Terminological Turgidity: Its use as a lexical comparator

Rick Duley
North Perth, Western Australia
rickduley@gmail.com

Abstract

Scientists use big words. This paper proves it! This paper also follows the search for a means to compare the use of big words between texts and to weigh up the worth of that test in marking student prose.

1 Introduction

Everyone knows that Scientists use big words. People complain that Scientists use too many big words. Did anybody evaluate writing on the basis of big word use? Not that I could find. So, following the contention of Tom Watson of IBM that one should stamp out gobbledygook (Hargis, 2000, p.127), I set out 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.

Following previous work (Duley, 2004a; Duley, 2004c), I reasoned that 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 total count of big words or the number of unique big words;
- Whether or not the comparator should be based on the document as a whole or on some system of sampling the document;
- Whether or not the comparator should be based on the number of Types in the document or the number of Tokens.

With three degrees of freedom and given that $2^3=8$, I was faced with eight possible formulae predicated by the establishment of an acceptable and recognisable system of notation (e.g. the use of *'T'* for Types precludes its use for Tokens? Further, if one is to use *'D´* for Difficult[1] words then which character does one use to represent Unique Difficult words?).

1.1 Notation and Definitions

Some commentators use *'N"* and *'U'* to represent Types and Tokens respectively (see (Durán, Malvern, Richards, & Chipere, 2004)). However,

in this context it was preferred to follow the work of others (see (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[2] in the document (Types).

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

- *'D'* — the total number of difficult words (see Endnote 1) 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 sample comprising a specified number of words randomly selected[3] from the document.

Table 1 demonstrates the derivation of the corresponding identifiers for each of the eight variables of lexical turgidity.

Table 1: 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)	$\left(\frac{U}{N}\right)P$	$\left(\frac{U}{N}\right)S$	

2 Defining a 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 298 documents collected from a range of fields and including articles from professional journals, textbooks, technical and user manuals,

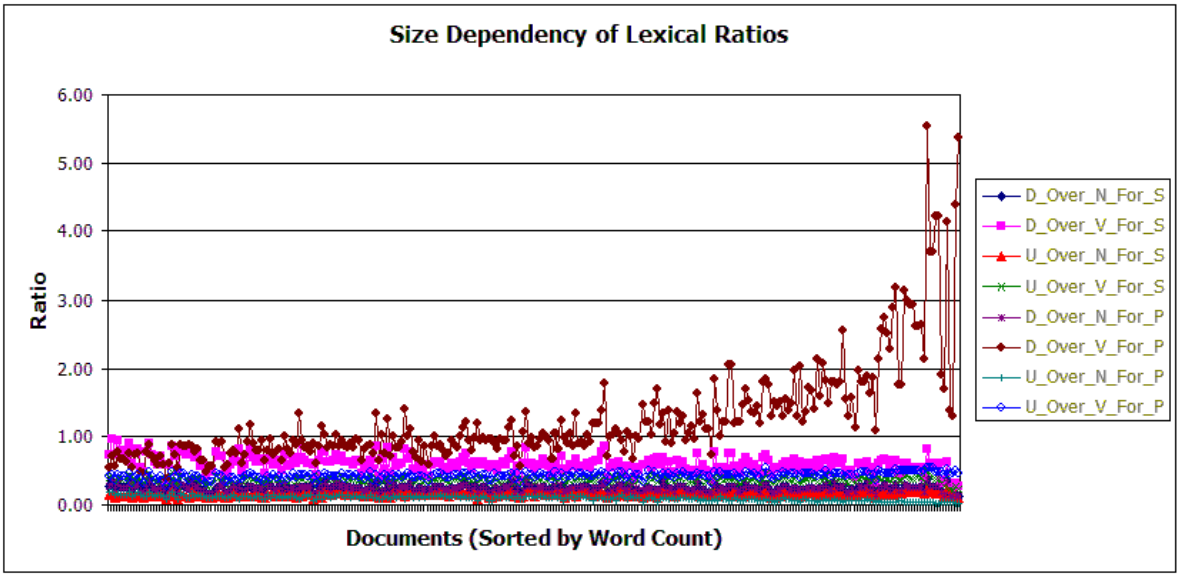

Figure 1: Lexical Variation Ratio Dependence on Document Size - All variables

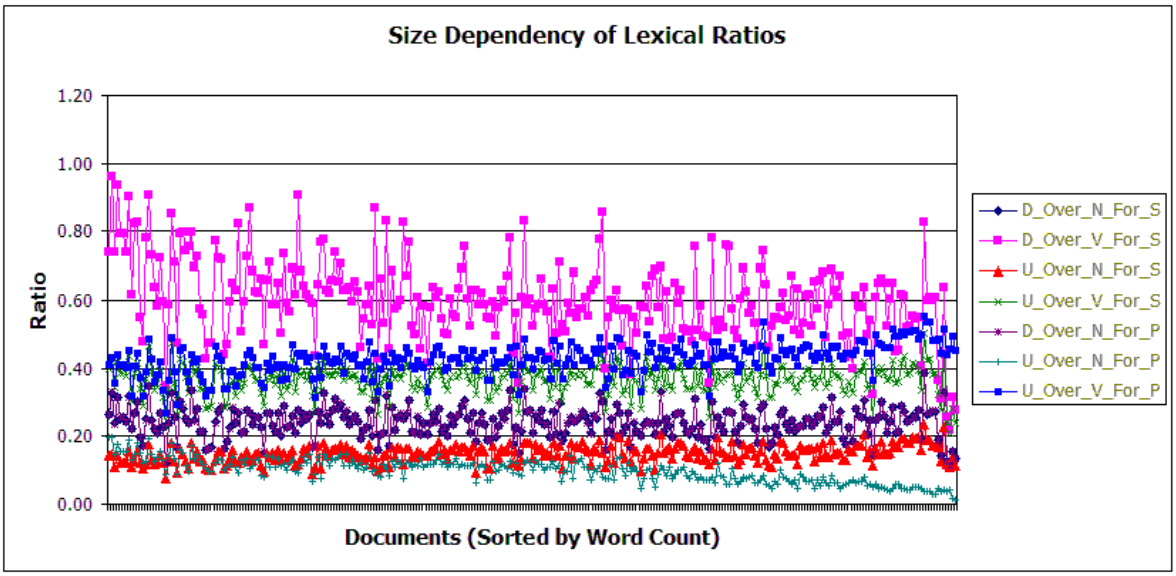

Figure 2: Lexical Variation Ratio Dependence on Document Size - (D/V)P eliminated

and reports on computer-industry-related topics. These totalled over 3·5 million words and were deemed to present a broad-spectrum sample of (computer) professional written communication. Figure 1 graphs all eight variables against increasing document size.

2.1 Elimination of 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 1 it is obvious that (D/V)P increases with *'N'* and its elimination produces the result in Figure 2.

Statistical significance of the variable slopes represented in Figure 2 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 3.

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 impossible to differentiate between the variables just from the graphs. Choosing between the remaining five variables[4] must be on a rather more precise basis.

2.2 Final Selection of A Comparator

Three factors were chosen as the conditions on which the final choice would be made:

- Minimal slope of a trendline to ensure maximum independence from a relationship with document size;
- Minimal variance of the data from the trendline for greater precision;

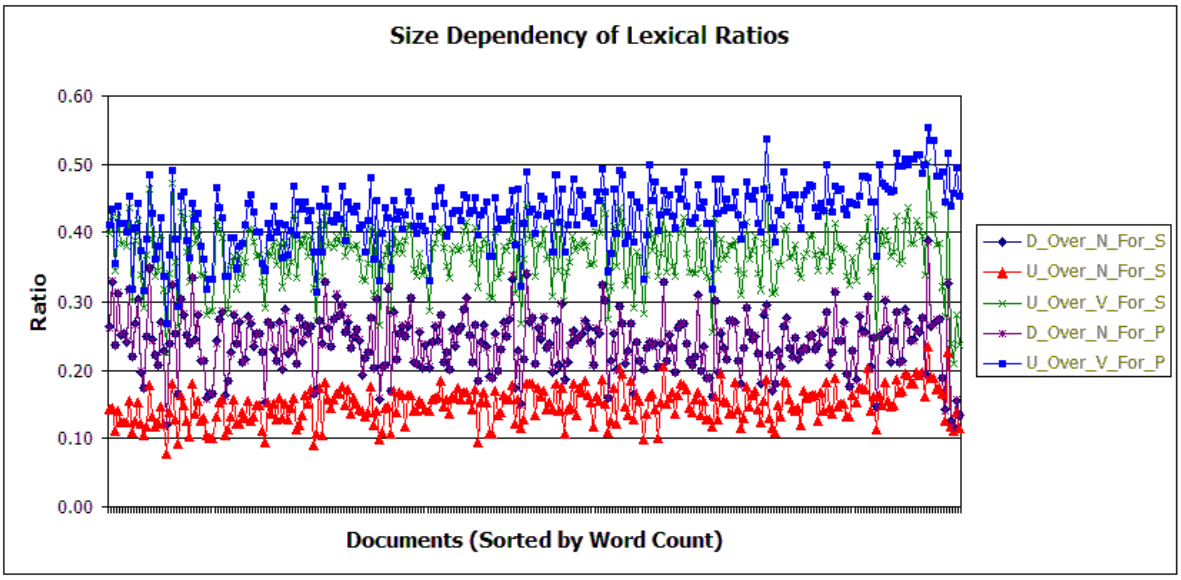

Figure 3: Lexical Variation Ratio Dependence on Document Size - (D/V)S and (U/N)P eliminated

- Minimal work involved in collecting the data — a factor just as important for computing efficiency as for the ease with which the comparator could be manually assessed from a transcript.

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). Resultant Coefficients and P-values (confidence level 95%) are given in Table 2 along with the covariance for each variable. From these data the following decisions were made:

- Eliminate (U/N)S because of the significance of the slope;
- Eliminate (U/V)P because of an appreciably greater slope than the remaining four.

At this point it had to be remembered that the scientific discipline here is Linguistics, not Astrophysics. Numbers expressed to eleven decimal places are simply inappropriate in the measurement of Readability where none or one is usually sufficient.

For this reason the choice between the remaining three variables was purely pragmatic and based on achieving a minimum workload. Calculation of (D/N)P requires no sampling, and no processing to discern the uniqueness of the difficult words.

Finding One: A comparator for lexical

Table 2: Statistical Criteria for Comparator Selection

	Slope	P-value	Covariance
(D/N)S	-3·00671E-07	9.53924E-05	0·00186414
(U/N)S	2.·92433E-08	0·528198124	0·000658077
(U/V)S	-3·18891E-07	5·26114E-05	0·001956248
(D/N)P	-3·02294E-07	8.99162E-05	0·00187095
(U/V)P	3·74653E-07	3.3674E-06	0·00206275

turgidity, independent of document size, may be calculated by dividing the number of Difficult Words in the document by the total number of Natural English Words in the document.

3 Scientists and Big Words

Establishment of a comparator (now referred to as *Degree of Difficulty or 'D'*) 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 another collection of documents. This collection (293 documents comprising 2236298 Natural English words) included:

- Books with calculated readability indicating their suitability for:
  - Those under ten years of age;
  - Those between 10 and 12;
  - Those between 12 and 14;
  - People 15 years of age and older (classical adult literature);
- Documents produced by undergraduate students including:
  - Freshman (First Year) students in Semester One;
  - Freshman (First Year) students in Semester Two;
  - Senior (Fourth Year) students as internship reports;
  - Senior (Fourth Year) students for internal assessment;
- Computer professional journal articles gathered unread from four well-known journals. These are presented anonymously although the journals may be described as:
  - Coffee Table Style (Group A and

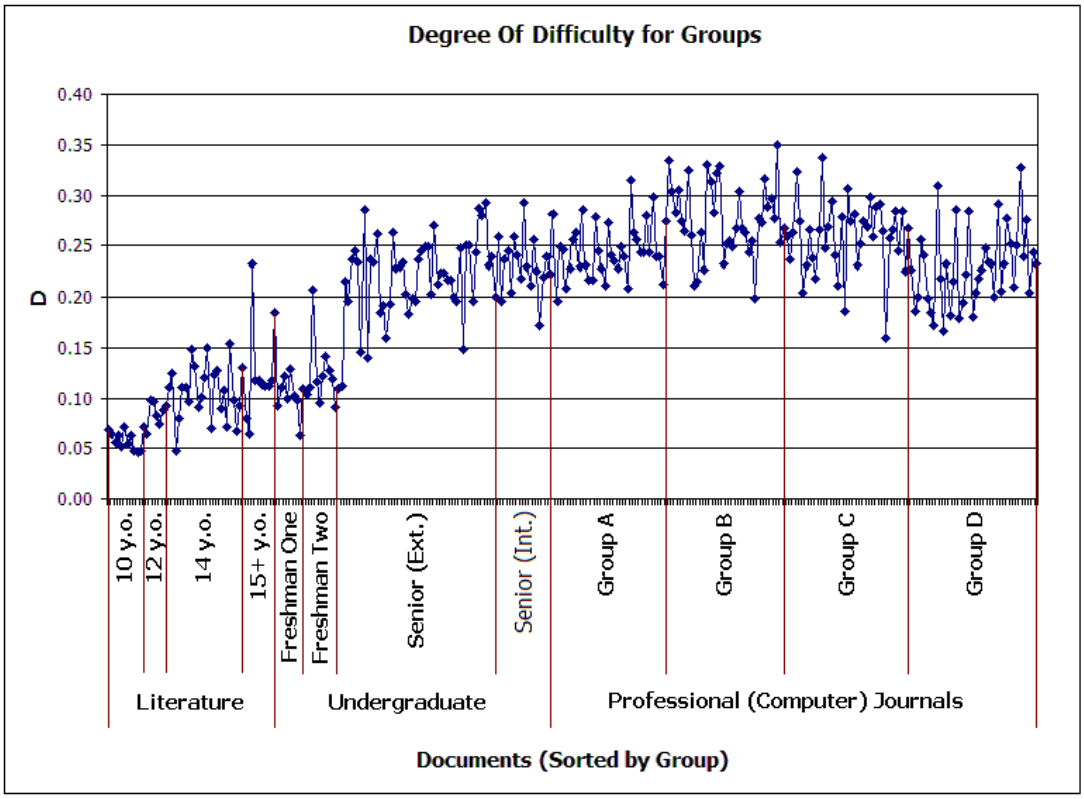

Figure 4: Degree of Difficulty for Characteristic Document Groups

Group D);

- Rather more erudite and arcane (Group B and Group C).

This data is presented in Figure 4.

3.1 Scientific vs Natural Language

Consider the section of Figure 4 sub-titled *'Literature'*. This area depicts the same progression of development as was detected in a similar experiment using standard Readability statistics and described in (Duley, 2004c, section 5.2):

- *D* for 10-year-olds is lower than *D* for 12-year-olds;
- *D* for 12-year-olds is lower than *D* for 14-year-olds;
- *D* for 14-year-olds approaches that shown for classical adult literature (for those 15+ years of age); However,
- *D* for classical adult literature is markedly lower than that for the section sub-titled *'Professional (Computer) Journals'*.

Outliers in this area of the graph provide further interesting insight into the capability of *D*:

- The highest value for *D* in the Adult section (15+) — *D*=0·23 — 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 — *D*=0·18 — refers to *"The Road Away From Revolution"* by Woodrow Wilson.
- The lowest value for *D* in the same section — *D*=0·06 — is not prose but a collection of the verse of Dylan Thomas. Furthermore, the lowest value for *D* in the 14 y.o. section — *D*=0·05 — refers to *"The Child's Garden of Verses"* by Robert Louis Stevenson.

Finding Two: Scientists do use more big words than Non-scientists.

3.2 Tracing Development in Student Writing

Consider the section of Figure 4 sub-titled *'Undergraduate'*. *D* for the documentation produced by the Freshmen is similar to the levels for adult literature — a Natural English Language level — while *D* for the Senior groups is comparable to that for the Professional (Computer) Journals.

Finding Three: *D* can detect a change in the documentation produced by undergraduates as they progress through their programmes.

3.3 *D* for the Professional Papers

Considering the descriptions of the four groups of papers given on page 5, it is interesting to note that Group A and Group D have lower average values of *D* than do Group B and Group C.

4 Conclusions

Evaluation of the Degree of Difficulty of the language of a document (calculated by dividing the total number of Difficult Words in the document by the total number of Natural English words in the document) has been shown to be a useful lexical comparator. *D* is almost

totally independent of the length of the document allowing comparison of virtually any two documents. *D* is sufficiently sensitive to detect changes in the content of the work presented for assessment by undergraduate students at either end of a four-year academic programme.

4.1 The Enigma of D

Linguists regularly refer to the Lexical Variation (or Type/Token Ratio (TTR)) in a document and calculate it as:

$$LV = \frac{V}{N}$$

where V represents the number of Types in the text and N the number of Tokens. This value may be readily shown to be dependent on the length of the document in question. (Figure 5 shows the decrease in TTR as the number of words in a sample of text increases.) On the other hand, D, as sought and as defined, does not show a significant relationship to text length despite the fact that its derivation may be expressed similarly to that of LV:

$$D = \frac{{}^{D}V}{N}$$

where ${}^{D}V$ is the count of Difficult Words rather than Tokens.

4.2 "Caveat Emptor"

Rising *D* values cannot be taken to prove that student prose has progressed from Natural to Professional during the academic programme. To deduce that would be to fall into the classical trap, *"Cows have four legs; my cat has four legs; therefore, my cat is a cow!"*

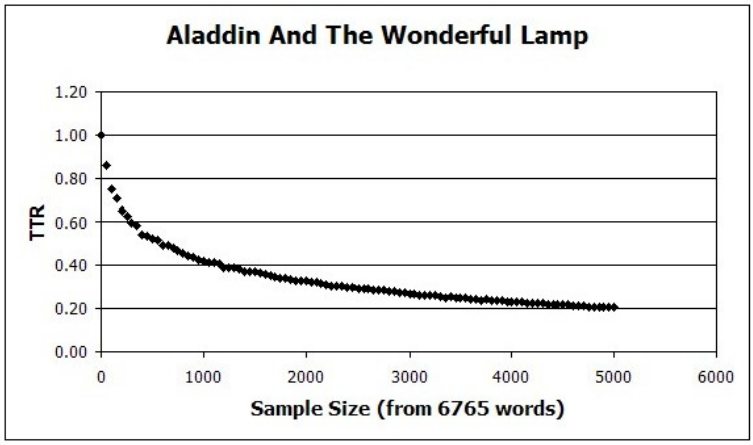

Figure 5 : TTR vs Word Count

D values can only show whether or not senior students use more difficult words in their prose than do junior students. *D* values do not even show whether the difficult words used are the right difficult words.

To put it another way, the demonstrated fact that scientists use more big words than non-scientists, combined with the fact that *D* shows the use of big words, does not make *D* a measure of scientist-ness. *D* can only be an indicator of the possible presence of scientific jargon. Whether or not the presence of scientific jargon is a benefit or a hindrance is another moot and is probably context-dependent.

As with any other readability statistic, *D* must be evaluated in the context of other indicators.

Reference List

Writing Tips, (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.

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

Bucks, R. S., Singh, S., Cuerden, J. M., & Wilcock, G. K. (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.

Duley R.; *"Natural Language Analysis: Its application to the improvement of student authoring skills"*. (Unpublished doctoral work, August 21, 2004a) Murdoch University School of Engineering Science - Rockingham, WA.

Duley, R.; *"Implementing Readability Evaluation in 'Analyse'"*. (Published: October, 2004b) Murdoch University School of Engineering Science - Rockingham, WA.

Duley R.; *"Interpreting Analyse Output"*. (Unpublished doctoral work, December 12, 2004c) Murdoch University School of Engineering Science - Rockingham, WA.

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

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.

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

[2] In this context, the term *'Natural English Language words'* refers to words extracted from the original document text by the software application *Analyse* as described in (Duley, 2004a; Duley, 2004c; Duley, 2004b)

[3] *Analyse v5.0*, as used at the time of writing, operated on the sample selection basis described in (Duley, 2004c, section 4.2).

[4] Five variables are represented in Figure 3, 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.=0·999503). Immediately, one can question the need for sampling.