Averaged Lexical Variation and Degree of Difficulty: Two new ways to trace writing development

Rick Duley

North Perth, Western Australia 6006

rickduley@gmail.com

ABSTRACT

Monitoring the development of student writing skills requires lexical comparators that are independent of document length. With such comparators it is possible to relate prose artefacts across groups, between individuals and over time. It is also possible to compare student work against an arbitrarily defined standard. These requirements cannot be met by conventional assignment marking. This paper presents an outline of the problems that led to the development of the lexical analysis software package *Analyse*. It was as a consequence of this software development that Averaged Lexical Variation and Degree of Difficulty were devised. These are presented as lexical comparators sufficiently sensitive to be able to identify those developments in student writing skill while meeting the criterion of being independent of document length.

KEYWORDS:

Lexical variation, readability, difficult words, writing

1 INTRODUCTION

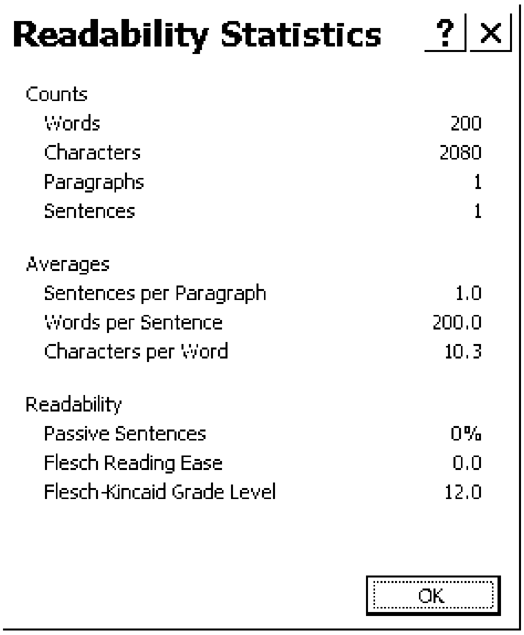
Engineers spend a lot of their time writing. Whether writing takes up 40% of the engineer's time (Beer & McMurrey, 2005, p.2) or just 30% (Petroski, 1993, p.419) commentators agree it is a lot of time. Engineering curricula, then, should include substantial effort to ensure that students graduate with high standards of writing skill. This effort requires a consistent means to monitor the development of that skill and the normal marking of assignments is inadequate to that end. In the first place the marking of assignments necessarily concentrates on content; secondly, marking is a subjective exercise providing no basis for comparison to a standard or between cohorts of students or between individual students.

Neither are standard readability tests, of which there are over 100 (Klare, 2000, p.151), satisfactory. In fact, readability statistics are a poor means of judging the quality of writing. Each formula for evaluating readability produces some number (a grade level, a reading age or simply a number on an undefined scale). While it is easy to show that the three most common readability grading systems — Flesch Reading Ease (FRE), Flesch-Kincaid Grade Level (FKGL) and Gunning Fog Grade Level (GFGL) — are independent of document size, there is little correlation between formulae. Commentators can be scathing:

> *Grade level estimates are almost meaningless. The 1948 Flesch Reading Ease was based on student test performance from 1925 (!) and on US adult educational attainment in 1940 when the average American had 8.5 years of formal education. Now that's up to 13.5 years.* (Hochhauser, 2006)

For any given document, the rating from one formula may vary 40% from the average of ratings from a set of formulae. Research has shown a readability difference of six grade levels among four formulae (Mailloux, Johnson, Fisher, & Pettibone, 1995, p.221). This multiplicity is compounded by computerisation. When computers calculate readability grades the results vary between programs even if the same formula is used. Quite simply, this is because different ways of counting syllables or words will give differing fundamental data and, therefore, different outcomes. Even then the results should be treated with caution:

> *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)



```
Readability Statistics                 ? ×

Counts
  Words                              200
  Characters                        2080
  Paragraphs                           1
  Sentences                            1

Averages
  Sentences per Paragraph            1.0
  Words per Sentence               200.0
  Characters per Word               10.3

Readability
  Passive Sentences                   0%
  Flesch Reading Ease                0.0
  Flesch-Kincaid Grade Level        12.0

                                    OK
```

Figure 1: MSWord Readability Tests

To make matters worse, software that should be trustworthy is not necessarily so. One of the best (or worst) examples used to be Microsoft Word. Figure 1 shows results from the Word 2000 Grammar Checker for a random collection of 200 difficult words (average of 3.8 syllables per word). As one might expect the *'sentence'* is unintelligible gobbledegook yet it scores a grade level of 12 (i.e. suitable for a 17-year-old). Even the associated the results are unreliable; no document with a FKGL of 12.0 can have an FRE of 0.0 and 2080 Characters in 200 Words gives 10.4 Characters per Word not 10.3!

(N.B. New versions of MS Word have corrected this fault.)

Grammar Expert Plus 1.5[1] (GXP) grades the *'sentence'* at a GFGL of 118.0 (*Analyse* rates it at 107.6). That is supposed to mean that the reader needs 100+ years of education to comfortably handle the *'sentence'*. This is probably difficult for a 17-year-old! For comparison, *TextStat 3.0*[2] (TS3) gives the *'sentence'* an FRE of -343.71 (*Analyse* rates it at -318.32). At that level, readability is out of the question. (An FRE as low as 20 might indicate a document from a government department that knows it has something to hide!) Obviously we have a case for *'Caveat Emptor'*!

Other analytical software, such as GXP and TS3, may be more accurate than Word but evaluation of student work means multiple files and the software packages deal with one file at a time. This extra problem led to the development of *Analyse* which searches a folder tree and successively analyses the text files it finds[3]. *Analyse* is batch-operated, i.e. it runs in the background progressively analysing files without operator intervention. As it does so, *Analyse* pre-edits the text files, a matter which has, in the past, been something of an issue. For example, 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.

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 with the following findings:

- Earlier research by the author suggested allowing one word and three syllables for the appearance of a numeric word in a document. It also showed that such allowance makes no statistically significant difference to the resultant readability statistics. *Analyse*, therefore, simply replaces numerals (and numeric operators) with spaces to save work.

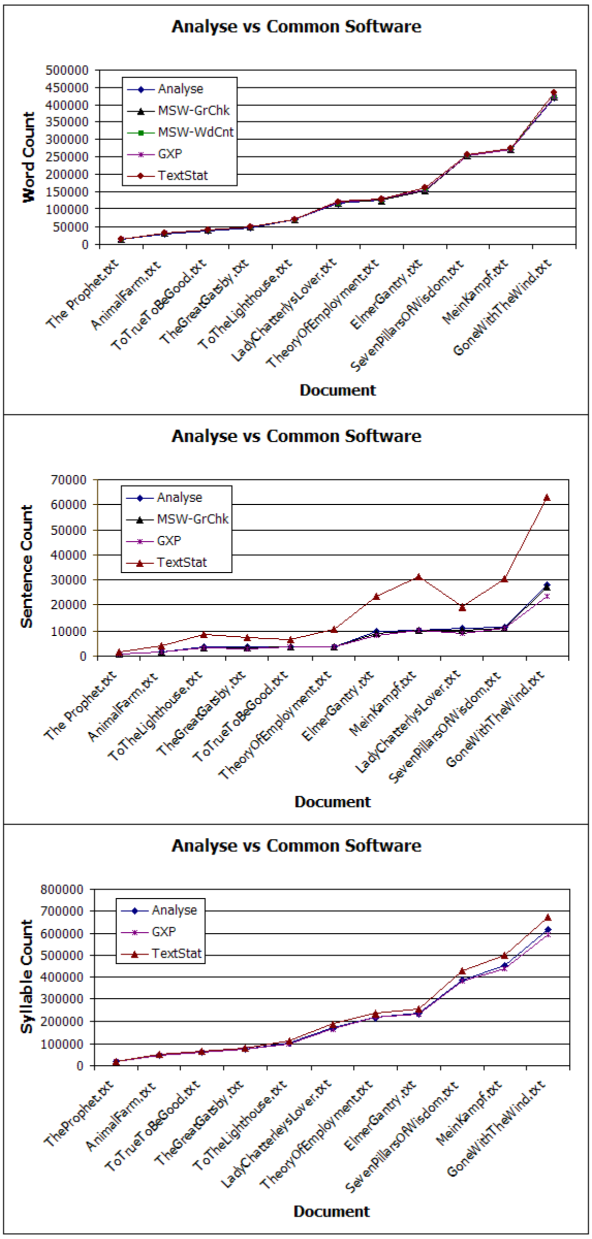


Figure 2: Three Basic Data

- Despite the potential for abbreviations to skew readability statistics, research showed that making allowances for abbreviations in the word and syllable counts makes no statistically significant difference to the resultant readability statistics. *Analyse*, therefore, simply replaces abbreviations with spaces to save work.
- Counter-intuitively, pre-editing to remove tables, lists and a variety of other textual items from texts also made no statistically significant difference to the readability statistics produced from substantial documents. Any such effect is assumed to be related to the size of the documents involved — in substantial documents the number of words contained in tables etc. may be a trivial 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. *Analyse* simply reads tables etc. as normal text and leaves the handling of such matters to the user.)

There is no claim that *Analyse* is absolutely precise, its collection of fundamental data compares favourably with other packages as shown in Figure 2. Furthermore, *Analyse* will work on any number of files and on files of any size which can be accommodated by the computer's RAM. It was the exploitation of these characteristics that led to my investigation into Lexical Variation.

2 INVESTIGATING LEXICAL VARIATION

Lexical Variation[4] (LV) is also referred to as Lexical Diversity. It 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)

LV 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. Neither does extending the vocabulary used necessarily lead to better writing. As Nelson 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'*[5].

2.1 Types, Tokens and Type/Token Ratios

There are many measures of the richness of vocabulary (Tweedie & Baayen, 1998) and many researchers have found it 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 measure of LV, the Type/Token Ratio (TTR), 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*[6] *are part of the linguistic competence, which in different studies has proved to be covariant, for instance, with grammatical competence.* (Laurén, 2002)

TTRs 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 tokens, *'N'*) and two unique words (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, static *'U'* and increasing *'N'*; for any number of types where the number of tokens approaches infinity then

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

Equation 2: Lower Limit of TTR.

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, Malvern, Richards, & Chipere, 2004, p.221)

So, in practice, the TTR for a document will lie between the extremes 1.0 and 0.0 but the problem is that values do not have a linear relationship with discourse length. Figure 3 shows the characteristic curve obtained when TTR is plotted against word count. You can see that it is invalid to compare documents of different lengths on the basis of TTR. However, vocabulary use is an important factor in the development of writing skill and many researchers have applied themselves to this problem — the creation of a valid comparator. Usually they have used some mathematical model (see (Durán et al., 2004) and (Hultman, 1993) for examples of recent work; see also (Tweedie & Baayen, 1998) for a range of examples). This paper does not present a review of this theoretical work but rather takes advantage of the abilities of *Analyse* to present actual results. This is not a presentation of what others think might happen or of what the author thinks might happen but, rather, what was found to happen.

2.2 Variations in the TTR Curve

Computerising TTR analysis made it possible to derive accurate TTR values for documents of extended length. In Figure 3 the extreme right-hand point represents TTR for a database management system technical manual consisting of 460489 words. Counting that many words manually is a task at which most researchers would baulk — let 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:

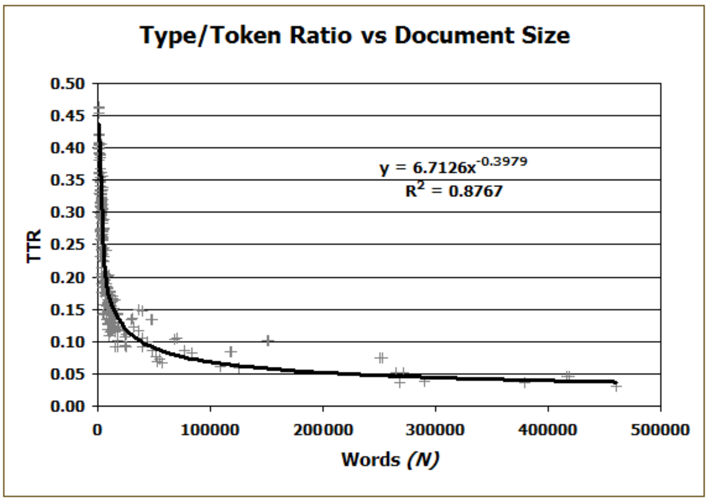


Figure 3: TTR vs Document Size

- There is a relationship 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:

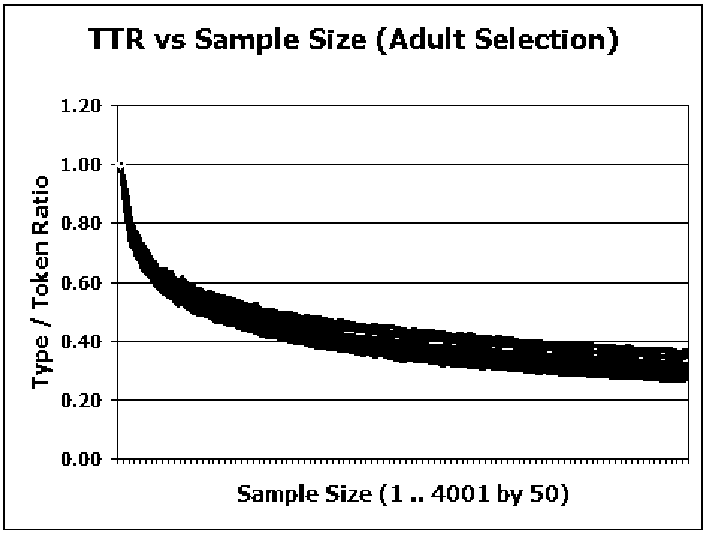


Figure 4: TTR/n for Adult Literature

> ... 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. One hundred 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 data from a different textual genre.

2.2.1 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.

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

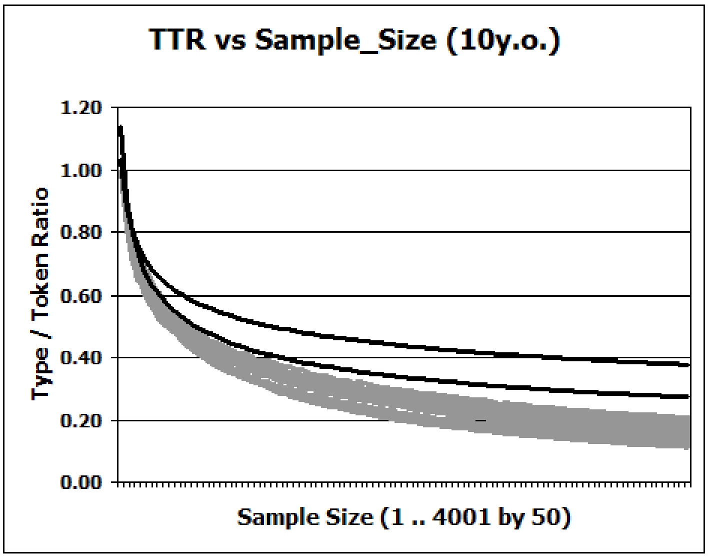


Figure 5: TTR/n for 10yo

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.

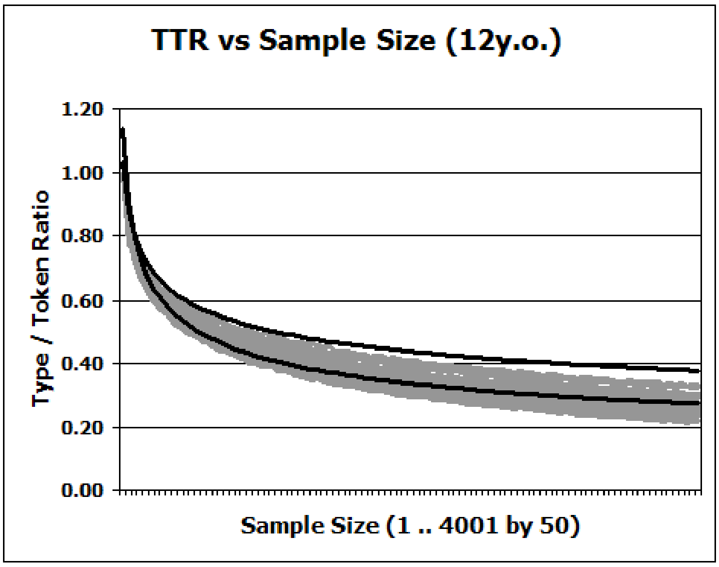


Figure 6: TTR/n for 12yo

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 material with different TTR/N traces. This raised the question of whether or not this holds for other groups.

Articles were collected from four well-known computer industry journals.

- 33 documents in Group One (a highly technical software practitioners' journal — 2,000..33,000 words);
- 35 documents in Group Two (a low-intensity computer industry publication, mostly airline reading — 1,000..5,000 words);

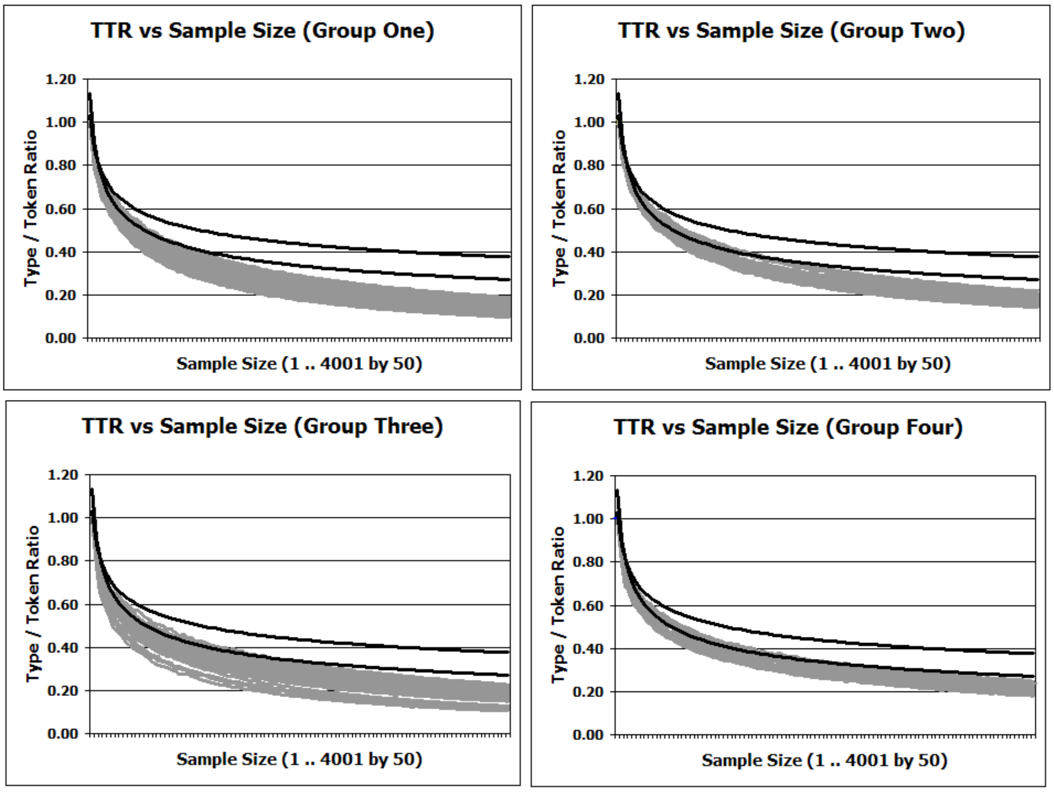


Figure 7: TTR/n for Journal Articles

- 30 documents in Group Three (a low-intensity software industry publication, mostly light study reading — 1,500..6,500 words);
- 31 documents in Group Four (a highly technical hardware practitioners' journal — 3,500..16,000 words).

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 range;
- 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.

2.3 Tracing Linguistic Development

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

2.3.1 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 upper and lower range curves derived from Figure 4 the two numbers would represent a standard range 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. Firstly, student prose artefacts, which *Analyse* was designed to review, are unlikely to be of vast size and the search is for a realistic and appropriate standard which does not require vast size. 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 of whether or not there a simpler way which produces a sufficiently accurate result?

2.3.2 Working with Samples

In tracing any trend there is an obvious need for a standard measurement. Because of the manner in which the range lines used as a standard were derived, i.e. random sampling, the answer to the above question 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[7]. 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 curve for any given sample size **n** for any given document may be compared with the range lines for the same sample size **n** in any given standard set of documents — 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 number of words in the document the sample will, while being representative, produce misleading information (see Endnote 7).

2.3.3 Selecting a Standard Sample Size

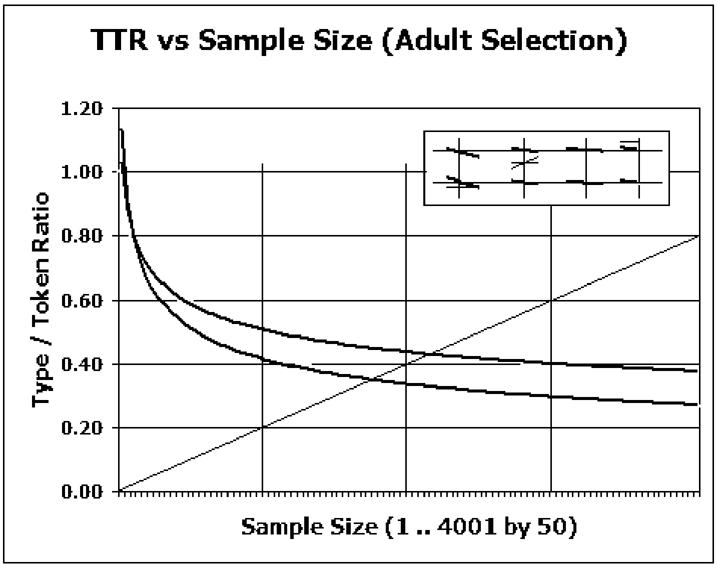


Figure 8: Selecting a Standard Sample Size

In seeking the greatest possible difference between the upper and lower standard range lines, 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.)

Remembering that the purpose of this research is to investigate the work of undergraduate students, sample size must relate to the artefact size expected. Our major student prose artefacts are expected to reach ≈4-5000 words but in some courses, notably programming courses, a majority of the prose artefacts produced by freshman students 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 processing, *Analyse* takes (by default) 100 random samples of 1000 words each from each document analysed.

2.3.4 Effect of Averaging Lexical Variation

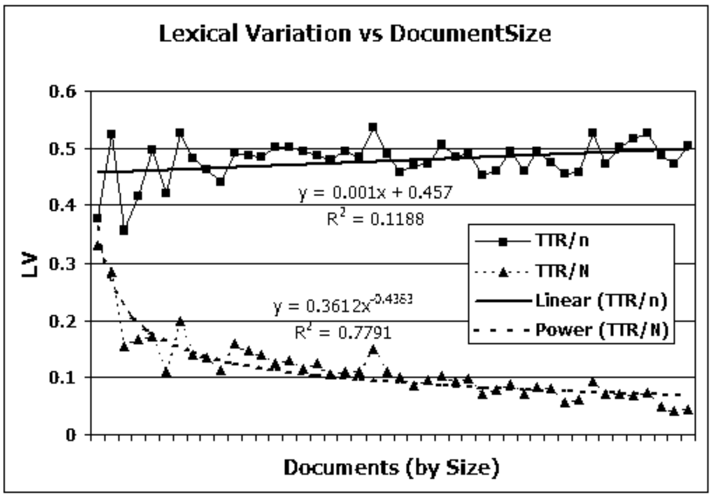


Figure 9: Lexical Variation by Sample and Document

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 for the hypothesis of no slope gave a p-value of 0.204, indicating no significant slope and the relationship of LV to document size has been eliminated. Averaged Lexical Variation (ALV) scores are, therefore, comparable regardless of the size of the document. This has shown 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. Interestingly, what freshman students write involves considerably less range of vocabulary than the material they read as 12-year-olds. It is also 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 <u>guide</u> not an absolute indication.)

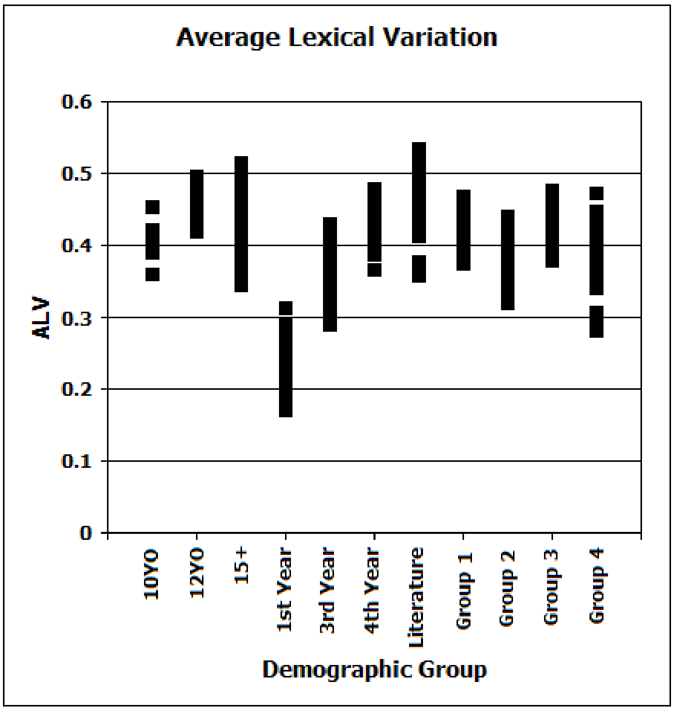


Figure 10: Averaged Lexical Variation by Demographic Group

Now, if ALV indicates the vocabularic range used, we must look at what else can we discover about the vocabulary itself.

3 A CLOSER LOOK AT VOCABULARY

Everyone knows that Scientists use big words. People complain that Scientists use too many big words. So, following the contention of Tom Watson of IBM that one should stamp out gobbledygook (Hargis, 2000, p.127), an investigation was undertaken to find out: (1) if the accusation was based on fact; (2) if there is a valid comparator for big word use in a document; (3) if that comparator could reveal a facet of development in student writing. There were four factors which definition of any comparator must take into account:

- That there should be no relationship between any comparator and the number of words in the document. This would enable the comparator to be used between documents regardless of differences in size;
- Whether or not the comparator should be based on the document as a whole or on some system of sampling the document; and
- Whether or not the comparator should be based on the number of unique difficult words in the document or the total number of difficult words.

With three degrees of freedom and given that $2^3=8$, I was faced with eight possible formulae and required the establishment of an acceptable and recognisable system of notation (For example, does the use of *'T'* for Types precludes its use for Tokens?

Table I: Identifiers for Variables

	Population (P)	Sample (S)	
Types (V)	$\left(\frac{D}{V}\right)_P$	$\left(\frac{D}{V}\right)_S$	Difficult (D)
Tokens (N)	$\left(\frac{D}{N}\right)_P$	$\left(\frac{D}{N}\right)_S$	
Types (V)	$\left(\frac{U}{V}\right)_P$	$\left(\frac{U}{V}\right)_S$	Unique Difficult (U)
Tokens (N)	$\left(\frac{U}{N}\right)_P$	$\left(\frac{U}{N}\right)_S$	

Further, if one is to use *'D´* for Difficult[8] words then which character does one use to represent Unique Difficult words?).

Some commentators use *'N'* and *'U'* to represent Types and Tokens respectively (Durán et al., 2004). However, in this context it was preferred to follow the work of others such as (Bucks, Singh, Cuerden, & Wilcock, 2000, p.77) and define:

- ***'N'*** — the total number of words on the document (Tokens);
- ***'V'*** — the number of uniquely identifiable Natural English Language words in the document (Types).

This reservation of *'U'* then permits the allocation of:

- ***'D'*** — the total number of difficult words (see Endnote 8) in the document;
- ***'U'*** — the number of uniquely identifiable difficult words in the document;
- ***'P'*** — a sample comprising all the words in the document (Population);
- ***'S'*** —a specified number of words randomly selected from the document — duplication allowed — (Sample).

Table I demonstrates the derivation of the corresponding identifiers for each of the eight variables of lexical turgidity.

3.1 Defining the Comparator

Analyse was modified to calculate and report on each of these eight variables. In the first instance it was run on a collection of 296 documents collected from a range of fields and including articles from professional journals, textbooks, technical and user manuals, and reports on computer-industry-related topics. These totalled over 3.5

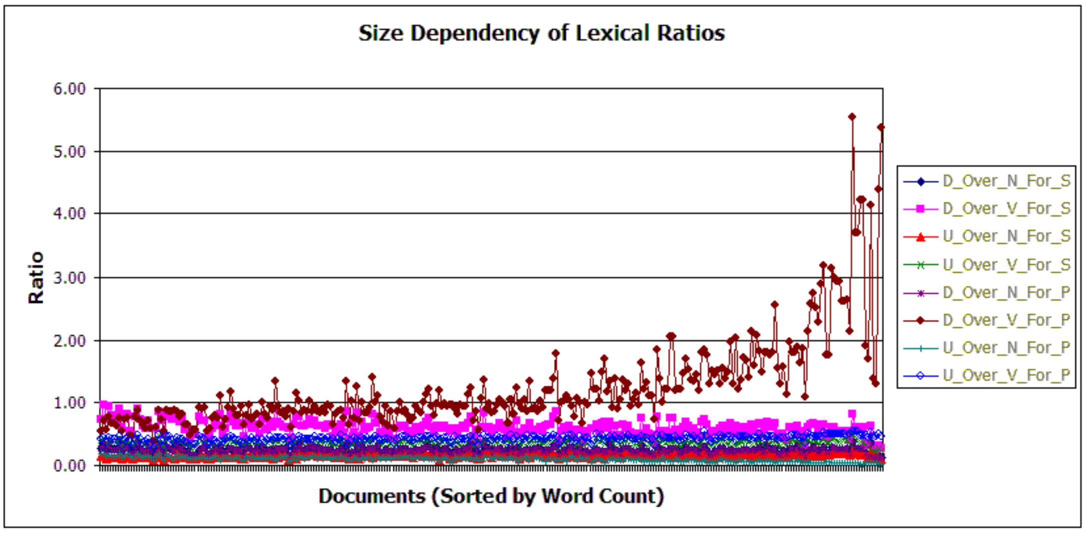


Figure 11: Size Dependency of Lexical Ratios

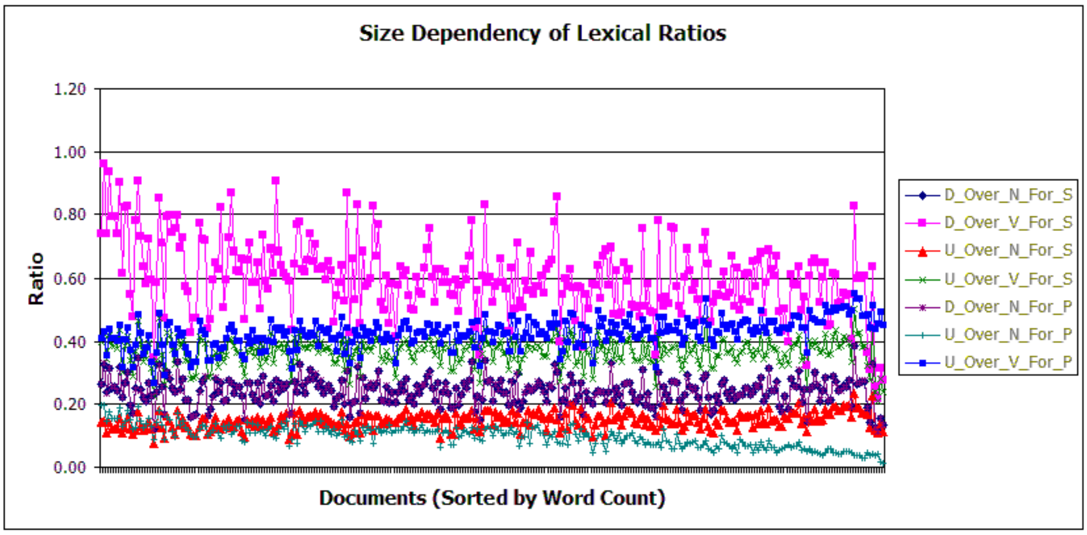


Figure 12: Size Dependency of Lexical Ratios

million words and were deemed to present a broad-spectrum sample of (computer) professional written communication. Figure 11 graphs all eight variables against increasing document size.

3.1.1 Selection from Variables

In the first stage of the process of elimination it was clear from the graphs that some variables failed to meet the criterion of independence of document size. In Figure 11 it is obvious that $(D/V)_P$ increases with *'N'* and its elimination produces the result in Figure 12.

Statistical significance of the variable slopes represented in Figure 12 was not evaluated at this time. There was still a choice of seven variables and two of them, ($(D/V)_S$ and $(U/N)_P$), stand out for two reasons; (1) they indicate a negative trend where the others appear neutral or positive and, (2) they indicate greater slope than the others. Their elimination produces the result in Figure 13.

Since the documents in this sample range in size from 1083 words to 387496 words and the vertical scale covered is almost all within a range of 0.1 for each variable, it is

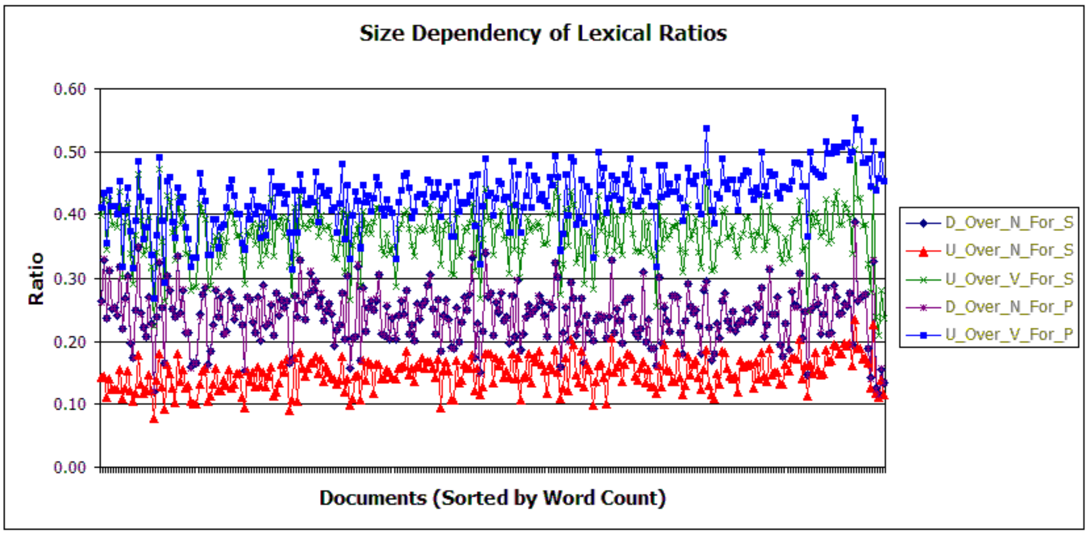


Figure 13: Size Dependency of Lexical Ratios

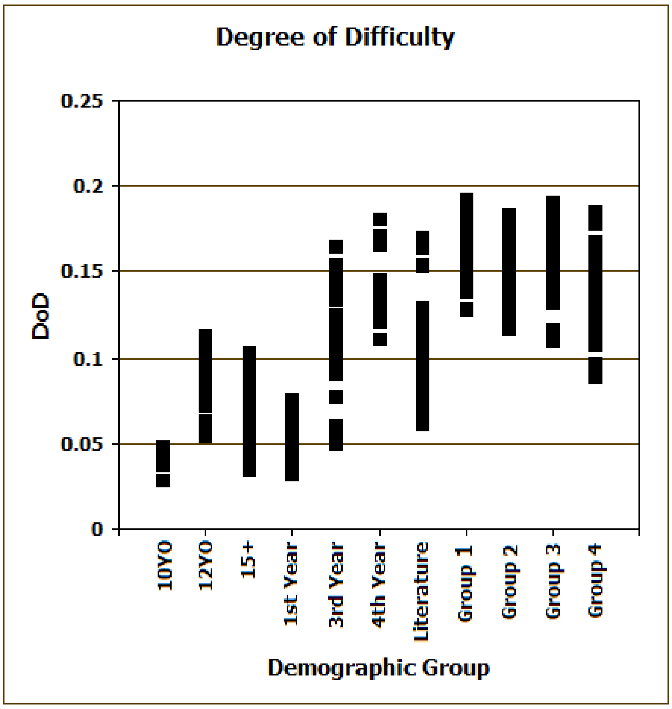


Figure 14: Degree of Difficulty by Demographic Group

impossible to differentiate between the variables just from the graphs. Choosing between the remaining five variables[9] had to be on a rather more precise basis.

Regression tests were carried out between each of the five variable data sets and the corresponding word count range (the X-axis of the graph). Two-tailed t-tests gave the p-values given in Table II. The obvious choice was to take the number of unique difficult words divided by the number of words in a sample — $(U/N)_S$. This decision was easy to implement since *Analyse* already had the required sampling functionality to generate ALV ratings. *Analyse*, therefore, takes 100 samples of 100 words (permitting duplication) from the document and returns the mean result of $(U/N)_S$.

Table II: P-values for Comparators

	P-value
$(D/N)_S$	9.53924E-05
$(U/N)_S$	0·528198124
$(U/V)_S$	5·26114E-05
$(D/N)_P$	8.99162E-05
$(U/V)_P$	3.3674E-06

3.2 Using Big Words

Establishment of a comparator (now referred to as *Degree of Difficulty or ' DoD '*) allowed research focus to shift to the original hypothesis that scientists use more big words than non-scientists. To establish this, *Analyse* was run on the testbed used for Figure 10 and the results are graphed in Figure 14. As with Figure 10, some things are readily discernable:

- DoD for 10-year-olds is lower than DoD for 12-year-olds;
- DoD for 12-year-olds is comparable to that 15+-year-olds;
- DoD for both 12-year-olds and 15+-year-olds approaches that shown for classical adult literature; However,
- DoD for classical adult literature is markedly lower than that for the Professional

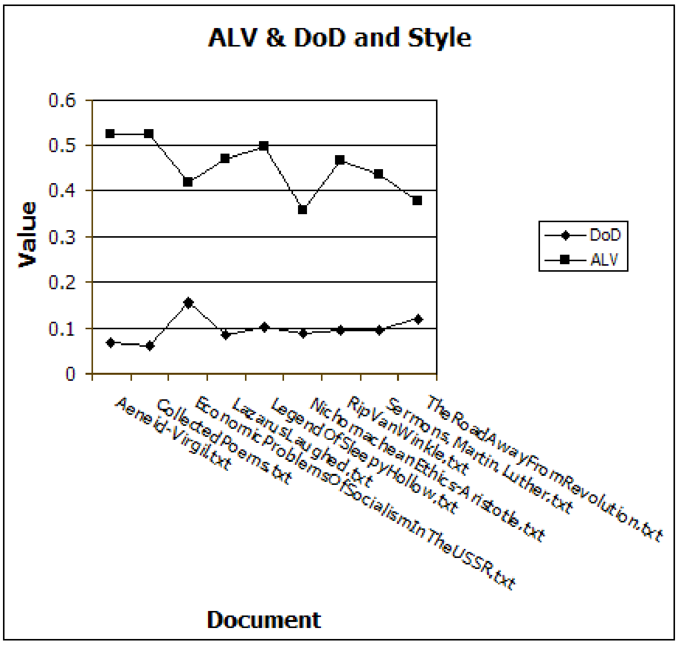


Figure 15: ALV & DoD and Style

Computer Journals in Groups A to D.

- First Year students use simple language but by the Fourth Year the DoD is comparable to professional writing.

So ALV and DoD can indicate developments in writing skill. They can also give us some insight into writing style. Consider Figure 15:

- The highest value for DoD (0.15) refers to the document *"Economic Problems of Socialism in the U.S.S.R."* by Joseph Stalin. However, lest the reader consider such lexical stultification a prerogative of the Eastern Bloc it should be noted that the next lower value in that section of the graph (0.12) refers to *"The Road Away From Revolution"* by Woodrow Wilson.
- Both of these documents also exhibit low ALV (0.42 and 0.38 respectively). Only some writing by Aristotle has lower ALV (0.36).
- The lowest value for DoD (0.06) is not prose but a collection of the verse of Dylan Thomas which also exhibits high ALV (0.52).
- Values returned for two documents by the same author (*"The Legend of Sleepy Hollow"* and *"Rip van Winkle"* by Washington Irving — both scored DoD = 0.1 and ALV = 0.5) may indicate some potential for forensic linguistics.

4 SUMMARY AND CONCLUSIONS

This paper has outlined the problems that led to the development of *Analyse*, a software package for the linguistic analysis of Natural English Language text files. It has presented the basis for, and the development of, two new identifiable characteristics of prose — Averaged Lexical Variation and Degree of Difficulty. ALV and DoD were developed on an empirical basis rather than a theoretical one. This was made possible by particular characteristics designed into the software package *Analyse* — the abilities to analyse large prose artefacts and to analyse any number of them. As such, these comparators are founded on revealed language characteristics rather than on mathematical models of supposed characteristics. ALV and DoD are submitted to the linguistics community as valid prose comparators; statistically proven independent of document size and sufficiently sensitive to identify differences between the writings of demographic groups. Further, they have been shown to be useful to the teachers of writing skill as objective tools for the comparison of student prose — either between groups or against a defined standard.

ACKNOWLEDGEMENTS

This paper would not have been published without the gratefully noted assistance of two other people:

- Martin Dowie of BAE Systems (UK) developed the software package *Ada.Directories.Searches* which permitted the reading and analysis of multiple text files — a major advantage of *Analyse*; and
- Marian Kemp of the Teaching and Learning Department of my university who watched over the statistical interpretation of the generated data.

REFERENCE LIST

Writing Tips - Fog Index, (1998), Published by: University of Minnesota (USA). Retrieved June 8, 2004 from http://www.fpd.finop.umn.edu/groups/ppd/documents/information/writing_tips.cfm.

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.

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.

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.

Readability Tests, (n.d.). Retrieved June 8, 2004 from http://developer.gnome.org/documents/style-guide/x3568.html.

Beer, D., & McMurrey, D. (2005). *"A Guide to Writing as an Engineer"*. Hoboken, NJ (USA): John Wiley & Sons.

Bucks, R. S., Singh, S. et al. (2000). "Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analyzing lexical performance". *Aphasiology,* 14(1): 71-91.

Durán, P., Malvern, D. et al. (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.

Hargis, G. (August, 2000). "Readability and Computer Documentation". *ACM Journal of Computer Documentation,* 24(3): 122-131.

Hochhauser, M. (September, 1997). "Some Overlooked Aspects of Consent Form Readability". *IRB: A Review of Human Subjects Research,* 19(5): 5-9.

Hochhauser, M. (August 24, 2006). Software Feedback. E-mail to R. Duley.

Hultman, T. (1993). *"Hur gick det med OVIX?"*. In U. Teleman (ed.), *Språkbruk, grammatik och språkförändring.* (pp. 55-64). Lund (SWE): Lund University.

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. et al. (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.

Petroski, H. (September, 1993). "Engineers as Writers". *American Scientist,* 81(5): 419-423.

Tweedie, F. J., & Baayen, R. H. (1998). "How Variable May A Constant Be?: Measures of Lexical Richness in Perspective". *Computers and the Humanities*, 32: 323-352.

ENDNOTES

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

[2] See http://www.lunerouge.com

[3] See http://www.freewebs.com/rickduley/Analyse_7-2.htm

[4] 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)

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

[6] 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).

[7] 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.

[8] A *'Difficult'* word is defined as one which has three or more syllables (Writing Tips - Fog Index, 1998; Readability Tests, n.d.; Hochhauser, 1997, p.6).

[9] Five variables are represented in Figure 13, yet a visual examination of the illustration gives the illusion of there being only four. This is because the coordinates for $(D/N)_P$ and $(D/N)_S$ are almost exactly the same (Corr.=0·9995).