Replication and Corpus Linguistics
Lexical Networks in Texts

Dr Paul Doyle
English Language and Literature
National Institute of Education
Nanyang Technological University, Singapore
pgdoyle@nie.edu.sg

1 Introduction

This paper examines the issues of replication, a key tenet of the scientific method, and its importance for corpus linguists. The issue is one that several have advocated (Phillips, 1985; Sinclair, 1991; Leech, 1992; Stubbs, 2001; Sampson, 2002), yet hardly any replication studies have been conducted on the major claims and findings in the field, the only known exception being Lee (2000).

The paper explores the issue of replication in corpus linguistics through a critical examination of the methodology used by Phillips (1985) in an investigation of lexical networking as an aspect of text structure in science texts.

2 Replication and the Scientific Method

Before looking critically at lexical networks and the methodology used to identify them in text, it is useful to consider first what the term 'replication' implies, and then how it has been applied in the field of linguistics and corpus linguistics.

Replication is a method of validating research in the sciences, and is generally considered to be a criterion for the acceptance of new theories and knowledge within those domains. In the context of testing hypotheses through empirical research, replication means to repeat an experiment in such a way that scientists can be reasonably certain that the results of two experiments are comparable and, therefore, that they are measuring the same phenomenon. To ensure that the original experiment has been accurately reproduced, the same methodological procedures must be used and the same control of variables applied, otherwise scientists are vulnerable to the contention that they have measured something different in the second experiment. In addition, there is the requirement that scientists measure the same aspects of the phenomena under investigation. If a set of measurements differs from an earlier set, further experiments may be required to establish whether a hypothesis has been disproved or a particular experiment was insufficiently rigorous.

This approach to investigation is usually referred to as the *scientific method* and, while it has come under attack as a misconception of what scientists actually do by philosophers such as Lakatos, Kuhn and Feyerbrand, nevertheless within scientific communities it is still deemed to be the acceptable way of doing empirical research. If

linguistic research is to be considered 'scientific work' in the sense used here, it needs to demonstrate greater concern for issues such as replication.

2.1 Science and Linguistic Data

A currently emerging debate that cuts to the heart of the linguistic endeavour is over the nature of language data and the appropriate method for acquiring linguistic knowledge. The ascendancy in the mid-1960s of generative linguistics, with its exclusive focus on native speaker intuition as data and introspection as the methodology for acquiring that data, has been increasingly challenged by a gradual re-emergence of empirical approaches. An empirical linguistics is one that valorises language performance data held in a corpus and statistical techniques for analysing that data to support linguists' intuition about language. Sampson (2002) is an advocate of this resurgence and has forcefully argued the case for empirical techniques:

> Listen, look. Summarize what you hear and see in hypotheses which are general enough to lead to predictions about future observations. Keep on testing your hypotheses. When new observations disconfirm some of them, look for alternative explanations compatible with all the evidence so far available; and then test these explanations in their turn against further observational data. This is the empirical scientific method (Sampson, 2002: 1).

Tischer et al. (2000) in a cross-disciplinary view of text and discourse analysis, observe that:

> A fundamental rule for any scientific work says that the manner in which the results were obtained must be verifiable. This requirement derives, in essence, from the postulate that scientific discovery is not merely self-discovery: *research must be generalizable and transparent, and (where possible) capable of being replicated and repeated*. (Titscher, Meyer, Wodak and Vetter, 2000: 11 – my emphasis).

Thus, three aspects of the scientific method are fundamental to linguistic analysis. Linguists must be able to verify their methodologies for achieving results, to generalize from these results, and ideally be able to repeat or replicate their research.

2.2 Corpus Linguistics and Replication

For corpus linguistics to be considered truly empirical, it surely has to adhere to the notion of scientific work described above. Surprisingly, however, to date very few corpus linguists have carried out replication studies. Some, it is true, have considered the issue at a theoretical level. Stubbs (2001) makes a strong case for corpus linguistics possessing these key values, when he states that "both data and methods

are publicly accessible", for data are held in "publicly available corpora" and computational methods may be "embodied in software" or "defined in student textbooks" (Stubbs, 2001: 123). But he refines the requirement of replication when he states that:

> It is sometimes necessary to check an analysis by replicating it on exactly the same data, but it is also necessary to check the findings on different but comparable data, in order to see whether they were an artefact of one single data set. (Stubbs, 2001: pp 123-124).

I believe it is crucial that what corpus linguists might mean by "the same data" or "different but comparable data" is made clear. In simple terms, this could mean that the same corpus should be analysed in comparative studies, otherwise no comparison between results is possible. For example, a replication study could entail comparing one set of measurements of word frequencies in the BNC with another. On the face of it, this would appear to be a trivial exercise, for surely a computer will merely produce the same results twice? Stubbs (2001) argues as much when he claims:

> … replication means something more complex that doing exactly what someone else has already done, and strict replication of an experiment is probably rare in all sciences. With computer-assisted work, it could be pointless: if you run the same data through the same computer program, you will get the same output each time. This will tell you nothing about whether the program is working correctly, or whether the procedure is sensible. (Stubbs, 2001: 140).

Unfortunately, this is not the case. We have to be concerned, here, over Stubb's use of "strict" and "probably". Firstly, it is unclear to me how "replication" can be modified by "strict" and maintain any consistency of meaning. Secondly, Stubbs' argument here is greatly undermined by that "probably", for surely he has checked whether or not "strict replication" is rare or not?

Familiarity with software design reveals the fact that implementing an algorithm in a computer language can lead to differences in output depending on the programming language chosen. Moreover, as McEnery (2002) notes, much corpus linguistics research is based upon hand coding which is not easily expressed in supposedly machine independent algorithms. Ostensibly the same cluster analysis techniques implemented in different statistical applications can produce different results on the same data. Wishart (2001) has noted that results from cluster analysis procedures in *SPSS* are sensitive to how the cases are sequenced before applying the cluster analysis method to the data. In my own research, I have found that *WordSmith Tools* (Scott, 1996*)* produces, over time, different results despite using the same data and performing the same actions on that data (Doyle, 2003).

If there has been a substantial period of time between studies (as is the case discussed in this paper), the statement "if you run the same data through the same computer program, you will get the same output each time" over simplifies the matter. It is precisely the fact that the same computer programs are not used which is the default situation in both corpus linguistics and computational linguistics. Lee (2000) encountered the same issue when trying to replicate Biber's 1988 study of dimensions of variation in a corpus using factor analysis. The original data tapes were no longer available to Lee, and the software implementation of the factor analysis algorithms Lee used was different to that used by Biber (Lee, 2000). The upshot of this is that software programs which acquire the textual data and those which analyse the data may no longer be available to subsequent researchers , even if the algorithms are.

Stubbs also notes that:

> A replication will therefore generally be aiming to check whether compatible results arise if some variable is altered… a *deeper form of replication* requires testing the procedures on different texts and therefore cannot be a mere repeat of the original experiment. (Stubbs, 2001: 141 – *my emphasis*)

I think this depends very much on what kind of analysis one is undertaking. If the analysis is between corpora, then there may be some justification in going for the "deeper replication" mentioned by Stubbs, but if the focus is on a single text, then the selection of different texts is clearly inappropriate.

In the context of a corpus-based text analysis, replication would require using the same text on which to base the analysis. A text in a corpus has typically been processed in some way: for example, it might have been POS tagged to disambiguate word senses, or parsed to differentiate syntactical structures. Thus, a word (or other linguistic unit) from the text may be tagged such that the word and its tag could be considered one unit of data: for example `amplifier<NP>`. This textual processing is what differentiates corpora from text collections as objects worthy of study in the domain of corpus linguistics; it is also, I would argue, part of what constitutes the data. While the computer makes it easy to separate the tags from the words in any actual analysis, the text stored on the computer is nevertheless different to that on the printed page. As a result, it is entirely possible for a text in one corpus to be substantially different from the 'same' text in another corpus, since they may be tagged or otherwise augmented in different ways.

Yet even the notion of the 'same text' (without annotation or tagging) is debatable. If you scan a sizeable text (such as a book) into a computer twice, there will be differences between the two versions. If you do this using different software, these differences will be even greater.

3 Investigating Aboutness With A Corpus-Based Methodology

In this section, I describe an attempt to replicate a corpus-based methodology for exploring aboutness in science texts (Phillips, 1985, 1989). The methodology utilised a corpus-based approach and cluster analysis as a statistical tool for identifying significant collocations among content words in order to explore aspects of text structure in science textbooks which were taken from the nascent COBUILD corpus.

In the case of the methodology adopted by Phillips, the corpus was not analysed as a single entity; instead, whole texts were analysed individually, so that while these texts were said to make up a corpus of scientific text, the integrity of each text was preserved during analysis. For both theoretical and practical reasons, the texts were not tagged or parsed (Phillips, 1985: 65), nor were the typical concerns of corpus compilers today, balance and representativeness, addressed. Each text in the corpus was treated as a self sufficient set of data. Moreover, unlike most other corpus linguistic analyses (including the original COBUILD project from whence the texts came), the analysis did not claim to attribute findings to the English language or the written form of the language, or some dimension of language (Biber, 1988), but rather only to each text. From the separate but related analyses of the texts in the corpus, Phillips was then able to make some general statements about the degree of lexical networking in scientific text and thus establish that there was organisation in text at the lexical level.

3.1 Aboutness, Lexical Networks and Collocation

Phillips believed that "the crucial point concerning aboutness is that it is a type of meaning arising from the global structuring of text" (Phillips, 1985: 30). His research hypothesis, therefore, was that:

> knowledge-free analysis of the terms in a text... will reveal evidence of systematic and large-scale patterning which can be interpreted as contributing to the semantic structure of text and hence as constituting a major device through which the notion of content arises. (Phillips, 1985: 26)

By "knowledge-free analysis", Phillips meant an analysis that did not depend on semantic notions originating with the introspection of the researcher, but rather dealt empirically with the "syntagmatic patterning of the textual substance" itself (Phillips, 1985: 26). If, as he claimed, the text projected its own reality rather than referred to an external one, it would be incorrect to begin a process that was intended to find structure in a text by imposing pre-conceived categories not derived from the text itself.

Having arrived at this position, Phillips went on to develop a methodology based on the patterns of association between textual elements, or *lemmata* (hereafter referred to by the more common term, *lemmas*). Using this method to analyse science textbooks, Phillips was able to make two claims. Firstly, he identified what he termed "lexical

networks" which could be related to the topics of chapters in the textbooks. Secondly, the relationship between these networks in different chapters revealed a level of global text structure. Figure 1 illustrates both these claims, showing two networks from different chapters of a textbook.

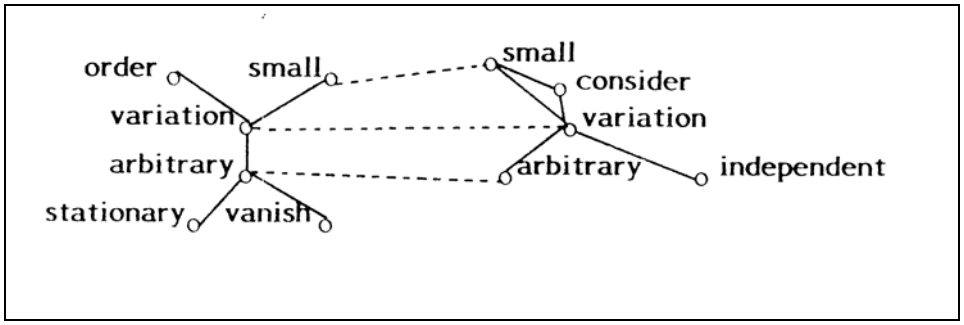

Figure 1 Lexical Networks from Two Chapters of an Undergraduate Textbook (Phillips, 1989).

The network on the left in Figure 1 is from Chapter 3 of *Classical Mechanics* (Kibble, 1973), while the network on the right is from Chapter 11. Phillips' argued that, given the re-occurrence of the lemmas *arbitrary*, *variation* and *small* in the two networks (shown here as dashed lines), there was a significant relationship between the lexical networks. Such relationships were sufficient evidence, he believed, for the claim that text is lexically organised.

Commenting on the two linked networks shown in Figure 1, Phillips related these empirical phenomena, which had been uncovered by a combination of computerised identification of collocates and cluster analysis of their frequency of co-occurrence, to aboutness: "What is observed here is the notion of "arbitrary small variation" operating in slightly different contexts in the two chapters." (Phillips, 1989: 58). Similar analyses have since been carried out to identify "topical nets" or "landscapes" in spoken texts such as business meetings (Collins and Scott, 1996), and "collocational networks" in plant biology research articles (Williams, 1998).

3.2 Reproducing a Corpus

It would seem clear that, in order to replicate this study of lexical networks in science textbooks, it is necessary to use the same corpus comprising the same texts. To create his corpus, Phillips was able to use science textbooks which had been collected for the COBUILD dictionary project and scanned with a Kurzweil scanner (an account of how this was done and some of the problems that were encountered is given in Sinclair, 1987). For reasons reported in Doyle (2003), it was believed that analysis of one of the eleven texts would, for the purpose of replication, be sufficient to produce meaningful results.

If we can verify that the starting point, that is, the original data is the same, then we are actually analysing the same text and therefore any statements about the comparative worth of the 1985 technique against the 2003 technique are valid. Since

the aim was to compare the lexical networks revealed by the two methods, one text should be sufficient to reveal similarities or differences in the findings, even if the results could not be generalised beyond that text. By replicating the analysis of lexical networks in this text, a basis for comparing results achieved using different software to implement the methodology would be established. The main reason for holding with this view is that, as mentioned earlier, Phillips did not really deal with his corpus as a single entity, thus there is no requirement to do so in this replication study. Despite this, it has to be admitted that a more comprehensive result could be attained by comparing all eleven analyses that he carried out.

One of the original texts (Ahmad and Spreadbury, 1973, *Electronics for Engineers*) was available and, fortuitously, this was the only text for which a full set of dendrograms showing the results of cluster analyses on lemma co-occurrence data for each chapter of this text was available (Phillips, 1985: 255-265). Thus, the approximately 60,000 words that make up *Electronics for Engineers* were scanned and proof read to produce a digital version of the text. To recreate Phillips' data, his own procedure had to be followed exactly: "The entire text with the exception of the preface and the appendices is contained in the sample." (Phillips, 1985; 107). In practical terms, this meant scanning 232 pages, or 116 sheets of A4 size photocopies of the facing page layout. Once these 116 facing pages had been recognised and proofed in the OCR software, the results were saved as ten separate text files, one for each chapter and for each quiz. The ten edited text files were then combined in *Word* to form a complete copy of the original text, excluding its preface, contents page, answers to problems, appendices and index, and excluding all diagrams. It should be noted that recreating a text by scanning was not without its problems and challenges. OCR technology has certainly improved since the 1980s, but not sufficiently to rely on the computer. In addition, texts such as this which contain columns and diagrams and formulae on nearly every page still present major problems.

To facilitate reference, the following naming conventions will now be used throughout the subsequent discussion. The printed version of the book, *Electronics for Engineers* (Ahmad and Spreadbury, 1973) will be referred to as ELEN. The two sets of data being compared will be distinguished by referring to the original scanned version of the text as ELEN1 (Phillips, 1985) and to my scanned version of the text as ELEN2 (Doyle, 2003).

3.3 Recreating Lemmas

Once the text has been acquired through scanning and OCR, the next stage was to recreate the list of lemmas to be analysed for collocation with each other and their frequency of occurrence in each chapter of ELEN. As has been argued elsewhere (Sinclair, 1991), the process of lemmatization is not a precise, universally agreed one. Decisions about the inclusion or exclusion of word forms are subjective. For example, strict lemmatization would combine *amplifier* and *amplifiers*, but not *amplified* and *amplifying*, since the latter are verb forms while the former are noun forms. In practice, however, such demarcations are not clear cut and depend on context. In ELEN, both *amplified* and *amplifying* occur frequently as part of noun phrases, but

never as verb forms and thus combining them within the lemma amplify would appear entirely justified (lemmas are shown underlined from here on).

When trying to recreate the data in Phillips (1985), around a hundred such decisions taken by the initial researcher had to be retraced and checked to ensure that "the same data" (i.e. lemmata frequencies for each chapter of the text ELEN) was used. Thus, it was not surprising to find that the frequency figures for words in the original data (Phillips, 1985) and in my data (Doyle, 2003: 294-295) were not identical.

One way of establishing whether the two sets of data are the same is to compare the frequencies reported by the *CLOC* software for the 198 lemmas in ELEN1 (Phillips, 1985: 242-243) with the frequencies as shown by *WordSmith Tools* for the same lemmas in the data from ELEN2. Figure 2 illustrates one way of doing this by graphing the amount of variance between two counts against the number of lemmas exhibiting this amount of variance. The graph shows that, for 198 lemmas considered, 81 (41%) had the same frequency count, whereas 117 lemmas showed some variance (59%) in frequency between the two studies. However, 80 of those 117 lemmas (a further 40% of all lemmas) showed variance no greater than ± 2, which demonstrates a high degree of similarity. Appendix A shows the complete set of lemmas and their individual percentage of variance.

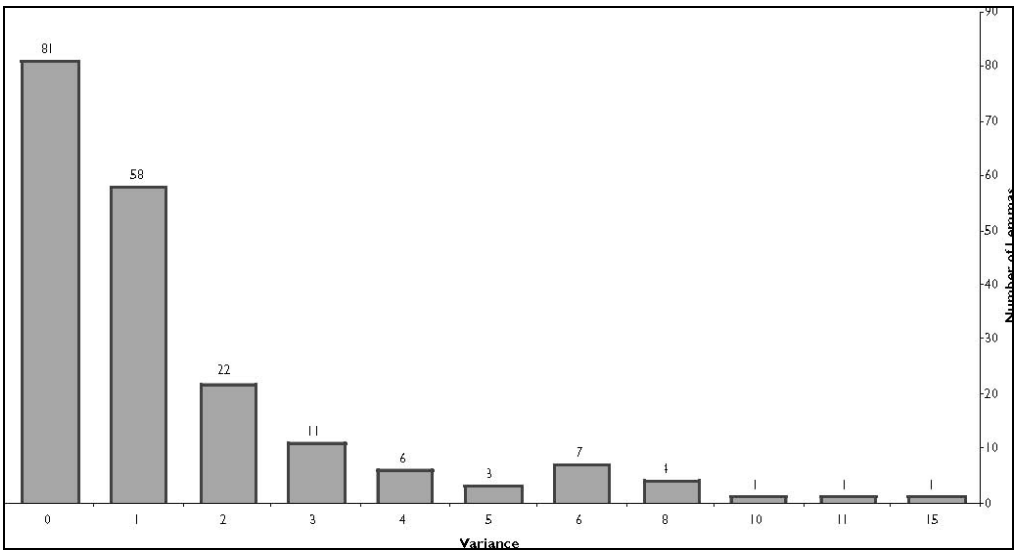

Figure 2 Graph of Variance in Lemma Frequency in Two Versions of a Text.

Another way of comparing the two sets of frequencies is to use a statistical measure of their variance. Table 1 gives basic statistics for the two sets of lemma frequencies.

	Author	**N**	**Mean**	**Std. Deviation**	**Std. Error Mean**
LEMMA COUNT	ELEN1	198	62.462	101.851	7.257
	ELEN2	198	62.452	101.531	7.234

Table 1 Statistics for Lemma Frequency Count for ELEN1 and ELEN2.

The standard deviation is extremely high for these two sets of data (101.85 and 101.53 respectively) due to the fact that the data is not normally distributed. There are a few cases of lemmas with frequencies of several hundreds, while most lemmas have a frequency between 10 and 50. Statistical measures such as mean and standard deviation have to be interpreted relatively rather than absolutely, because a list of lemma frequencies in a text will always be highly skewed: there will be a few lemmas with very high frequencies and many with low frequencies. This reflects the basic facts of word frequency distribution in texts and underlies the critical assumption of this study that the relations between high frequency content words are central to the creation of topic specific clustering. Thus it is the differences between the mean, standard deviation and the standard error mean which should be considered for the lemma frequencies given in Table 1. In every case, these are very low: the difference for the mean is 0.01, for the standard deviation (SD) 0.32 and for the standard error mean 0.023. This suggests that the two sets of data are very similar. To determine this statistically, a further test can be carried out on the 198 frequency counts.

Tables 2 and 3 give details for a comparison of these two sets of frequencies using the t-test statistic implemented in *SPSS*. Since the same lemmas are counted in both studies, the appropriate test for comparing the two sets of data is the paired t-test.

	N	**Correlation**	**Sig.**
Pair 1 Phillips & Doyle	198	.999	.000

Table 2 Paired Samples Correlations for Lemma Frequencies by Phillips and Doyle.

	Mean	**Std. Deviation**	**Std. Error Mean**	**95% Confidence Interval of the Difference**		**t**	**df**	**Sig. (2tailed)**
				Lower	**Upper**			
Pair 1 Phillips -Doyle	1.015E-02	3.2810	.2338	-.4509	.4712	.043	196	.965

Table 3 Independent Samples T-Test for Lemma Frequencies in Phillips and Doyle.

The hypothesis here will be a two-tailed one as there is no reason to suppose that the lemma frequency counts will be consistently higher for my study. Table 2 shows that there is a strong relationship between the two sets of data (r = 0.999), as might be expected (it is unlikely that frequency counts for the lemmas would vary significantly from one observation to the next). Nevertheless, it is necessary to look at the value of *t* to determine whether there is a significant difference between the two sets of data. Table 3 shows that there is no significant difference in the variance between individual lemma counts. The difference between the mean of the lemma frequencies for Phillips' data and the mean of the lemma frequencies for my data is 0.102, with a t-value of 0.43 with an associated probability level of 0.965. The confidence interval

is quite small – we can say, with 95% confidence that the true population mean difference is between -0.451 and 0.471. Thus we cannot discount the null hypothesis that there is no significant difference between these figures. In other words, it has been established that the lemma frequencies for ELEN1 and ELEN2, while not identical, exhibit significant similarity.

As a result of these graphical comparisons and statistical tests, it can be stated that there is negligible difference between the two sets of lemma frequency counts for the 198 lemmas analysed. Given this similarity, it can be claimed that there are strong grounds to believe that Phillips' and my decisions about lemmatizing are identical. This issue has been considered at length because replication of Phillips' original study is not really feasible unless the two sets of lemma counts are sufficiently similar. If this were not the case, it would be difficult to claim that lexical networks identified by the current analysis are in any way comparable to those in the earlier study.

3.4 Retrieving Collocates from Text

Having confirmed that the lemma frequencies are comparable, we now need to retrieve the significant collocates for these lemmas and their frequency of co-occurrence. For every chapter of ELEN2, the 198 lemmas were analysed for significant collocation with each other. In practical terms, this entailed creating a concordance for all 198 lemmas in the seven chapters that comprise ELEN2, such as the one for power shown in Table 4.

1	. Then a current driven	power	amplifier stage will have
2	put resistance (¤3.12) and in a	power	amplifier (¤3.15). Other us
3	polar transistor is used in the	power	amplifiers of our domestic
4	licon devices, and higher for a	power	device which has a large ju
5	y is safely under the allowable	power	dissipation curve.
6	alanche may start). (c) Maximum	power	dissipation, P.This may
7	alculate the heat flow rate as:	Power	dissipated = . Thus the tem
8	(it may be enclosed with other	power	dissipating devices so T, m
9	an be neglected so that a given	power	dissipation is given by the
10	llector voltage; and c, maximum	power	dissipation. When the trans
11	ut resistance, voltage gain and	power	gain for the amplifier of f
12	Derive the voltage, current and	power	gain relationships and then
13	hat is the stage efficiency and	power	gain? (d) The transistor ha
14	than unity, but its voltage and	power	gains are considerable and
15	.7 ~l (¤ 1.6).To calculate the	power	in this load, the currents
16	d, it must equal the electrical	power	input to the transistor whi
17	at is the optimum load and what	power	is developed in it by 15 mA
18	we can modify it, this form of	power	matching would give rise to
19	criterion in chapter 1 for best	power	matching. ft is a deficienc
20	o make the assumptions that the	power	supply is well decoupled an
21	type (about 4 deg c/ watt for a	power	transistor in a T03 case, a
22	k box' and the input and output	powers	are marked.... (d) The
23	igher frequencies and at higher	powers	; it is easier to use in

Table 4 Concordance of Lemma *Power* in Chapter 3 of ELEN 2 (*WordSmith Tools*).

As the software used by Phillips, *CLOC* (Reed, 1981), was no longer available, *WordSmith Tools* (Scott, 1996) was used instead.

	LEMMA	F	L4	L3	L2	L1	N	R1	R2	R3	R4
1	**amplifier**	3	0	0	0	0	-	2	0	0	1
2	and	12	1	3	2	4	-	0	0	1	1
3	are	2	0	0	0	0	-	1	1	0	0
4	**current**	2	0	0	2	0	-	0	0	0	0
5	*dissipation*	4	0	0	0	0	-	4	0	0	0
6	for	4	0	0	3	0	-	0	1	0	0
7	*gain*	4	0	0	1	0	-	3	0	0	0
8	*given*	2	0	0	0	1	-	0	0	1	0
9	*higher*	2	0	1	0	1	-	0	0	0	0
10	in	5	0	0	2	0	-	1	1	1	0
11	**input**	2	0	1	0	0	-	1	0	0	0
12	is	6	1	0	0	0	-	1	3	1	0
13	it	3	1	0	0	0	-	1	0	0	1
14	**load**	2	0	1	0	0	-	0	0	1	0
15	*matching*	2	0	0	0	0	-	2	0	0	0
16	*maximum*	2	0	0	0	2	-	0	0	0	0
17	may	2	1	0	0	0	-	0	0	0	1
18	of	2	0	0	0	1	-	0	1	0	0
19	other	2	0	0	0	1	-	0	0	1	0
20	so	2	1	0	0	0	-	0	0	1	0
21	*stage*	2	0	1	0	0	-	0	1	0	0
22	that	2	0	1	1	0	-	0	0	0	0
23	the	15	4	0	2	3	-	0	0	4	2
24	then	2	1	0	0	0	-	0	0	0	1
25	this	3	0	1	0	0	-	0	1	1	0
26	to	3	0	1	0	0	-	0	1	0	1
27	**transistor**	4	0	0	0	0	-	1	0	0	3
28	**voltage**	4	1	2	1	0	-	0	0	0	0

Table 5 Collocates of power, Frequency >1, in ELEN2 Chapter 3, Sorted Alphabetically.

Each concordance is analysed for collocates by counting up all the words that occur above a specific threshold frequency within a text window, or span, of a specified number of words to the left and right of the node lemma. Phillips used a span of 4 and threshold frequency of 2 on the rationale that this would give maximum information about patterning in the text, whilst filtering out the 'noise' of nonce collocations (Phillips, 1985: 111). It is debatable, however, whether two occurrences of a lemma within the text window around the node lemma is sufficiently high enough to be deemed "significant" (in a future paper I will argue that a threshold frequency of three should be considered instead).

Wordsmith Tools can retrieve collocates from a concordance and produce two lists of these collocates, one in alphabetical and one in frequency order. The output is not filtered in any way, as seen in Table 5, which shows the alphabetical listing of 28 collocates for the lemma power and their frequency (F) of collocation. The items in bold are words that correspond to the lemmas which Phillips selected at random in his study (Phillips, 1985). Thus the output contains a large number of high frequency function words, which are not excluded here by the stop list function in *Wordsmith Tools*, a weakness of the software as it stands. In addition, there are a number of words superfluous to the analysis, because their overall frequency in the whole text was less than ten. Finally, there are seven words (in italics) that are not function words, which have a text frequency of ten or more in the whole text, but which Phillips excluded in his random sample of the lemmas in ELEN1. For the purpose of replication, only the word forms of the 198 lemmas identified earlier were of interest. Therefore, all the high frequency function words and all the content words that are ruled out by either of the two reasons stated above must be deleted from the list of collocates and their frequencies lists, as shown in Table 6.

	LEMMA	**F**
1	amplifier	3
2	given	2
3	higher	2
4	input	2
5	transistor	4
6	voltage	4

Table 6 Eligible Collocate Frequencies for Lemma power

Once these extraneous items had been removed, the list of collocates and their frequencies were saved as a text file for subsequent import into *Microsoft Excel* where they were manipulated to create the matrix format of data required by a cluster analysis. Thus, in terms of replication, when retrieving the collocating lemma frequencies it was merely necessary to ensure that the same span (± 4 words) and threshold frequency of collocation (2) were entered in *Wordsmith Tools* before concordancing each of the 198 lemmas in each chapter in ELEN2.

3.5 Identifying Lexical Networks Using Cluster Analysis

The next stage of the methodology requires the researcher to identify lexical networks by looking for patterns of collocation among the 198 lemmas and their co-occurrence frequencies for each chapter of ELEN2. To do this, Phillips argued for the use of hierarchical cluster analysis, an exploratory statistical technique that identifies groups in data. Hierarchical cluster analysis is an umbrella term for a variety of sophisticated statistical procedures and related measures of association, or distance measures. A review of the literature on cluster analysis reveals that there is an inconsistent use of terminology for these procedures, though not all statisticians make the reader aware

that this is so. Everitt (1993), for example, gives alternative names for virtually every clustering procedure discussed. This problem is compounded by the names assigned to procedures in software implementations of the relevant algorithms. A further concern, identified by Wishart (2003) is that not all software implementations of clustering procedures are the same.

Thus, replication involved finding a software implementation of the specific hierarchical cluster analysis method Phillips used, Ward's method, since the software originally used was no longer available (*Clustan 3rd Edition*, Wishart, 1978). One of the issues here is knowing whether a commercial statistical application such as *SPSS for Windows* (SPSS Inc., 1997-2002) can be trusted for its implementation of statistical algorithms. Wishart (2001), for example, has critiqued the implementation of the k-means clustering procedure in a number of software packages, including SPSS, and states:

> … it is possible for different stable k-means cluster solutions to be obtained under different starting conditions or simply by changing the case order, so you need to explore a range of classifications produced by the procedure and examine their criterion values. (Wishart, 2003).

Phillips (1985) reported that Ward's Method, (aka Increase in Sum of Squares), gave more satisfactory results than the Density Search (aka Centroid) technique. Ward's Method has been implemented in several software packages, principally *SAS* (SAS Institute, Inc. 1989), *SPSS* (SPSS Inc. 2000), *STATISTICA*, Minitab (Minitab Inc. 1999), *Clustan/PC* and *ClustanGraphics* (Wishart, 1994 and 1999). For the purpose of replication, in this research I used *SPSS*.

The need for replication also required that a specific similarity coefficient had to be used, a similarity coefficient being a measure of the closeness of any two variables in the data. Phillips claimed that the "Euclidean metric" must be used if Ward's method is to produce meaningful results, citing Wishart as his authority (Phillips, 1985: 83). However, if Ward's Method is selected in *SPSS*, the software displays a warning message if any measure other than Squared Euclidean Distance is selected. Phillips has no mention of a squared Euclidean measure. There is some confusion in the literature on this measure. Everitt recommends that Euclidean distance be used, but contrasts this with the use of squared Euclidean distance by Wishart (Everitt, 1993: 67). It appears that the Euclidean measure has two variants. Due to this, I decided to use both Euclidean and Squared Euclidean distance measures and compare the results. There was no noticeable difference in the number or composition of clusters produced by the two distance measures, and so I elected to use squared Euclidean distance as advised by *SPSS*.

Nevertheless, it seems to me that the use of cluster analysis procedures in different software implementations to cluster collocational data is a further instance where the claim that, if you run the same data through the same computer program, you will get the same output each time (Stubbs, 2001: 140), appears to be somewhat naive, and quite probably false.

4 Results and Discussion

One dimension of comparison between the results that Phillips achieved in 1985 and my own study is the number and composition of lexical networks found in each chapter of the text analysed. Overall, the two studies showed similar results. Table 7 presents the total number of lemmas analysed, the number of clusters identified and the percentage of lemmas included in the clusters for each chapter in ELEN, for both the 1985 study and the present (2003) study.

ELEN Chapter	Phillips 1985			Doyle 2003		
	Total Lemmas	Clusters	% Lemmas Clustered	Total Lemmas	Clusters	% Lemmas Clustered
1	71	11	63.4	67	14	65.7
2	92	22	81.5	88	25	86.4
3	88	14	72.7	88	18	63.6
4	88	15	75.0	80	18	82.5
5	80	15	61.3	73	17	65.3
6	72	12	72.2	69	17	71.0
7	74	18	74.3	71	20	73.2

Table 7 Comparison of Lemmas Clustered by Two Studies.

There is a clear tendency for my analysis to produce more clusters than that of Phillips (1985), and in that respect the analytical technique used appears to be more ‘conservative’.

Taking a closer look, we can focus on the analysis of lemmas in Chapter 1 of the text ELEN and compare the lemmas in each cluster between the two studies. Table 8 presents Phillips’ data and mine for Chapter 1 of ELEN. (Figure A-2 in the Appendix reproduces the dendrogram output from SPSS from which the number of clusters can be identified). Phillips submitted 71 lemmas for cluster analysis, while I was able to submit only 67 of these after following his restrictions on threshold frequencies. Phillips records 11 clusters containing 45 lemmas, whereas I found 14 clusters containing 44 lemmas. A high degree of similarity is seen in these results, but they are clearly not identical. The lemmas gathered in Phillips’ two largest clusters (clusters **5** and **11**) are represented in the present study by a greater number of clusters: **e** and **f**, and **k**, **l** and **m**.

ELEN1 Phillips 85	**ELEN2** Doyle 03
1 capacity shunting	**a** capacity shunting
2 half expected	**b** half power
3 signal unwanted	**c** signal unwanted
4 low frequency case high	**d** low frequency case high circuit
5 data factors make following parts containing denominator inductors	**e** denominator expression data make factor
	f followed part containing inductor transistor
6 similarly figure power decibel	**g** decibel know

ELEN1 Phillips 85	**ELEN2** Doyle 03
7 respectively say	**h** respectively say
8 parallel resistor	
9 effective rise give	**i** effective rise give
10 derive find expression seen	**j** derive see find
11 large value capacitor need first circuit know input amplifier component voltage related	**k** large value capacitor need
	l first amplifier input voltage
	p component related
	q equal short device

Table 8 Comparison of Clusters Identified in Chapter 1 of ELEN.

It is clear from this table that certain key lemmas have 'drifted' from one cluster to another: for example, circuit was found to be associated with the lemmas shown in cluster **11** by Phillips (1985), but in the present study, it has joined the lemmas shown in cluster **d**. Similarly, know, also from cluster **11**, has joined cluster **g** and expression from cluster **10** has joined cluster **e**.

These results are typical of those for all seven chapters of ELEN (a full discussion can be found in Doyle, 2003). What they clearly illustrate is that running the same data through the same algorithms gives very similar but not identical results. The drifting of a few lemmas from one cluster to another shown in Table 8 implies that the lexical networks identified in the original study (Phillips, 1985) may not actually be 'there' in the text but, instead, an artefact of the cluster analysis technique. Given that one of the common warnings in most introductory texts on cluster analysis is that the technique tends to produce clusters regardless of the relationships that may or may not exist in

the data, this must be a major concern for corpus linguists using clustering techniques on collocational data.

5 Conclusion

Several linguists have written of the value of an empirical linguistics (Stubbs, 2001; Sampson, 2002), and some have also commented that the benefit of a data-oriented approach lends scientific credibility to the description of language. But how 'scientific' is most corpus linguistics work if so few replication studies are published?

In this paper, I have tried to show that the apparently simple matter of running the same data through the computer again does not inevitably produce the same results. This I believe calls into question the certainties with which some results are presented under the guise of 'scientific' rigour in the field of corpus linguistics. We must at least acknowledge that replication needs to be further explored in relation to the basic methods of corpus based enquiries into the nature of language.

Appendix A

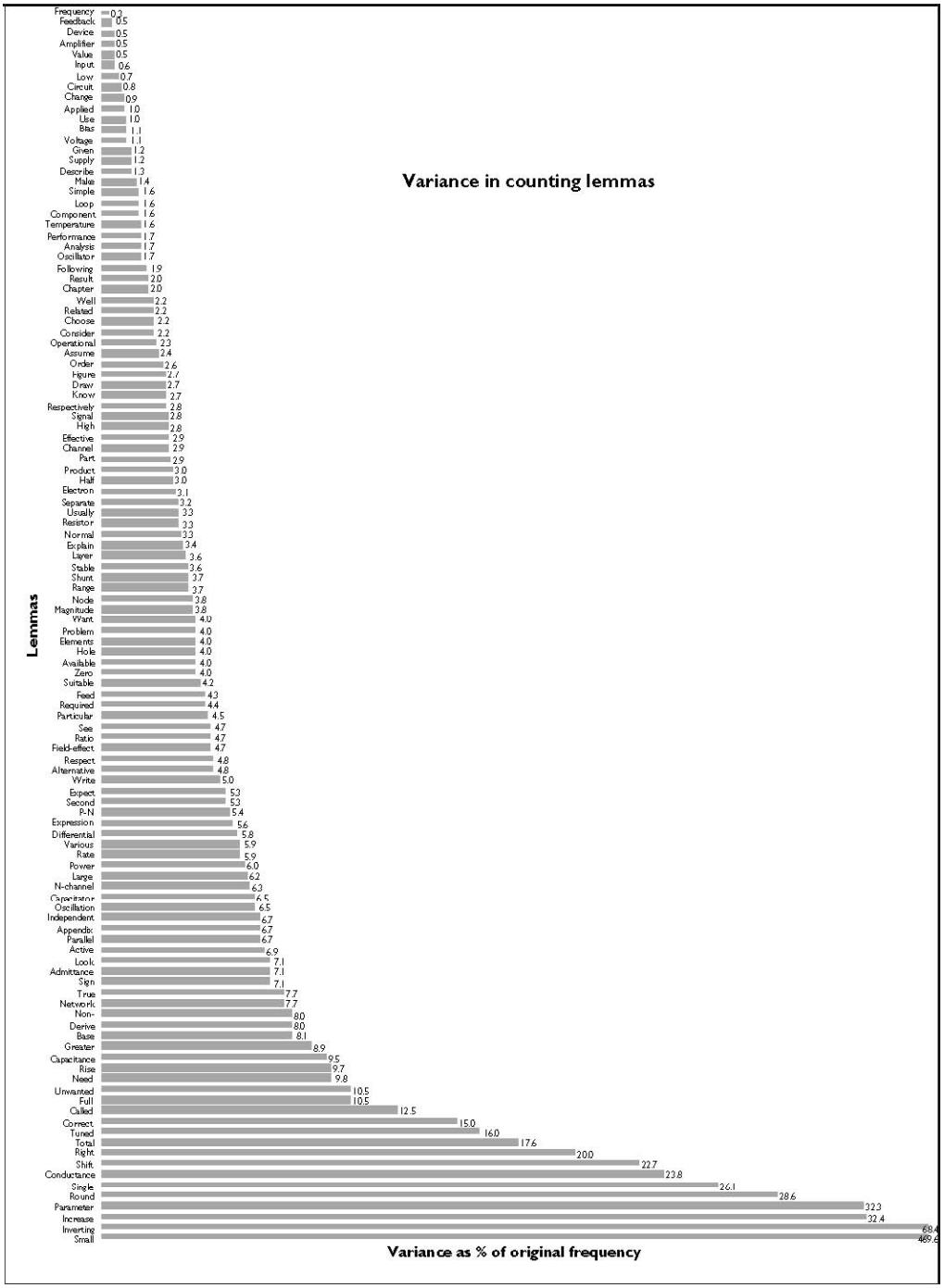

Figure A-1 Variance Between Frequency Counts for 117 Lemmas As Percentage of Original Count Noted By Phillips,1985 (Doyle, 2003: 173).

Lemmas

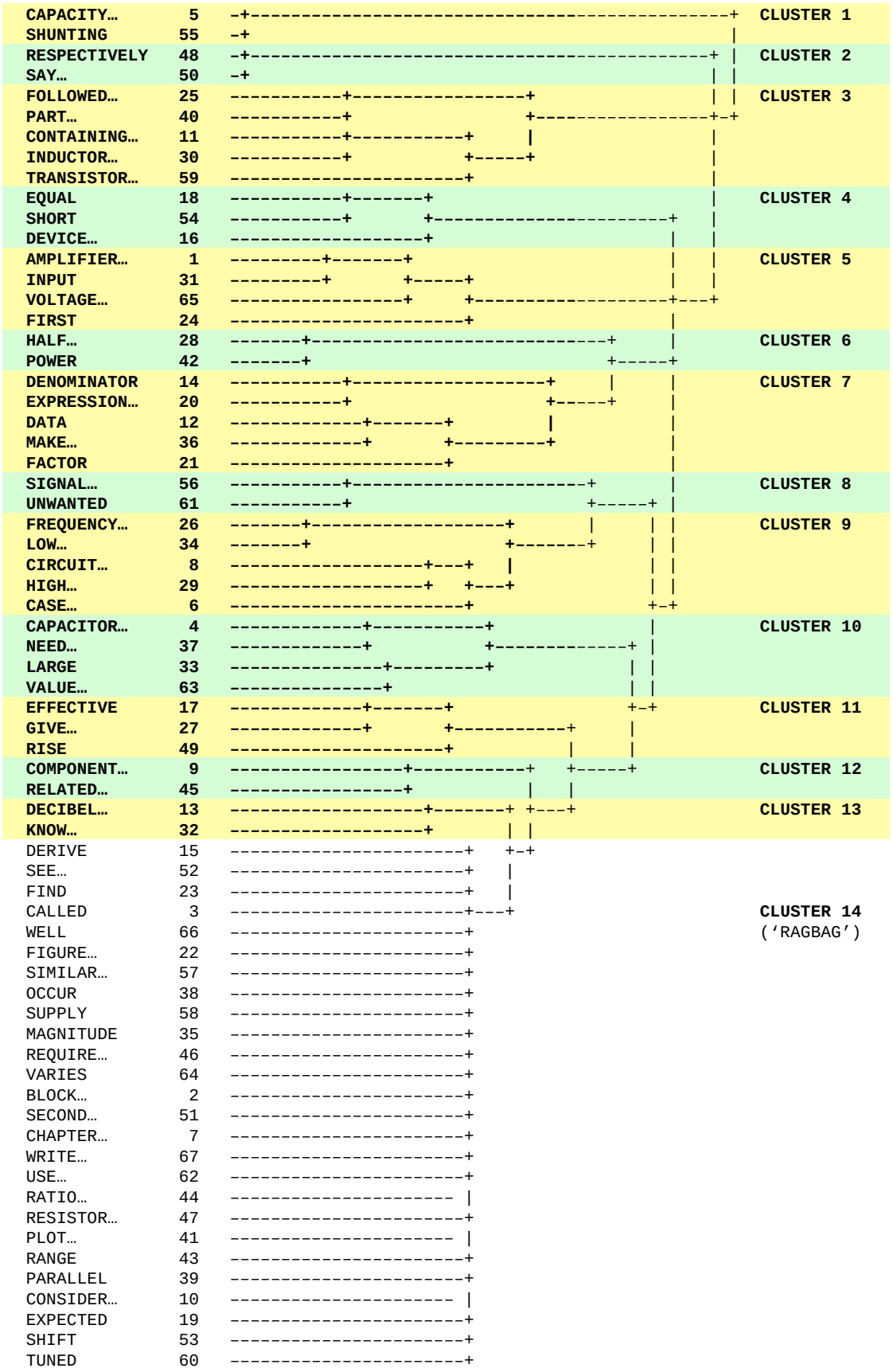

Figure A-2 Dendrogram of Cluster Analysis of 67 Lemmas in Chapter 1 ELEN2 (Doyle, 2003: 321).

References

Biber, D. (1988) *Variations Across Speech and Writing* (Cambridge: Cambridge University Press).

Collins, H. and Scott, M. (1996) Lexical landscaping in business meetings, DIRECT Paper 32. Available online from http://www.liv.ac.uk/~tony1/direct.html (accessed March 1st 2004).

Doyle, P. (2003) *Replicating Corpus Linguistics: A Corpus-Driven Investigation of Lexical Networks in Text* (Unpublished thesis, Lancaster University).

Everitt, B. S. (1993) *Cluster Analysis* (London: Edward Arnold).

Kittredge, R. and Lehrberger, J. (eds.) (1982) *Sublanguage: Studies of language in Restricted Semantic Domains* (Berlin: Walter de Gruyer).

Lee, D. (2000) *Modelling variation in spoken and written language: the multidimensional approach revisited* (Unpublished thesis. Lancaster University).

Leech, G. (1992) Corpus Linguistics and Theories of Linguistic Performance, in J. Svartvik (ed.) *New Directions in Corpus Linguistics* (Berlin: Mouton De Gruyter), 105-134.

McEnery, T. (1996) *Corpus Linguistics* (Edinburgh: Edinburgh University)

Phillips, M. (1985) *Aspects of Text Structure: An Investigation of the Lexical Organisation of Text* (Amsterdam: North Holland).

Phillips, M. (1989) *Lexical Macrostructure in Text* Birmingham Monographs, 9 (University of Birmingham)

Sampson, G. (2002) *Empirical Linguistics* (London: Continuum).

Scott, M. (1996) *WordSmith Tools* (Oxford: Oxford University Press).

Sinclair, J. M. (1991) *Corpus, Concordance, Collocation* (Oxford: Oxford University Press.

Sinclair, J.M. (1987) *Looking Up* (Cambridge: Cambridge University Press).

Stubbs, M. (2001) *Words and Phrases* (Oxford: Blackwell).

Titscher, A., Meyer, C., Wodak, R. and Vetter, D. (2000) *Text and Discourse Analysis* (London: Routledge).

Williams. G. C. (1998) Collocational networks: interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics* 3, 1, 151-171.

Wishart, D. (2003) Critique of k-Means Clustering Technique. Webpage online at http://www.clustan.com/k-means_critique.html (accessed June 22 2005).

Wishart, D. (2001) k-Means Clustering with Outlier Detection, Mixed Variables and Missing Values, in *Proceedings of GfKl 2001*.