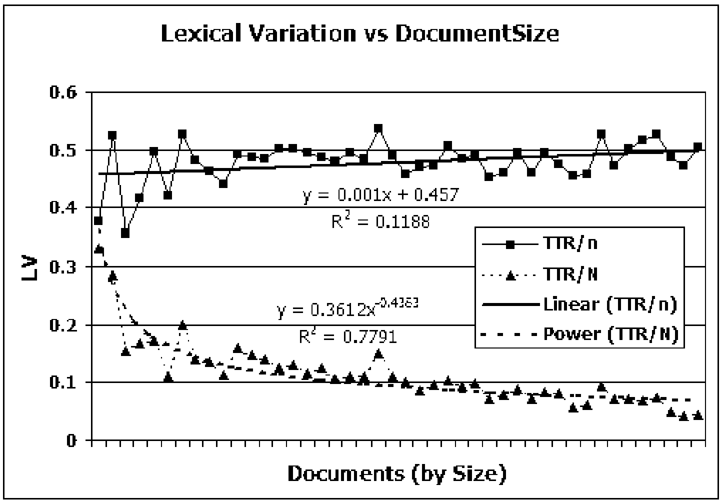

Figure 9: Lexical Variation by Sample and Document

Finding Five:

Multiple random sampling to create an Averaged Lexical Variation (ALV) eliminates the previously proven dependency of LV on document length. ALV scores are, therefore, comparable regardless of the size of the document. This has considerable comparative value. Figure 10 shows ALV data plotted for different demographic groups. Each vertical bar represents the range of ALV scores for its group. These results show visible variation between the work of freshman and senior students. What freshman students write involves considerably less range of vocabulary than the material they read as 12-year-olds.

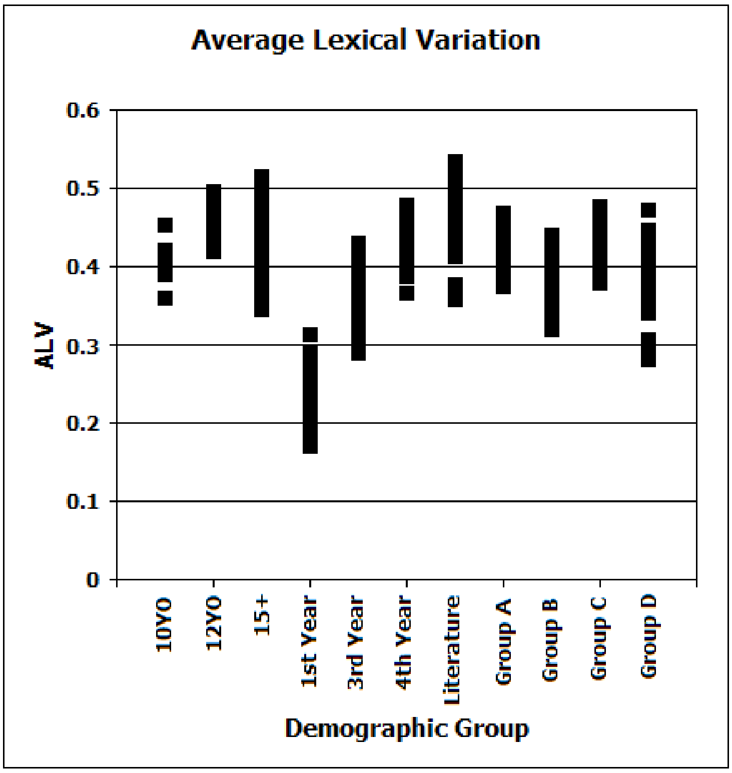

Figure 10: Averaged Lexical Variation by Readership Group

Groups A to D represent material from four widely-read computing journals.

- 33 documents in Group A (a highly technical software practitioners' journal — 2,000..33,000 words);
- 35 documents in Group B (a low-intensity computer industry publication, mostly airline reading — 1,000..5,000 words);
- 30 documents in Group C (a low-intensity software industry publication, mostly light study reading — 1,500..6,500 words);
- 31 documents in Group D (a highly technical hardware practitioners' journal — 3,500..16,000 words).

It is interesting to note that by the Senior year the students were producing work comparable in vocabularic range to the writing of professionals. N.B. This is not to say that the work was of comparable value, only that the vocabularic range was comparable. As with any other readability statistic, ALV is a guide not an absolute indication.

Averaging Readability Scores

Multiple random sampling, it should be noted, does not have the same effect on FRE, FKGL, and GFGL scores. Scores based on random samples were t-Tested against those based on full word count (327 text samples) producing (P(T<=t) two-tail) values of 0.5, 0.5, and 0.5 respectively (hypothesised mean difference = 0). Random sampling, therefore, had no effect on the readability scores obtained. Since the documents ranged in size from 1001 words to 784118 while the sample size (**n**) stayed at 1000, this is further evidence that readability scores are not dependent on document size.

SUMMARY AND CONCLUSIONS

This paper has described a series of experiments undertaken during the development of a lexical analysis software package. Each has provided empirical insights into the nature of lexical analysis hitherto unavailable because of the labour-intensive nature of that analysis:

- Deleting and not counting numbers and numeric operators from a document, and making an allowance for the words and syllables they represent, has no effect on the readability statistics for substantial documents;

- Deleting and not counting abbreviations from a document, making no allowance for words and syllables they represent, has no effect on the readability statistics for substantial documents;
- Pre-editing documents to remove tables, lists etc. has no effect on the readability statistics for substantial documents;
- Any apparent relationship between readability scores and the length of the document is attributable to document selection;
- Multiple random sampling, and the derivation of an Averaged Lexical Variation (ALV) score, eliminates the previously apparent, strong relationship between LV and document length. ALV scores are, therefore, comparable regardless of the size of the document. It should be remembered, however, that ALV data is no more definitive than formulaic readability statistics and, like them, can only find valid application as an indicator in the evaluation of prose.

REFERENCE LIST

Introduction to Content/Function Words, (2000), Published by: University College London (UK). Retrieved November 25, 2004 from http://www.speech.psychol.ucl.ac.uk/training2/intro.html.

Caslon Analytics profile: online readability , (2003), Published by: Caslon Analytics. Retrieved July 14, 2004 from http://www.caslon.com.au/readabilityprofile1.htm.

Function Words: The Columbia Guide to Standard American English, (2004), Published by: Bartleby. Retrieved November 25, 2004 from http://www.bartleby.com/68/67/2667.html.

Reading in the Content Area, (2004), Published by: Washington County Public Schools, MD (USA). Retrieved July 14, 2004 from http://www.alt.wcboe.k12.md.us/mainfold/technolog/instruct/msde07/module14/mod14_activity_c.htm.

Sentence Stress in English, (2004), Published by: English Club. Retrieved November 25, 2004 from http://pronunciation.englishclub.com/sentence-stress.htm.

Content and Function Words, (2004), Published by: University of Liverpool (UK). Retrieved November 25, 2004 from http://www.liv.ac.uk/CSD/helpdesk/faqs/copycatch/words.htm.

English Language & Literature: Content & Function Words, (n.d.), Published by: Wirral Metropolitan College, Liverpool (UK). Retrieved November 25, 2004 from http://www.wmc.ac.uk/English/function.html.

Full Report of Research Activities and Results, (n.d.). Retrieved November 23, 2004 from http://www.regard.ac.uk/research_findings/R000238260/report.pdf.

Duin, A. H., & Graves, M. F. (1987). "Intensive vocabulary instruction as a prewriting technique". *Reading Research Quarterly*, 22(3): 317-330.

Durán, P., Malvern, D., Richards, B., & Chipere, N. (2004). "Developmental Trends in Lexical Diversity". *Applied Linguistics*, 25(2): 220-242.

Flesch, R., (n.d.) "*How to Write Plain English*" , Published by: Canterbury University, Christchurch (NZ). Retrieved April 27, 2004 from http://www.mang.canterbury.ac.nz/courseinfo/AcademicWriting/Flesch.htm.

Flesher, T. (August, 2003). "Writing to Learn in Mathematics". *The WAC Journal*, 14: 37-48.

Gregory, E., (n.d.) "*Readability*" , Published by: Agricultural Communications. Retrieved July 14, 2004 from http://agcommwww.tamu.edu/market/training/power/readabil.html.

Hochhauser, M. (December, 1999). "Some pros and cons of readability formulas". *Clarity*, 44: 22-28.

Johnson, K., (1998) "*Readability*" . Retrieved June 8, 2004 from http://www.timetabler.com/reading.html.

Klare, G. R. (August, 2000). "Readable Computer Documentation". *ACM Journal of Computer Documentation*, 24(3): 148-168.

Laurén, U., (2002) "*Some lexical features of immersion pupils' oral and written narration (Working Papers 50:63-78)*" , Published by: University of Vaasa, FI. Retrieved July 19, 2004 from http://www.ling.lu.se/disseminations/pdf/50/Lauren.pdf.

Mailloux, S. L., Johnson, M. E., Fisher, D. G., & Pettibone, T. J. (September, 1995). "How Reliable is Computerized Assessment of Readability?". *Computers in Nursing*, 13(5): 221-225.

Merriman, B., Ades, T. &Seffrin, J.R. (May, 2002). "Health Literacy in the Information Age: Communicating cancer information to patients and families". *CA - A Cancer Journal for Clinicians*, 52(3): pp.130-133. Retrieved June 28, 2004 from http://caonline.amcancersoc.org/cgi/content/full/52/3/130.

Nelson, A., (2002) "*Coaching Points: Lexical Diversity*" , Published by: CRA Inc. Retrieved November 23, 2004 from http://www.crawblogs.com/commlog/archives/000533.html.

Spiegel, M. R., Schiller, J., & Srinavasan, R. A. (2001). *"Probability and Statistics"*. New York, NY (USA): McGraw-Hill.

ENDNOTES

[1] Wells, G.C. (1985) *Language Development in the Pre-school Years*. Cambridge: C.U.P. cited in (Durán et al., 2004, p.242)

[2] This criterion ruled out diagram-based techniques such as Fry Reading Age (see http://www.timetabler.com).

[3] *Analyse* has revealed inadequacies on the performance of the FORCAST Formula (see (Johnson, 1998, p.6; Reading in the Content Area, 2004; Gregory, n.d.; Caslon Analytics Profile: Online Readability, 2003; Gregory, n.d.; Caslon Analytics Profile: Online Readability, 2003)). FORCAST Grade Levels tend to settle at approximately 10.0 and the formula has now been dropped from *Analyse*.

[4] See http://www.wintertree-software.com

[5] See http://www.lunerouge.com/freeware/freeware.htm

[6] *Analyse* uses a definition of a syllable based on the presence of a vowel. The *'sentence'* "S." has no vowel so *Analyse* would count no syllable.

[7] Acronyms present an even more difficult problem. For example the acronym *'KISS'* (Keep It Simple Stupid) represents four words and six syllables. Furthermore, they often spell natural English words (e.g. kiss) and so whether or not, in fact, they are acronyms is contextually defined. It was decided that acronyms, by nature, are new words and should be treated as such. Therefore, no word and syllable count adjustment for acronyms was ever made by *Analyse*.

[8] Redish, J. C. and Selzer, J. (1985). The place of readability formulas in technical communication. *Technical Communication*, 32(4): 46-52 cited in (Klare, 2000, p.153)

[9] Schriver, K. A. (1997). *Dynamics in document design*. New York, NY (USA): Wiley cited in (Klare, 2000, p.153)

[10] Hultman, T. G. & Westman, M. (1997) *Gymnasistsvenska*. Lund: Liber Läromedel - cited in (Laurén, 2002)

[11] Lexical Variation (or Lexical Diversity) is not to be confused with Lexical Density. Lexical Variation is the simple ratio cited; Lexical Density is the ratio of the number of *'lexical words'* — lexical words being nouns, main verbs, adjectives and adverbs — to the total number of words in a discourse. (Laurén, 2002)

[12] Terminology varies especially across languages. I have assumed that Lauren is referring to 'Content Words' and 'Function Words'. For example, "Children and foreign travellers learn content words first when they begin to speak. These are the ones which carry the lexical meaning — hotel, beer, double room. Function words carry the grammatical meaning — the, in, where, when." (English Language & Literature: Content & Function Words, n.d.) See also (Content and Function Words,

2004; Introduction to Content/Function Words, 2000; Function Words: The Columbia Guide to Standard American English, 2004) but also (Sentence Stress in English, 2004).

[13] Each word from the document is numbered individually and stored in a data structure. Words for each sample are randomly selected from the structure by number — they are selected solely by number so there may be duplications. These words are stored in a dynamic structure in which duplications are rejected — thereby creating a structure containing the unique words. TTR is then calculated by dividing the number of words drawn for the sample by the number of words in the dynamic structure. This process is repeated the specified number of times for each sample size and the mean of those results calculated and filed. Note that this permits a *'sample'* to be generated which is larger than the document being sampled in which case *'N'* will continue to rise while *'U'* for the *'sample'* can never exceed *'U'* for the document. TTR, then, tends to zero.