Do we Need Statistics when we have Linguistics?

(Precisamos de Estatística quando temos a Lingüística?)

Pascual Cantos Gómez
(Universidad de Murcia)

Abstract: *Statistics is known to be a quantitative approach to research. However, most of the research done in the fields of language and linguistics is of a different kind, namely qualitative. Succinctly, qualitative analysis differs from quantitative analysis is that in the former no attempt is made to assign frequencies, percentages and the like, to the linguistic features found or identified in the data. In quantitative research, linguistic features are classified and counted, and even more complex statistical models are constructed in order to explain these observed facts. In qualitative research, however, we use the data only for identifying and describing features of language usage and for providing real occurrences/examples of particular phenomena. In this paper, we shall try to show how quantitative methods and statistical techniques can supplement qualitative analyses of language. We shall attempt to present some mathematical and statistical properties of natural languages, and introduce some of the quantitative methods which are of the most value in working empirically with texts and corpora, illustrating the various issues with numerous examples and moving from the most basic descriptive techniques (frequency counts and percentages) to decision-taking techniques (chi-square and z-score) and to more sophisticated statistical language models (Type-Token/Lemma-Token/Lemma-Type formulae, cluster analysis and discriminant function analysis).*

Key-words: *Quantitative analysis; Statistics; Language modelling; Linguistic corpora.*

Resumo: *A estatística é conhecida por ser uma abordagem quantitative de pesquisa. No entanto, a maioria da pesquisa feita nos campos da linguagem e da lingüística é de natureza diferente, qual seja, qualitativa. De modo sucinto, a análise qualitativa difere da quantitativa pelo fato de a primeira não é feita tentativa de atribuir freqüências, porcentagens e outros atributos semelhantes, às características lingüísticas encontradas ou identificadas nos dados. Na pesquisa quantitativa, as características lingüísticas são classificadas e contadas, e modelos estatísticos mais complexos ainda são* *construídos a fim de explicar os fatos observados. Na pesquisa qualitativa, contudo, usamos os dados apenas para identificar e descrever características da linguagem em uso e para fornecer exemplos / ocorrências reais de um fenômeno particular. Neste trabalho, tentaremos mostrar como métodos quantitativos e técnicas estatísticas podem suplementar análises qualitativas da linguagem. Nós tentaremos apresentar alguns métodos quantitativos que são de grande valor para trabalhar empiricamente com textos e com corpora, ilustrando diversas questões com vários exemplos, passando das técnicas mais básicas de descrição (contagem de freqüência e porcentagens) para técnicas de tomada de decisão (qui-quadrado e z-score) e para modelos lingüístico-estatísticos mais sofisticados (fórmulas de Forma-Ocorrência / Lema-Ocorrência / e Lema-Forma, análise de cluster e discriminant function analysis.)*

Palavras-chave: *Análise quantitativa; Estatística; Modelagem lingüística; Corpora lingüísticos.*

1. Introduction

The title itself is the reverse of Hatzivassiloglou's[1]. As a statistician, he discussed whether linguistics knowledge could be of any help and contribute to a statistical word grouping system. Our aim here is the opposite: to try to illustrate with numerous examples how quantitative methods can most fruitfully contribute to linguistic analysis and research. In addition, we do not intend here to offer an exhaustive presentation of all statistical techniques available to linguistics, but to demonstrate the contribution that statistics can and should make to linguistic studies.

Among the linguistic community, statistical methods or more generally quantitative techniques are mostly ignored or avoided because of the lack of training, fear and dislike too. The reasons: (1) these techniques are just not related to linguistics, philology or humanities; statistics falls into the province of sciences, mathematics and the like; and/or (2) there is a feeling that these methods may detroy the "magic" in literary text.

[1] Haztivassiloglou, V. (1994) "Do we need Linguistics when we have Statistics? A Comparative Analysis of the Contributions of Linguistic Cues to Statistical Word Grouping System". *The Balancing Act. Combining Symbolic and Statistical Approaches to Language.* Ed. J.L. Klavans and P. Resnik. Cambridge: The MIT Press. 67-94.

George Zipf (1935) was one of the first linguists to prove the existence of statistical regularities in language. His best known law proposes a constant relationship between the rank of a word in a frequency list and the frequency with which it is used in a text. To illustrate this, consider the 30th, 40th, 50th, 60th and 70th most-frequently occurring words taken from a sample of the *Corpus Collection B* (published by Oxford University Press): all the values (*constants*) come out at around 20,000 (see *Table 1*). This is because the relationship between *rank* and *frequency* is inversely proportional. In addition, Zipf thought that the *constants* are obtained regardless of subject matter, author or any other linguistic variable.

Word	**Rank**	*	**Frequency**	=	**Constant**
were	30	*	700	=	21,000
she	40	*	538	=	21,520
them	50	*	402	=	20,100
been	60	*	329	=	19,740
did	70	*	275	=	19,250

Table 1: Zipf's law on the relationship between *rank* and *frequency*

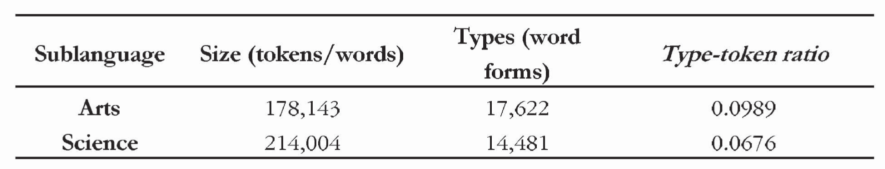
Similarly, another Zipf law showed the inverse relationship between word length and its frequency. In some languages, such as English, for example, the most commonly used words are monosyllabic ones. This effect seems to account for our tendency to abbreviate words whenever their frequency of use increases, i.e. the reduction of 'television' to 'TV' or 'telly'. It would also seem to be an efficient communication principle to have the popular words short and the rare words long.

These examples show how some linguistic patterns are regular and independent of speaker, writer, or subject matter, and how linguistic behaviour conforms closely to some expectations: quantitative or statistical patterns. In the next section, we shall try to exploit the most basic descriptive data: *frequencies*, and illustrate some potential applications to linguistic research.

2. Frequency counts

A preliminary survey of a text or linguistic corpus is to produce a frequency list of its items (tokens, types or lemmas[2]). At it simplest, the frequency list shows the types that make up the text(s) or corpus, together with their instances of occurrence. It can be produced in several different sequences[3]. Despite its simplicity, it is a very powerful tool. So for example, frequency lists of huge corpora enable lexicographers to take important decisions on which words a dictionary should include and which particular meanings. Similarly, authors of L2-materials might use this data to decide which words, phrases, expressions or idioms are most relevant in teaching an L2. This evidence of usage is without any doubt a unique and most important source for any enterprise in language description.

It is not just general language, but also sublanguages, that is, specific varieties of language used in certain communicative domains, such as business, medicine, sports, etc, or the study of genders or specific authors that can profit from frequency list analysis. This analysis can, for instance, shed some new light on stylistic variation issues regarding such diverse writers as Henry James and George Orwell, to name but two. To illustrate this, let us compare two sublanguages: arts (literary criticism) and science (biology). The table below (*Table 2*) summarises the output of a lexicon extraction program, showing the size of the lexicons produced for each corpus and giving the *type-token ratio*[4] for each.



Sublanguage	**Size (tokens/words)**	**Types (word forms)**	***Type-token ratio***
Arts	178,143	17,622	0.0989
Science	214,004	14,481	0.0676

Table 2: Arts and science corpus samples

[2] For a better understanding of the terms *tokens, types* and *lemmas*, consider the following word sequence: *sings, singing, sang, sing, sings, sung, singing, sung* and *sang*, where we have nine words or **tokens**, five different word forms or **types** (*sing, sings, singing, sang* and *sung*) and a single base form or **lemma**, namely *sing*.

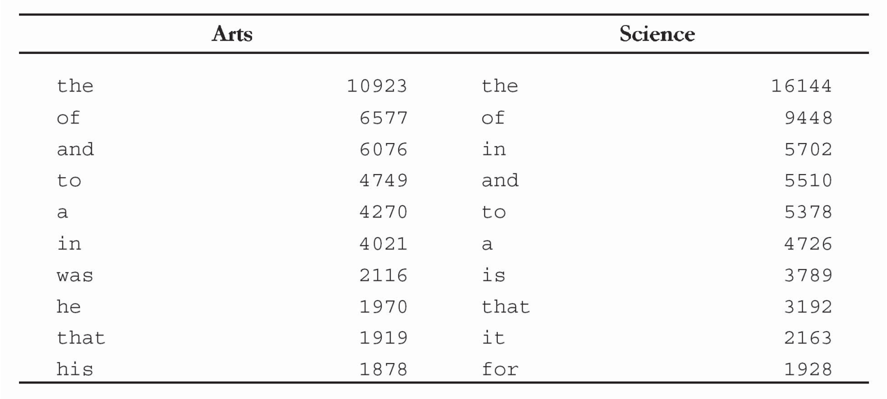
[3] For an exhaustive typology of frequency lists see, among others, Cantos (1995).

[4] The *type-token ratio* is the quotient obtained when dividing the total amount of *types* by the *token* total. See also footnote *23*.

The *type-token ratio* is an extremely valuable indicator here[5]. It shows that, although the two language samples are different in size (the science sample has 35.861 tokens more; it is therefore 16.75 % larger), we find, on average, almost ten new items for every one hundred words (tokens) in the arts sublanguage (*type-token ratio* = 0.0989 * 100 = 9.89 ≈10), whereas its counterpart sublanguage offers only six (0.0676 * 100 = 6.76). Furthermore, on the basis of this evidence, it seems that the sublanguage of science reaches a higher degree of lexical closure than its arts counterpart.

This preliminary approach shows that the sublanguage of arts is, on the whole, richer and more varied regarding the use of different vocabulary items. It resorts to more different words forms (types) and its lexical closure is also more difficult to establish. In contrast, the science sublanguage seems less varied lexically speaking and so, in lexical terms, it would appear that, on the basis of the evidence presented, it tends very strongly towards premature closure, whereas the other sublanguage does not.

Another interesting finding of frequency lists relates to lexical selection, or determining, by means of the evidence of usage, which are the most frequent or relevant items of a particular sublanguage, author, etc. Regarding our two sublanguages (science and arts) we obtain the following top 10 items (*Table 3*).

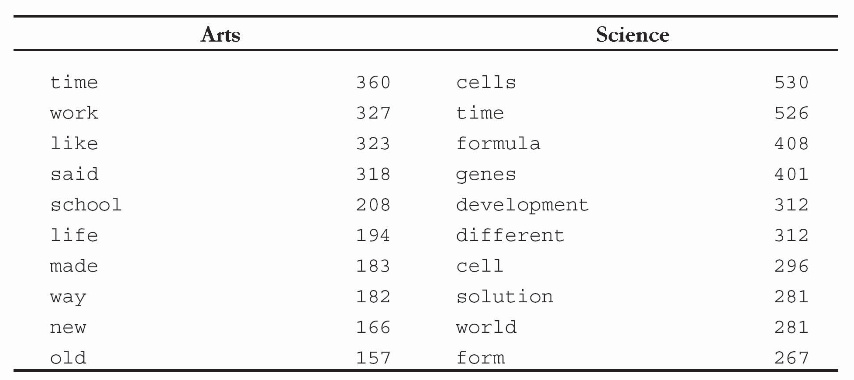


Arts		Science	
the	10923	the	16144
of	6577	of	9448
and	6076	in	5702
to	4749	and	5510
a	4270	to	5378
in	4021	a	4726
was	2116	is	3789
he	1970	that	3192
that	1919	it	2163
his	1878	for	1928

Table 3: The top 10 words used in arts and science

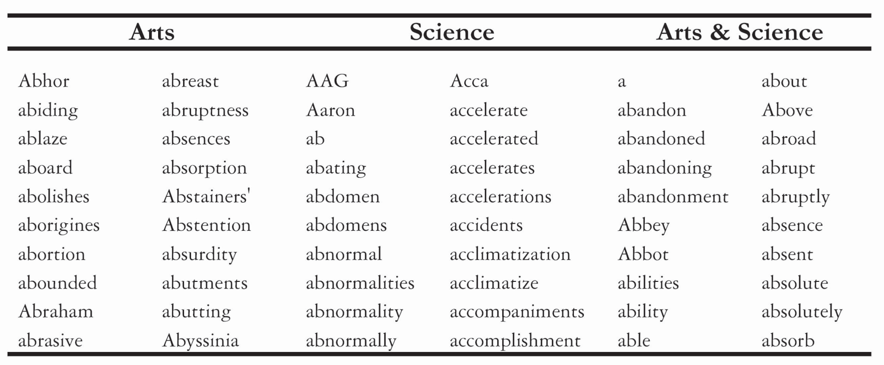
[5] See Cantos (2000) for a detailed discussion on the limitations and drawbacks of the *type-token ratio*; see also McKee, Richards and Malvern (2000), Chipere et al. (2001) and Baayen (2001:4-5). Interesting is also Scott's notion of "standardised *type-token ratio* (1996).

The main problem with this information is that the use of raw frequencies highlights the very common words such as *the, of*, *in*, etc., despite the fact that their comparatively high frequencies of occurrence are unlikely to provide conclusive evidence of any specifically used vocabulary in any sublanguage. These are words that, on the basis of frequency of occurrence alone, would be found to occur within most sublanguages, and it can perhaps be read more usefully if the purely grammatical words (close-class items) are discarded. This leaves us with (*Table 4*):



Arts		Science	
time	360	cells	530
work	327	time	526
like	323	formula	408
said	318	genes	401
school	208	development	312
life	194	different	312
made	183	cell	296
way	182	solution	281
new	166	world	281
old	157	form	267

Table 4: The top 10 content words used in arts and science



Arts		Science		Arts & Science	
Abhor	abreast	AAG	Acca	a	about
abiding	abruptness	Aaron	accelerate	abandon	Above
ablaze	absences	ab	accelerated	abandoned	abroad
aboard	absorption	abating	accelerates	abandoning	abrupt
abolishes	Abstainers'	abdomen	accelerations	abandonment	abruptly
aborigines	Abstention	abdomens	accidents	Abbey	absence
abortion	absurdity	abnormal	acclimatization	Abbot	absent
abounded	abutments	abnormalities	acclimatize	abilities	absolute
Abraham	abutting	abnormality	accompaniments	ability	absolutely
abrasive	Abyssinia	abnormally	accomplishment	able	absorb

Table 5: Extract of arts-only items, science-only items and shared items

	Arts (arts – science = only arts)	**Science (science – arts = only science)**	**Arts and Science (arts 1 science)**
Types	10.622	7.481	7.000

Table 6: Summary of arts-only items, science-only items and shared items

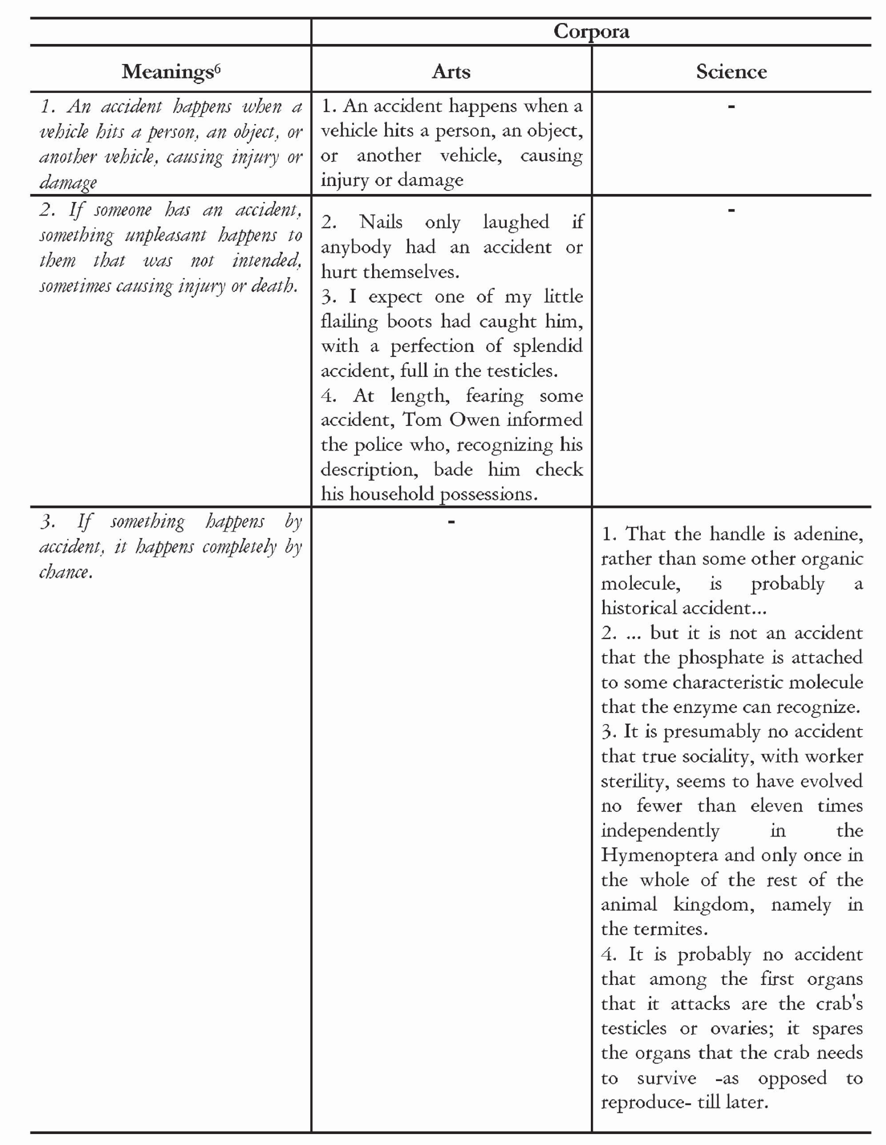
The specific vocabularies of both sublanguages become immediately apparent, and we note striking differences. Interesting, however, is the coincidence on the highly frequent use of *time* in both sublanguages (360 in arts and 526 in science).

A further type of study, using raw frequency lists, could be establishing lexical items exclusively used in each sublanguage and those used in both sublanguages (*Tables* 5 and 6).

From a lexicographical and semantic point of view it could be interesting to investigate the shared items (see the case of *time* above). An inevitable starting hypothesis is that the same words used in different contexts are likely to bear different meanings. As an example we can examine the use of the singular noun *accident.* This type occurs four times in each corpus. Here are the full sentences of their occurrences and their distribution according to sublanguage and meaning (*Table* 7).

This very brief but revealing lexical analysis confirms our initial hypothesis that the same words used in different contexts are likely to carry different meanings. We see how the use of *accident* in scientific communication is restricted to a single meaning (*3. If something happens by accident, it happens completely by chance*), compared with the other two meanings of the noun occurring in the arts corpus (*1. An accident happens when a vehicle hits a person…causing injury or damage*; *2. If someone has an accident, something unpleasant happens …causing injury or death*).

The merits of this apparently simple technique -frequency listing- do not end here. It is, in our opinion, a potentially non-exhaustible resource and an excellent starting point for descriptive linguistic research, as it sometimes turns the invisible into the visible. The close observation of a frequency list may be the first step for the formulation of a hypothesis.



	Corpora	
Meanings[6]	**Arts**	**Science**
1. An accident happens when a vehicle hits a person, an object, or another vehicle, causing injury or damage	1. An accident happens when a vehicle hits a person, an object, or another vehicle, causing injury or damage	-
2. If someone has an accident, something unpleasant happens to them that was not intended, sometimes causing injury or death.	2. Nails only laughed if anybody had an accident or hurt themselves. 3. I expect one of my little flailing boots had caught him, with a perfection of splendid accident, full in the testicles. 4. At length, fearing some accident, Tom Owen informed the police who, recognizing his description, bade him check his household possessions.	-
3. If something happens by accident, it happens completely by chance.	-	1. That the handle is adenine, rather than some other organic molecule, is probably a historical accident... 2. ... but it is not an accident that the phosphate is attached to some characteristic molecule that the enzyme can recognize. 3. It is presumably no accident that true sociality, with worker sterility, seems to have evolved no fewer than eleven times independently in the Hymenoptera and only once in the whole of the rest of the animal kingdom, namely in the termites. 4. It is probably no accident that among the first organs that it attacks are the crab's testicles or ovaries; it spares the organs that the crab needs to survive -as opposed to reproduce- till later.

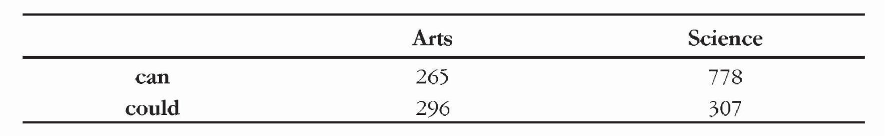
Table 7: Sense distribution of *accident* in the arts and science corpus

[6] Taken from the *Collins Coubild English Dictionary* (Sinclair et al. 1995).

3. Significance testing

As mentioned earlier, frequency lists of word forms are never more than a set of hints or clues to the nature of a text. By examining a list, one can get an idea of what further information would be worth acquiring before starting an investigation.

Returning to our arts and science sample corpora, let us suppose now that we are interested in examining how two modal verbs, *can* and *could,* are distributed in both sublanguages and compare their usage. The first thing to do is to make a simple count of each of these modals in the two corpora. Having done this, we arrive at the following frequencies (*Table 8*):



	Arts	**Science**
can	265	778
could	296	307

Table 8: Frequency distribution of *can* and *could* in arts and science

A quick look at these figures reveals that *can* and *could* are more frequently used in scientific communication than in arts. But with what degree of certainty can we infer that this is a genuine finding about the two corpora rather than a result of chance? We cannot decide just by looking at these figures; we need to perform a further calculation: a test of *statistical significance* and determine how high or low the probability is that the difference between the two corpora on these features is due to chance.

3.1. Chi-squared test

Among the various significance tests available to linguists, we find: the *chi-squared test*, the *t-test*, *Wilcoxon's rank sum test*, etc. The *chi-squared test* is probably the most commonly used one in corpus linguistics, as it has numerous advantages for linguistic purposes (McEnery and Wilson 1996: 70): (a) it is more accurate than, for example, the *t-test*; (b) it does not assume that the data is *normally distributed*[7] (quite frequent with linguistic data); (c) it is easy to calculate, even without a computer statistics package; and (d) disparities in corpus size are unimportant.

Probably, the main disadvantage of *chi-square* is that it is unreliable with very small frequencies (less than 5). Succinctly, *chi-square* compares the difference between the actual observed frequencies in the texts or corpora, and those frequencies that we would expect (if the only factor operating had been chance). The closer the expected frequencies are to the observed frequencies, the more likely it is that the observed frequencies are a result of chance. However, if the difference between the observed frequencies and the expected ones are greater, then it is more likely that the observed frequencies are being influenced by something other than chance. For instance, if we take our example, a significant difference between the observed frequencies and the expected ones of *can* and *could* would mean a true difference in the grammar or style of the two domain languages: arts and science.

The first step is to determine the significance level or threshold of tolerance for error[8]. In linguistic issues, it is common use to fix the probability of error threshold of 1 in 20, or $p < 0.05$. Remember that *chi-square* compares what actually happened to what hypothetically would have happened if all other things were equal. The first thing to do is to calculate the column and row totals, giving (*Table 9*):

	Arts	**Science**	**Total**
can	265	778	1043
could	296	307	603
Total	561	1085	1646

Table 9: Frequency distribution (with column and row totals) of *can* and *could* in arts and science

[7] In a theoretical normal distribution, the mean (the sum of all scores divided by the total number of scores), the median (the middle point or central score of the distribution) and the mode (the value that occurs most frequently in a given set of scores), all three, fall at the same point: the centre or middle (mean = median = mode). Additionally, if we plot graphically the data we get is a symmetric bell-shaped graph.

[8] The probability that rejecting the null hypothesis (whenever the difference is not significant) will be an error.

Next, the expected frequencies are calculated. This is done by multiplying the cell's row total by the cell's column total, divided by the sum total of all observations. So, to derive the expected frequency of the modal verb *can* in the arts corpus, we multiply its cell row total (1043) by its cell column total (561) and divide that product by the sum total (1646):

$$\frac{1043 * 561}{1646} = 355.48$$

All the calculations of the expected frequencies of each cell are shown below (*Table 10*):

	Arts	Science	Total
can (observed)	265	778	1043
can (expected)	355.48	687.51	
could (observed)	296	307	603
could (expected)	205.51	397.48	
Total	561	1085	1646

Table 10: Observed data *versus* expected data for *can* and *could* in arts and science

Now, we need to measure the size of the difference between the pair of observed and expected frequencies in each cell. This is done with the formula[9]:

$$\frac{(O-E)^2}{E}$$

Where O = observed frequency and E = expected frequency. So, for instance, the difference measure for *can* (in the arts corpus) is:

$$\frac{(265-355.48)^2}{355.48} = 23.02$$

[9] Note that squaring the difference ensures positive numbers.

Next we calculate the difference measure for all cases (*can* in the art corpus, *can* in the science corpus, *could* in the art corpus and *could* in the science corpus), and add all these measures up. The value of *chi-square* is the sum of all these calculated values. Thus, the formula for *chi-square* is as follows:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

For our data above, this results in a total *chi-square* value of: 95.35. Having done this, it is necessary to look at a set of statistical tables to see how significant our *chi-square* value is. To do this one first requires a further value: the number of *degrees of freedom* (*df*)[10]. This is very simple to work out:

*df = (number of columns in the table - 1) * (number of rows in the table - 1).*

Which is for our case:

df = (2 –1) * (2 –1) = 1

We now look in the table of *chi-square* values[11] in the row for the relevant number of degrees of freedom (1 *df*.) and the appropriate column of significance level (*0.05* in linguistics). Returning to our example, we

			Significance level			
Df	**0.20**	**0.10**	**0.05**	**0.025**	**0.01**	**0.001**
1	1.64	2.71	3.84	5.02	6.64	10.83
2	3.22	4.61	5.99	7.38	9.21	13.82
3	4.64	6.25	7.82	9.35	11.34	16.27
4	5.99	7.78	9.49	11.14	13.28	18.47
5	7.29	9.24	11.07	12.83	15.09	20.52

Table 11: Extract of a *chi-square* distribution (*1-5 df*)

[10] This is a technical term from mathematics which we shall not attempt to explain here. For some non-technical and easy accessible explanations see Woods, Fletcher and Hughes (1986: 138-9) and/or Brown (1988: 118-9).

[11] The *chi-square* distribution tables can be found in the appendices of most statistics book/ manual, see for instance Oakes (1998: 266).

have a *chi*-value of 95.35 with $df = 1$, so according to the distribution table, we would need our *chi*-value to be equal to or greater than 3.84 (see *Table 11*), which is true. This means that the difference found between the two sublanguages regarding the use of *can* and *could* is statistically significant at $p < 0.05$, and we can therefore, with quite a high degree of certainty, say that this difference is not due to chance, but due to a true reflection of variation in the two sublanguages.

3.2. Z-score

Up to this point we have been dealing just with single items without paying attention to the co-text, that is the words on either side of the item under investigation. These co-textual words are known as collocations. Collocation has been defined as "the occurrence of two or more words within a short space of each other in a text" (Sinclair 1991: 170).

Concordance lines[12] hold the primary information needed for collocation analysis. Approaches for the extraction of collocations may range from simple frequency information for words that occur near to the keyword, to the application of sophisticated statistical techniques, which calculate the figures needed for the comparison, and use them to produce measures of significance.

The most basic and naïve form of collocation analysis provided by concordance packages is to produce a frequency list of all the words occurring within predetermined proximity limits (*span*). As an example, we extracted the occurrences of the lemma KEEP (*keep, keep'em, keeping, keeps, kept*) in the arts corpus, by means of the concordance program *Monoconc*[13], using a fixed span of six words on either side of the keyword (see *Appendix 1*). However, as already mentioned, the major drawback with the data obtained is that the use of raw frequencies highlights the very common words, despite the fact that they are unlikely to provide conclusive evidence of significant collocation patterns. So, we discarded the purely grammatical words (close-class items) and also deleted low frequency words, leaving just those that co-occur at least three times. This would leave us with just:

[12] See *Appendix 3*.

[13] *MonoConc* v. 1.5. (Athelstan Publications; http://www.athel.com/mono.html#mono).

going	*5*
hands	*4*
feet	*3*
house	*3*
pony	*3*
well	*3*

All of these could form fairly strong patterns with KEEP and would be worth investigating further.

This approach, though useful, is very simple. It just offers some quantitative data and is only the starting point for calculating their significance. The calculation of the data needed for collocation analysis is not very complicated, although several alternative methods are available. The starting point for any collocation analysis is a set of concordance lines for the words under investigation, long enough to contain the required span of words. The first decision to make is to choose an appropriate length of the span. Let us consider the following concordance line:

```
answering their questions and trying to keep a conversation going while cooking a
   -6       -5       -4      -3     -2    -1 node +1     +2         +3    +4       +5   +6
```

The word under investigation (KEEP) is referred to as the *node* and is used to generate the concordance lines. The words around the node are numbered according to the position to the node. Those words left to the node are expressed negatively, those to the right positively. So the span here is of twelve words, six words on either side. This set of concordance lines offers the basis for any further significance technique. All three major collocation significance tests, namely *z-score*, *t-score* and *mutual information*[14], rely on actual or observed frequency and expected frequency. So once the concordance lines have been obtained, we need to establish the actual frequency of the words within the span. In other words, we produce a

[14] There are important differences between the information provided by these three measures: more, perhaps, between *t-score* and the other two than between *z-score* and *mutual information* themselves. It is difficult, if not impossible, to select one measure that provides the best possible assessment of collocations, although there has been ample discussion of their relative merits (see, for example, Church et al. 1991; Clear 1993; or Stubbs 1995).

frequency list of the concordance lines, similar to the one already discussed and may, alternatively, also eliminate close-class items. Suppose this is what we get:

going	*5*
hands	*4*
feet	*3*
house	*3*
pony	*3*
well	*3*

At this point we need to calculate the expected frequency figures that will be compared to the actual frequencies to assess their significance. The calculation of the expected frequency for words occurring in the span is straightforward. First, we need a theoretical language model (i.e. a representative language sample or corpus of the language or domain that we want to investigate). This model will help us to predict how these words would be distributed if there were no particular pattern of collocation between them and the node. In other words, if we want to check whether the node (KEEP) is exercising some influence over the distribution of *going*, *hands*, *feet*, *house*, *pony* and *well*, we need to know how we would expect these words to behave in the absence of that influence.

Let us return to our frequency list of occurring words with KEEP within the span limit of 12, where the verb form *going* occurs 5 times (observed frequency). This means that *going* appears 5 times in the proximity of KEEP in a text sample of overall:

12 * 5 = 60 tokens (words)

On the other hand, its overall frequency in the entire corpus (arts subcorpus of *Corpus Collection B*), which consists of 178,143 tokens, altogether, is 89. If *going* is randomly distributed throughout the text, then its expected frequency in any 60 token text sample should be:

$$\frac{89}{178.143} * 60 = 0{,}029$$

That is, the expected frequency of *going* in any random selected sample of 60 tokens should be just 0.029 compared with its real observed 5 occurrences in the set of lines of KEEP. Though the difference is huge, we cannot decide anything yet just from these figure; we need to perform a further calculation: a test of *statistical significance* and determine how high or low the probability is that the difference between the observed and expected frequency is due to chance.

The *z-score* is probably the most familiar of the statistical significance measures used for collocation analysis. The calculation is reasonably easy:

$$z = \frac{O - E}{sd}$$

Where O = observed frequency, E = expected frequency and sd = standard deviation[15]. O is straightforward, E needs to be calculated, as explained previously, and sd uses the formula:

$$sd = \sqrt{N * (p * (1 - p))}$$

And where p = probability of occurrence of the co-occurring word in the whole text, and N = number of tokens in the set of concordance lines. Thus, the probability of *going* occurring in the whole text is:

$$p = \frac{89}{178.143} = 0{,}0005$$

And N (number of tokens in the truncated concordance lines) is:

N = *number of concordance lines* * *span*

$N = 5 * 12 = 60$

[15] The standard deviation provides a sort of average of the differences of all scores from the mean.

So, the *sd* can now be calculated:

$$sd = \sqrt{60 * (0{,}0005 * (1 - 0{,}0005))}$$
$$sd = 0.17$$

and its *z-score* gives:

$$z = \frac{5 - 0{,}029}{0{,}17}$$
$$z = 29.24$$

A useful cut-off measure for significance in this type of test is around 3 (Barnbrook 1996: 96). This leads us to conclude that the occurrence of *going* within the co-text of the lemma KEEP is not due to chance, but due to some kind of lexical 'attraction' (see *Appendix 2*).

4. Putting together quantitative and qualitative approaches

An important advantage of collocation analysis is that it allows us to focus our attention on specific aspects of the contexts of words already selected for investigation through concordance lines. Collocations can help organize the context into major patterns, and we can use our knowledge of those patterns to assess the general behaviour patterns of the language or the uses being made of particular words within the text[16]. This can make it much easier to identify phrases and idioms, to differentiate among different meanings of a single word form or to determine the range of syntactic features.

[16] See for instance Hunston and Francis' book on a corpus-driven approach to a lexical grammar of English (1999), Cantos (2001) and Cantos and Sánchez's forthcoming article on lexical hierarchical constructions of collocations.

In the examination of the results above the most significant collocations for KEEP were *going, hands, feet, house, pony* and *well*. However, and for practical reasons, we deliberately discarded all grammatical words and low frequency items. For the following analysis, we included close-class items, focusing particularly on prepositions, since they are likely to be relevant when dealing with verbs (think of prepositional or phrasal verbs, idioms, and the like). Collocation frequency data can be very useful in this respect. The following table (*Table 12*) shows the data for the verb lemma KEEP (with a frequency threshold of 3; see *Appendix 3* for concordance lines):

2 – Left		**1 – Left**		**1 - Right**		**2 - Right**	
6	he	23	to	16	the	6	with
3	that	5	he	7	a	3	a
3	and	4	and	7	up	3	of
		3	I	6	his	3	to
		3	would	6	him	3	hands
		3	could	3	in		
		3	of	3	to		

Table 12: Collocation data for KEEP

Given this data, a first approach could be to group the words in column *1-Right* (words that immediately follow KEEP) according to their parts-of-speech; we get:

- Determiners: *the, a* and *his* (there is no instance where *his* is a possessive pronoun, see *Appendix 3*)
- Prepositions: *up, in* and *to*
- Pronouns: *him*

The right hand side association power of KEEP can now be typified as:

- KEEP + *Preposition (up, in, to)*
- KEEP + *Pronoun (him)*
- KEEP + *Determiner (the, a, his)*

A quick glance at the *z-scores* for KEEP (see *Appendix 2*) reveals that the probability for *in* to co-occur with KEEP is quite low, compared with *to* and *up.* The latter two are statistically very significant, particularly *up*. It is difficult to make assumptions here, due to the small sample analysed, but the *z-scores* point to one hypothesis: KEEP + *up* and KEEP + *to* may form lexical units (prepositional verbs or phrasal verbs), such as in:

The jargon they **kept** *up was delicious for me to hear.*

Commentary on form is **kept** *to a minimum and is almost entirely superficial.*

However, *in* seems very unlikely to be part of the verb and is probably part of the prepositional phrase (PP) that follows the verb (see low *z-scores*; *Appendix 2*), as in:

Hoomey knew it was a suggestion for sugarlumps, which he **kept** *in his pocket.*

Regarding determiners, these do not occur alone, they precede or *determine* a noun or noun head, we can go further saying that KEEP is capable to associate on its right hand side noun phrases (NPs) of the type:

- NP → *Pr*; (*You* **keep** *him here, and say your prayers, and all will be well*)
- NP → *Det (Adj) N*; (*My uncle will go on* **keeping** *the horses if we want them.- He* **keeps** *a sleeping bag up there, stuffed behind the old ventilator pipes, and he sleeps in with her.- He was* **keeping** *his feet well out of the way, following where his horse ate*)

The above is true, as KEEP is a transitive verb. Consequently, we could with some degree of certainty say that the high co-occurrence frequency of pronouns and determiners with the verb KEEP is not due to the configuration of any particular phrase but due to the transitivity of KEEP.

Regarding its association with prepositions, we have three prepositions which are directly attached to the verb (*up, in, to*), and three other which occur within a word distance of two (*2-Right*: *with, of, to*). A first hypothesis could be that the co-occurrence of KEEP + *Preposition* (*up, in* or *to)* attracts other prepositions. If we look at the concordance list, this is only true for *up,* which attracts *with* in four out of six occasions, as in:

To **keep** *up with this pace, that is, just to carry out the work that...*

This results into the following syntactic frames for KEEP + *Preposition*:

- KEEP + *up*
- KEEP + *up* + *with*
- KEEP + *to*
- KEEP + *in*

Whereby the first three are very likely to form phrasal verbs or prepositional verbs, as already discussed, but not *in,* which is part of what follows, a PP in this case. In addition, KEEP + *up* + *with*, might be a phrasal prepositional verb.

The three non-directly attached prepositions (*with, of, to*) have different syntactic patterns with respect to those directly attached ones (*up, to, in*). *With, of* and *to* allow another constituent to be placed in between the verb KEEP and the preposition itself; see for instance:

> *... but where one* **keeps** *faith with it by negation and suffering...*
> *One Jew with a pencil stuck behind his ear* **kept** *gesticulating with his hands and…*
> *... that for so long helped to* **keep** *the Jews of Eastern Europe in medieval ghettoes*

The allowed 'intruder' is either a present participle or an NP:

- KEEP + *NP / Present Participle + with*
- KEEP + *NP* + *of*
- KEEP + *NP* + *to*

An interesting syntactic difference among these non-directly attached prepositions is that in the first two instances (KEEP + *NP/Present Participle* + *with* and KEEP + *NP* + *of*), the prepositions are part of a PP. That is, the PP that complements the preceding NP or Present Participle. Whereas in KEEP + *NP* + *to*, the verb seems to form a kind of discontinuous phrase, and the preposition might, therefore, be thought to be part of the verb:

> *…it was insisted that they* **keep** *themselves to themselves...*

We also find the determiner *a* in position *2-Right*, which indicates that whenever KEEP has the pattern:

- KEEP + ... + a...

The determiner introduces an NP, and this might be some hint of transitivity. The transitive instances with this pattern we have detected are the two phrasal or prepositional verbs:

- KEEP + *up* + *a*: … **keeping** *up a house*
- KEEP + *to* + *a*: *Commentary on form is* **kept** *to a minimum...*

However, in the case of KEEP + *in*, the NP that follows is not a direct object but part of an adverbial-PP headed by the preposition *in*:

... which he **kept** *in his pocket*

Finally, the exploration of the right hand side association of KEEP takes us to *hand*. Its high *z-score* (40.51) somehow gives evidence of some kind of idiomatic expression:

- KEEP + *Det(Poss)* + *hands: ... and* **keeping** *his hands away from the ever-questing...*

Let us now analyse the left hand side association. This side can be interpreted straightforwardly and indicates clearly that KEEP can only be a verb. The word forms heading KEEP are:

- The infinitive particle *to*: *You'd be able to* **keep** *order, sir.*
- The pronouns *I* and *he* (subjects of the verb): *I* **kept** *the past alive out of a desire for revenge.*
- The modal verbs *would* and *could*, both require a non-finite verb form to its right: *Some would* **keep** *me trotting round the parish all day.*
- The preposition *of* which requires a non-finite verb form, in instances such as: … *the habit of* **keeping** *themselves to themselves was not easily lost.*

Note that this has been by no means an exhaustive analysis of KEEP collocation patterns and phrases, just a mere illustration of the potential of using quantitative data in combination with qualitative linguistic data. In addition, for the sake of simplicity and clarity, we have deliberately reduced:

(a) the collocation data (*freq* ≥ 3) and (b) the concordance sets. As already noted, the identification of phrases goes far beyond simple syntactic frame identification and, it may take us to important syntactic, lexical and semantic discoveries about the behaviour of words, collocates and phrases: *the sociology of words.*

In the following final section, we shall try to introduce the reader to a more challenging side of statistics: language modelling[17].

5. Statistical language modelling

The term *model* as used here is understood to represent a simplified description of a natural language property.

5.1. Modelling the lexicon

Sánchez and Cantos (1997; 1998) use the term *model* to represent mathematically the transitive relationship[18] between types, tokens and lemmas. This relationship has led to interesting findings. So, while tokens grow linearly, types and lemmas do so in a curvilinear fashion[19] (*Figure 1*). After a thorough analysis, Sánchez and Cantos (1997) came up with three statements of prediction regarding tokens, types and lemmas. That is, the amount of types and lemmas a text or corpus is made of is related to the total amount of tokens. The first of these statements is the so-called *Type-Token formula*:

$$Types = K\sqrt{Tokens}$$

[17] Readers unfamiliar with statistics might find some parts of this next section difficult to grasp.

[18] If a relation R, whenever it holds between both x and y and between y and z, also holds between x and z, the relation is said to be transitive: $\forall x\ \forall y\ \forall z\ ((R\ (x,y) \rightarrow (R\ (y,z))$ ® $R\ (x,y))$ (Allwood et al. 1977: 89).

[19] For an ample discussion on the predictive power of the *Type-Token formula,* see Yang, Cantos and Song (2000), and Yang, Song, Cantos and Lim (*forthcoming*).

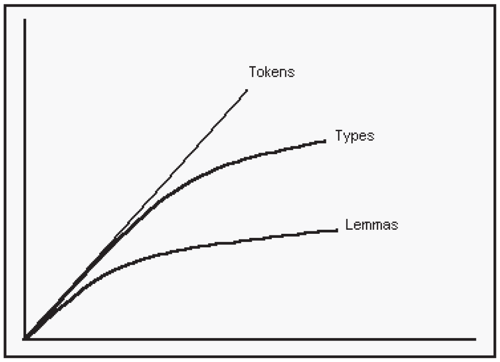


Figure 1: Graphical representation of the increase of tokens, types and lemmas

This formula states that the total amount of types can be modelled and predicted given the total tokens of a text. The *K* is a text dependent constant value and needs to be calculated beforehand using a small sample of the text or corpus under research[20]. In other words, the formula above models the relationship between types and tokens. This mathematical model enables researchers to estimate reliably the hypothetical number of types a corpus may entail, even before compiling the corpus. In much the same way, we can also calculate the lemmas:

$$Lemmas = K_L \sqrt{Tokens}$$

And also infer the lemmas from a type sample:

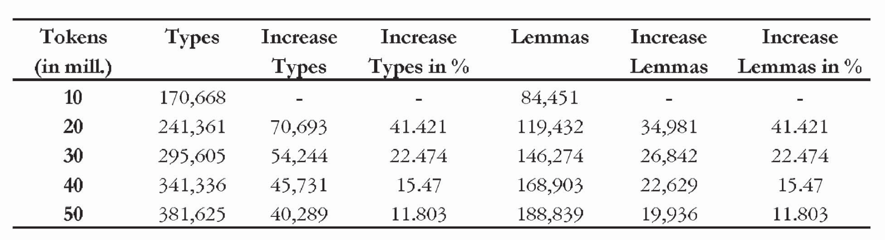
$$Lemmas = Types * \left(\frac{K}{K_L} \right)$$

The table below shows the predictions of types and lemmas for a hypothetical corpus of contemporary Spanish[21] (*Table 13*).

[20] The *K-value* is a *type-token* constant, whereas the K_L*-value* is a *lemma-token* constant.

[21] The calculations are based on a single small 250,000 token sample taken randomly from the *CUMBRE Corpus*, a corpus of contemporary Spanish (for more details see Sánchez et al. 1995).

The analytic technique for predicting types proposed and applied by Sánchez and Cantos (1997) is simple and straightforward and the resulting formulae are easy to use, flexible and can be applied quickly to any corpora or language samples. After thorough testing on various text samples of different sizes, the formulae have shown to be very reliable with a more than acceptable error margin of ±5%, and this speaks eloquently of their validity. Their most positive contributions can be summarised in the following points: (a) they are stable indicators of lexical diversity and lexical density[22]; (b) they overcome the reliability flaw of both the *token-type ratio* and *type-token ratio*[23] as they are not constrained or dependent on text length; and (c) they can be used as predictive tools to account for the total amount of types and lemmas that any hypothetical corpus might contain (see Sánchez and Cantos 1998).



Tokens (in mill.)	Types	Increase Types	Increase Types in %	Lemmas	Increase Lemmas	Increase Lemmas in %
10	170,668	-	-	84,451	-	-
20	241,361	70,693	41.421	119,432	34,981	41.421
30	295,605	54,244	22.474	146,274	26,842	22.474
40	341,336	45,731	15.47	168,903	22,629	15.47
50	381,625	40,289	11.803	188,839	19,936	11.803

Table 13: Projections of the *CUMBRE Corpus*

A further revealing issue is that the application of the formulae above on different text samples outputs idiosyncratic, unique and distinctive slopes. The contrastive graph below (*Figure 2*) clearly shows that, for example, Conrad's lexical density is superior to Doyle's and Shakespeare's. This evidence suggests that these formulae might also be valid for text, author and language classifications, among others.

[22] Succinctly, what is ment by *lexical diversity* or *lexical density* is *vocabulary richness*.

[23] Both the *token-type ratio* and the *type-token ratio* provide information on the distribution of tokens between the types in a text. The *token-type ratio* is the mean frequency of each token in a text, whereas the *type-token ratio* reveals the mean distribution of types in a text or corpus (if we eventually multiply this quotient by 100, we get the mean percentage of different types per one hundred words).

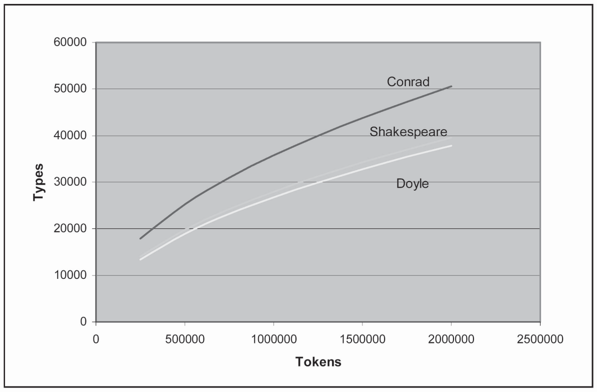


Figure 2: Comparing type distribution in Conrad, Shakespeare and Doyle

5.2. Modelling text-classification

For the following experiment, we (a) extracted (from the *CUMBRE Corpus*) 11 different text samples from textbooks and manuals for secondary education and university level education, relative to various subjects or linguistic domains, (b) obtained their total amounts of tokens and types, and (c) calculated their *K-values* (see above). The results are shown below in *Table 14*.

Sample	Tokens	Types	K-value
Architecture	64431	11225	44.22
Chemistry	22539	2771	18.46
Computing	18822	2344	17.09
Geography	48544	7341	33.32
History	29711	5671	32.90
Mathematics	18700	1907	13.95
Medicine	39639	5228	26.26
Natural Sciences	41650	5982	29.31
Philosophy	20385	3344	23.42
Physics	15233	2378	19.27
Sociology	75149	11522	42.03

Table 14: Tokens, types and K-values relative to 11 linguistic domains (*CUMBRE Corpus*)

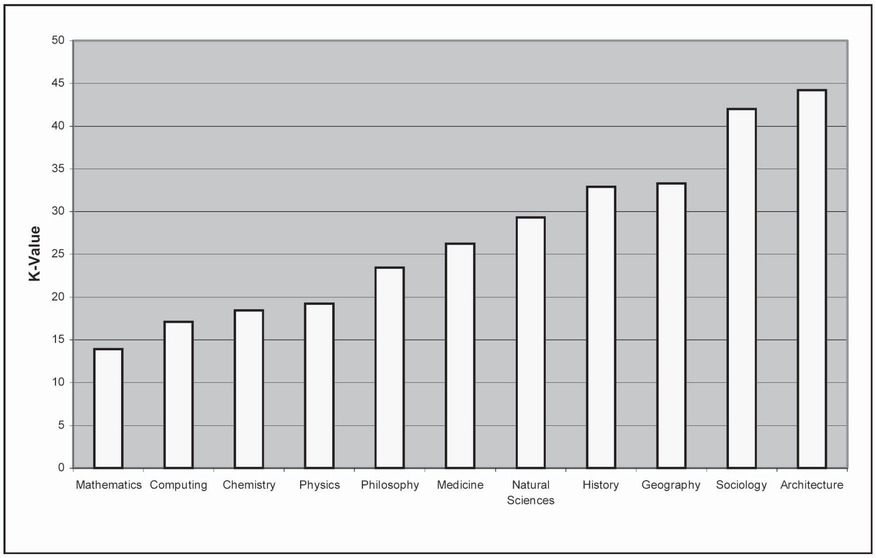


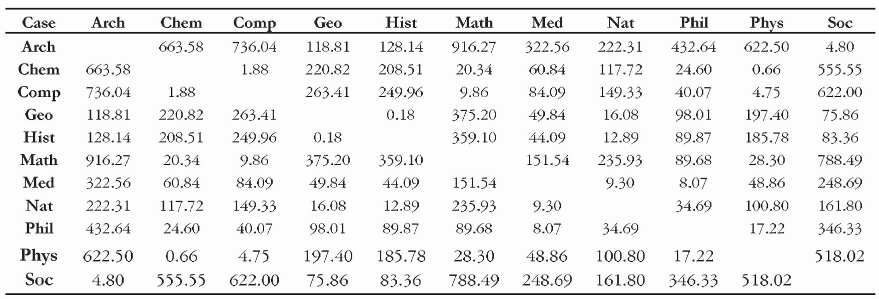
Figure 3: Text types and their lexical density (K-values)

The mean *K-value* for the 11 samples is 27.29 and its standard deviation 9.43. Comparing these figures with the individual *K-values* from the table above (*Table 14*) reveals a great deal of variability or dispersion among the various text samples. The sample on *physics* compared with *sociology* indicates huge differences in lexical density, not to mention *mathematics* versus *architecture*. However, *geography* and *history* seem to have a very similar lexical density. The histogram above (*Figure 3*) graphically displays the various text types ordered according to their lexical densities (*K-values*). Interesting here is the fact that the lexical density ordered scale moves smoothly from pure science subjects (*mathematics, computing, chemistry*, etc.) to more arts and humanistic content texts. Additionally, neighbourhood on the histogram might suggest subject relatedness: the more dissimilar the lexical density indices (*K-values*) the less the subjects relate to each other.

The *K-values* suggest that discrimination between *chemistry* (18.46) and *sociology* (42.03) texts might indeed be possible as both figures diverge significantly. However, a *K-value* based distinction between *chemistry* (18.46) and *physics* (19.27) seems less reliable, due to its closeness.

To construct a purely statistical discrimination model, we started experimenting with a statistical technique known as *cluster analysis*.

Succinctly, *cluster analysis* classifies a set of observations into two or more mutually exclusive groups based on the combination of interval variables[24]. The purpose of *cluster analysis* is to discover a system of organizing observations into groups, where members of the group share properties. *Cluster analysis* classifies unknown groups while *discriminant function analysis* (see below) classifies known groups. A common approach to performing a *cluster analysis* is to first create a table or matrix of relative similarities or differences between all objects and second to use this information to combine objects into groups. The table of relative similarities is called a proximity or *dissimilarity matrix*. *Table 15* displays this *dissimilarity matrix*[25].



Case	Arch	Chem	Comp	Geo	Hist	Math	Med	Nat	Phil	Phys	Soc
Arch		663.58	736.04	118.81	128.14	916.27	322.56	222.31	432.64	622.50	4.80
Chem	663.58		1.88	220.82	208.51	20.34	60.84	117.72	24.60	0.66	555.55
Comp	736.04	1.88		263.41	249.96	9.86	84.09	149.33	40.07	4.75	622.00
Geo	118.81	220.82	263.41		0.18	375.20	49.84	16.08	98.01	197.40	75.86
Hist	128.14	208.51	249.96	0.18		359.10	44.09	12.89	89.87	185.78	83.36
Math	916.27	20.34	9.86	375.20	359.10		151.54	235.93	89.68	28.30	788.49
Med	322.56	60.84	84.09	49.84	44.09	151.54		9.30	8.07	48.86	248.69
Nat	222.31	117.72	149.33	16.08	12.89	235.93	9.30		34.69	100.80	161.80
Phil	432.64	24.60	40.07	98.01	89.87	89.68	8.07	34.69		17.22	346.33
Phys	622.50	0.66	4.75	197.40	185.78	28.30	48.86	100.80	17.22		518.02
Soc	4.80	555.55	622.00	75.86	83.36	788.49	248.69	161.80	346.33	518.02	

Table 15: Matrix of dissimilarity of the text sample subjects

Looking at the matrix we find that the least dissimilarity or closest similarity of all is 0.18, between the *history* text sample and the *geography* one. We could say that these seem to form the pair that is most alike. *Physics* and *chemistry* have a very low dissimilarity index (0.66) and could be grouped, too. Since *history* is related to *geography* we could say that these form a cluster. On the opposite scale, we find the hugest difference between *mathematics* and *architecture* (916.27). After the distances between the text types have been found, the next step in the *cluster analysis* procedure is to

[24] The property of intervals is concerned with the relationship of differences between objects. If a measurement system possesses the property of intervals it means that the unit of measurement is the same thing throughout the scale of numbers. That is, a centimetre is a centimetre, no matter were it measured.

[25] The *dissimilarity matrix*, the *dendogram* and the *discriminant function analysis* have been calculated and produced using *SPSS v. 10*, a statistics package for the social sciences.

divide the text types into groups based on the distances. The results of the application of the clustering technique are best described using a *dendogram* or binary tree. The interpretation of the *dendogram* is fairly straightforward (*Figure 4*). For example, *Geo/His/Nat* form a group, *Chem/Phys/Comp/Math* form a second group and *Arch/Soc* make up a “runt” as it does not enter any group until near the end of the procedure. Our *dendogram* outputs 6 possible solutions:

Solution 1: 1 group: (1) *Geo/Hist/Nat/Med/Phil/Chem/Phys/Comp/Math/Arch/Soc*.

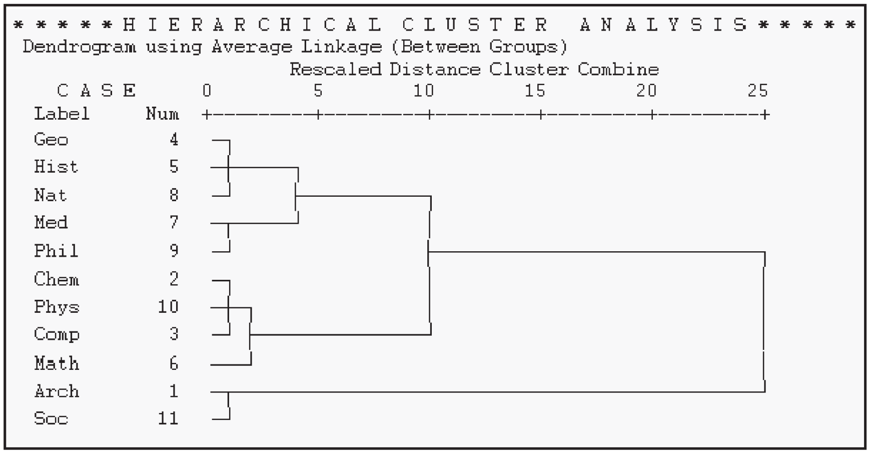
Solution 2: 2 groups: (1) *Geo/Hist/Nat/Med/Phil/Chem/Phys/Comp/Math* and (2) *Arch/Soc*

Solution 3: 3 groups: (1) *Geo/Hist/Nat/Med/Phil*, (2) *Chem/Phys/Comp/Math,* and (3) *Arch/Soc*

Solution 4: 4 groups: (1) *Geo/Hist/Nat*, (2) *Med/Phil*, (3) *Chem/Phys/Comp/Math,* and (4) *Arch/Soc*

Solution 5: 5 groups: (1) *Geo/Hist/Nat*, (2) *Med/Phil*, (3) *Chem/Phys/Comp*, (4) *Math* and (5) *Arch/Soc*

Solution 6: 11 groups: (1) *Geo*, (2) *Hist*, (3) *Nat*, (4) *Med*, (5) *Phil*, (6) *Chem*, (7) *Phys*, (8) *Comp*, (9) *Math*, (10) *Arch* and (11) *Soc*



```
* * * * * H I E R A R C H I C A L   C L U S T E R   A N A L Y S I S * * * * *
 Dendrogram using Average Linkage (Between Groups)
                         Rescaled Distance Cluster Combine
   C A S E      0         5        10        15        20        25
  Label     Num  +---------+---------+---------+---------+---------+
  Geo         4
  Hist        5
  Nat         8
  Med         7
  Phil        9
  Chem        2
  Phys       10
  Comp        3
  Math        6
  Arch        1
  Soc        11
```

Figure 4: Hierarchical *cluster analysis*

Obviously, the best solution is (6), as it models the classification of all 11 text types, whereas solution (1) is clearly the worst one, as it is unable to classify any text at all.

Cluster analysis methods always produce a grouping. The grouping produced by this analysis may or may not prove useful for classifying objects. To validate the results of a *cluster analysis* it has to be used in conjunction with *discriminant function analysis* on the resulting groups (solutions). *Cluster analysis* is a positive exploratory tool in order to elucidate possible grouping solutions and to construct at a later stage a group membership predictive model by means of the *discriminant function analysis*. This later technique is based on the interval variables (*K-values*). It begins with a set of observations where both group membership and the values of the interval variables are known. The end result of the procedure is a model that allows prediction of membership when only the interval variables are known: the *K-values*. A second purpose of *discriminant function analysis* is an understanding of the data set, as a careful examination of the prediction model that results from the procedure can give insight into the relationship between group membership and the variables used to predict group membership.

In order to construct a model using *discriminant function analysis*, we added 11 more text samples, one for each text type. Next, using the *cluster analysis* data, we constructed six models, one for each solution or grouping (see above).

The table below (*Table 16*) displays the discriminant model for solution 6 (11 text types). It shows the case number (each text sample), actual group[26], group assignment[27] (*Highest Group* and *2nd Highest Group*) and discriminant scores. Note that erroneous group assignment is marked with "**". The success rate (correct group assignment) is 81.81%; it failed on correctly assigning cases 10, 13, 15 and 16, which were, however, correctly classified in the second choice: *2nd Highest Group*.

[26] 1 = *Architecture*; 2 = *Chemistry*; 3 = *Computing*; 4 = *Geography*; 5 = *History*; 6 = *Mathematics*; 7 = *Medicine*; 8 = *Natural Sciences*; 9 = *Philosophy*; 10 = *Physics*; and 11 = *Sociology*.

[27] It refers to the automatic text-classification performance of the model, based on a single lexical density measure: *K-value*.

		Group Assignment		
Case Number	Actual Group	Highest Group	2nd Highest Group	Discrim. Score
1	1	1	11	23219
2	2	2	3	-12.424
3	3	3	2	-14.320
4	4	4	5	8.137
5	5	5	4	7.556
6	6	6	3	-18.664
7	7	7	9	-1.631
8	8	8	7	2.589
9	9	9	7	-5.561
10	10	2**	10	-11.303
11	11	11	1	20.189
12	1	1	11	21697
13	2	10**	2	-11.234
14	3	3	2	-13005
15	4	5**	4	7.418
16	5	4**	5	7971
17	6	6	3	-17.045
18	7	7	8	-.165
19	8	8	7	.984
20	9	9	7	-3.209
21	10	10	2	-10127
22	11	11	1	18930

Table 16: *Discriminant function analysis* for solution 6 (11 texts)

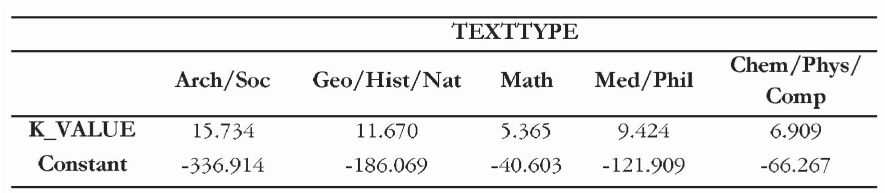
The next discriminant model based on solution 5 (*Table 17*) output a very promising 95.5% success rate (it only failed in classifying case 19).

These analyses are very revealing and it is now up to the reader to choose or decide which solution is best. We do think that the best model is solution 5, because of its reasonable discrimination power (it is able to discriminate 5 different text types: (1) *Geo/Hist/Nat,* (2) *Med/Phil,* (3) *Chem/Phys/Comp,* (4) *Math* and (5) *Arch/Soc*) and its accuracy (95.5%).

Solutions	Clusters	Success Rate
1	1	100%
2	2	100%
3	3	100%
4	4	95.5%
5	5	95.5%
6	11	81.81%

Table 17: Summary of the various solutions, cluster divisions and associated success rates

Another positive contribution of *discriminant function analysis* is that once the groups are known, we can construct a model that allows prediction of membership. This is done by means of the resulting *discriminant function coefficients*. The coefficients for solution 5 are (*Table 18*):



	TEXTTYPE				
	Arch/Soc	**Geo/Hist/Nat**	**Math**	**Med/Phil**	**Chem/Phys/ Comp**
K_VALUE	15.734	11.670	5.365	9.424	6.909
Constant	-336.914	-186.069	-40.603	-121.909	-66.267

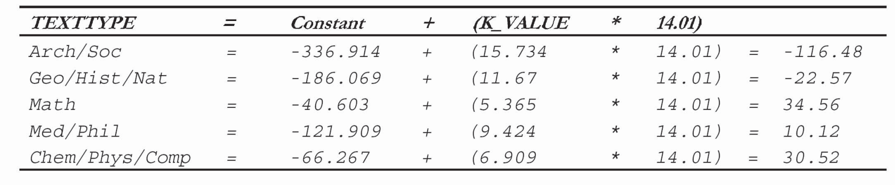
Table 18: Coefficients of discriminant function analysis *(*solution 5*)*

Thus, the *discriminant equation* would be:

*TEXTTYPE = Constant + (K_VALUE * x*)

Where *x* stands for any given *K-value*. Important to note here: do not mistake *K_VALUE* with given *K-value*. The former refers to a text specific coefficient that has been calculated by the *discriminant function analysis*, whereas the latter is the lexical density index discussed earlier in the text (see *Section V.1*).

To illustrate the discriminant and predictive power of the equation, take, for example, a hypothetical text with a *K-value* = 14.01, that is *x=14.01*. We instantiate *TEXTTYPE* and the coefficients *Constant* and *K_VALUE* accordingly and get the following results:



TEXTTYPE	***=***	***Constant***	***+***	***(K_VALUE***	*****	***14.01)***		
Arch/Soc	=	*-336.914*	+	*(15.734*	*	*14.01)*	=	*-116.48*
Geo/Hist/Nat	=	*-186.069*	+	*(11.67*	*	*14.01)*	=	*-22.57*
Math	=	*-40.603*	+	*(5.365*	*	*14.01)*	=	*34.56*
Med/Phil	=	*-121.909*	+	*(9.424*	*	*14.01)*	=	*10.12*
Chem/Phys/Comp	=	*-66.267*	+	*(6.909*	*	*14.01)*	=	*30.52*

Table 19: Results of the discriminant equations

Next, we just need to maximize the five results, that is, choose the maximum result. So, a text with a *K-value* = 14.01 would be most likely classified in first choice as being a *mathematics* text, as *Math* is the highest resulting coefficient (34.56); and in second choice, it would be classified as *Chem/Phys/Comp* (30.52). Similarly, the least likely group membership would be *Arch/Soc* (-116.48).

Interesting in this sense are *Figures 3* and *4*, and the *discriminant function analysis*. *Figure 3* represents visually the *K-value* ordered linguistic domains, where we can appreciate a logical and smooth text type transition, that goes from pure science (*mathematics*) to clear humanity contents (*sociology/ architecture*). This stratification is based on a sole lexical density feature: the *K-value*. In addition, *Figure 4* offers an exploratory grouped hierarchical structure of the text types, highlighting the major flaw of the *K-value*: its incapacity to distinguish between closely nearby *K-values*, as these are grouped into single clusters. Clearly, the *K-value* fails to distinguish between (a) *geography, history* and *natural sciences*; (b) *medicine* and *philosophy*; (c) *chemistry, physics* and *computing*; and (d) *sociology* and *architecture*. However, the final modelling of the data by means of the *discriminant function analysis* reveals that the *K-value* is valid and reliable for successful differentiation of (a) *geography/history/natural sciences*, (b) *medicine/philosophy*, (c) *chemistry/physics/ computing*, (d) *sociology/architecture* and (e) *mathematics* from each other. Though a potential text discriminator using *K-value* does not, in principle, output a very fine-grained classification, it does not invalidate the use of lexical density for text differentiation. The resulting text classification from the experiment is far from being erroneous or exaggeratedly generic. On the contrary, it discriminates clearly distinctive text type clusters with a promising accuracy rate.

6. Some final remarks

In this paper, we have tried to scrutinise studies from a number of diverse linguistic areas (lexicography, grammar, stylistics and also computational applications: probabilistic language modelling) and attempted to show the usefulness for statistics in each. In addition, we have also highlighted the issue of texts and corpora as sources of quantitative data. The important role of quantitative analysis and its interaction with qualitative analysis has been described and exemplified. We have tried to call the reader's attention to the singularity of statistical regularities and patterns in natural languages, that is, that while we are engaged in communication, we do not consciously monitor our language to ensure that these statistical properties entail. It would be impossible to do so. Yet, without any deliberate effort on our part, we shall find the same underlying regularities in any large sample of our speech or writing.

Statistics allows us to summarise complex numerical linguistic data in order to draw inferences from them: intuitive comparisons cannot always be used to tell us something significant about linguistic variation and so we considered the role of significance tests, such as *chi-square* and *z-score* in telling us how far differences between samples may be due to chance. The need to summarise and infer stems from the fact that there is variation in linguistic data; else there would be no place for statistics in linguistics.

Finally, we also envisage two underlying purposes in this article. First, by applying and discussing the statistical techniques used, the reader can evaluate the techniques employed: a critical evaluation on the appropriateness of the statistical methods used and the assumption they make for linguistic analysis, though no attempt has been made to present a thorough survey of statistical methods available to statistics, trying to make them more accessible for non-specialists (linguists and postgraduate students). Second, as some readers might be interested in planning their linguistic research using statistics, several of the techniques introduced in here might, partly, assist their aim in similar areas and topics.

The increasing accessibility of linguistic corpora and the belief that theory must be based on language *as it is* have placed empirical linguistics once again at the foreground of linguistics. The immediate implication of these assumptions is that linguists will increasingly demand the use of statistics in their research. The answer, hence to the title of this paper is definitively *yes*.

Acknowledgements

Thanks go to Dr. Javier Valenzuela and Richard Barbosa for their comments, advice and useful suggestions on an earlier draft of this paper.

Recebido em dezembro de 2001.

References

Allwood, J., L.-G. Andersson & Ö. Dahl 1977. *Logic in Linguistics*. Cambridge: CUP.

Barnbrook, G. 1996. *Language and Computers*. Edinburgh: Edinburgh University Press.

Baayen, R. H. 2001. *Word Frequency Distribution.* Dordrecht: Kluwer Academic Publishers.

Brown, D.J. 1988. *Understanding Research in Second Language Learning. A Teacher's Guide to Statistics and Research Design.* Cambridge: CUP.

Cantos, P. 1995. Tratamiento informático y obtención de resultados. *CUMBRE corpus lingüístico del español contemporáneo. Fundamentos, metodología y aplicaciones.* Ed. A. Sánchez, R. Sarmiento, P. Cantos and J. Simón. Madrid: SGEL: 39-70.

______ 2000. Investigating Type-token Regression and its Potential for Automated Text Discrimination. *Cuadernos de Filología Inglesa. Número monográfico: Corpus-based Research in English Language and Linguistics.* Ed. P. Cantos and A. Sánchez. Murcia: Servicio de Publicaciones de la Universidad de Murcia: 71-91.

______ 2001. An Attempt to Improve Current Collocation Analysis. *Technical Papers. Volume 13. Special Issue. Proceedings of the Corpus Linguistics 2001 Conference.* Ed. P. Rayson, A. Wilson, T. McEnery, A. Hardie and S. Khoja: 100-8.

Cantos P. & A. Sánchez (*forthcoming*) Lexical Constellations: What Collocates Fail to Tell. *International Journal of Corpus Linguistics.*

Chipere, N., D. Malvern, B. Richards & P. Duran. 2001. Using a Corpus of School Children's Writing to Investigate the Development of Vocabulary Diversity. *Technical Papers. Volume 13. Special Issue. Proceedings of the Corpus Linguistics 2001 Conference.* Ed. P. Rayson, A. Wilson, T. McEnery, A. Hardie and S. Khoja: 126-133.

Church, K.W., Gale, W., Hanks, P. & Hindle, D. 1991. Using Statistics in Lexical Analysis. Ed. U. Zernik. *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*. Hillsdale, NJ: Lawrence Erlbaum Associates: 115-164.

Clear, J. 1993. From Firth Principles: Computational Tools for the Study of Collocation. Ed. M. Baker et al. *Text and Technology*. Amsterdam: Benjamins: 271-292.

Haztivassiloglou, V. 1994. Do We Need Linguistics When We Have Statistics? A Comparative Analysis of the Contributions of Linguistic

Cues to Statistical Word Grouping System. *The Balancing Act. Combining Symbolic and Statistical Approaches to Language.* Ed. J.L. Klavans and P. Resnik. Cambridge: The MIT Press. 67-94.

Hunston, S. & G. Francis 1999. *Pattern Grammar. A Corpus-driven Approach to the Lexical Grammar of English.* Amsterdam: John Benjamins.

Mcenery, T. & A. Wilson 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Mckee G, D. Malvern & B.J. Richards 2000. Measuring Vocabulary Diversity Using Dedicated Software. *Literary and Linguistic Computing* 15 (3): 323-38.

Oakes, M. 1998. *Statistics for Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Sánchez, A. & P. Cantos 1997. Predictability of Word Forms (Types) and Lemmas in Linguistic Corpora. A Case Study Based on the Analysis of the CUMBRE Corpus: An 8-Million-Word Corpus of Contemporary Spanish. *International Journal of Corpus Linguistics* 2(2): 259-280.

_______ 1998. El ritmo incremental de palabras nuevas en los repertorios de textos. Estudio experimental y comparativo basado en dos corpus lingüísticos equivalentes de cuatro millones de palabras, de las lenguas inglesa y española y en cinco autores de ambas lenguas. *ATLANTIS* XIX(2): 205-223.

Sánchez, A., R. Sarmiento, P. Cantos & J. Simón 1995. *CUMBRE corpus lingüístico del español contemporáneo. Fundamentos, metodología y aplicaciones.* Madrid: SGEL.

Scott, M. 1996. *Wordsmith Tools*. Oxford: Oxford University Press.

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

Sinclair, J. ET AL. 1995. *Collins Cobuild English Dictionary*. London: Harper Collins Publishers.

Stubbs, M. 1995. Collocations and Semantic Profiles: On the Cause of the Trouble with Quantitative Methods. *Functions of Language* 2(1): 1-33.

Woods, A., P. Fletcher & A. Hughes 1986. *Statistics in Language Studies*. Cambridge: CUP.

Yang, D.H., P. Cantos & M. Song 2000. An Algorithm for Predicting the Relationship between Lemmas and Corpus Size. *International Journal of the Electronics and Telecommunications Research Institute (ETRI)* 22: 20-31.

Yang, D.H., M. Song, P. Cantos & S.J. Lim (*forthcoming*) On the Corpus Size Needed for Compiling a Comprehensive Computational Lexicon by Automatic Lexical Acquisition. *Computers in the Humanities.*

Zipf, G. 1935. *The Psycho-Biology of Language*. Boston: Houghton Mifflin.

APPENDIX 1: Co-occurring items with KEEP

the	42
to	38
and	21
of	18
he	15
his	14
in	14
a	12
that	10
up	10
him	9
with	9
was	8
I	7
it	7
but	6
for	6
on	6
could	5
from	5
going	5
if	5
is	5
themselves	5
by	4
hands	4
not	4
off	4
out	4
their	4
they	4
were	4
would	4
all	3
always	3
an	3
are	3
because	3
down	3
feet	3
Her	3
house	3
me	3
my	3
one	3
our	3
pony	3
she	3
some	3
we	3
well	3
which	3
You	3
your	3
awake	2
away	2
be	2
better	2
company	2
designing	2
fit	2
full	2
garden	2
had	2
himself	2
journal	2
letters	2
list	2
low	2
more	2
order	2
pencil	2
sheet	2
so	2
than	2
them	2
there	2
this	2
trotting	2
trying	2
uncle	2
us	2
where	2
who	2
wrote	2
able	1
about	1
above	1
acquaintances	1
admirer	1
after	1
alive	1
altar	1
although	1
angle	1
any	1
apart	1
author’s	1
avoiding	1
babysat	1
backdrop	1
bag	1
beautiful	1
been	1
behind	1
bent	1
Bones	1
Bones’s	1
both	1
boy	1
boys	1
Brawne	1
busy	1
can	1
can’t	1
cannot	1
careful	1
catalogue	1
cattle	1
change	1
charcoal	1
chattering	1
clean	1
clear	1
closed	1
cloths	1
collar	1
coming	1
Commentary	1
conversation	1
cooking	1
copies	1
correct	1
cost	1
Cret	1
dance	1
dark	1
daughter	1
day	1
dear	1
delicious	1
delivery	1
developments	1
did	1
didn’t	1
direction	1
distance	1
do	1
double	1
draw	1
drawing	1
dreams	1
during	1
ear	1
Eastern	1
eating	1
edge	1
editorials	1
elaborate	1
embarrass	1
enormous	1
Europe	1
evening	1
ever	1
every	1
eyes	1
faith	1
Firelight	1
form	1
frail	1
friends	1
gambling	1
gate	1
genitals	1
gesticulating	1
get	1
go	1
good	1
Government	1
great	1
habit	1
having	1
he’ll	1
hedges	1
helped	1
Henriette	1
here	1
hope	1
horses	1
how	1
image	1
insisted	1
instinct	1
instructors	1
instrument	1
instruments	1
interest	1
its	1
jargon	1
Jews	1
just	1
keep’em	1
labour	1
Lady	1
Lautrec	1
lavender	1
literary	1
long	1
lot	1
Lucky	1
made	1
making	1
man’s	1
masking	1
mc	1
meticulously	1
middle	1
Midnight	1
mind	1
minimum	1
ministering	1
mistakes	1
most	1
mother	1
mouth	1
Mr	1
multiplying	1
murdering	1
Mysterious	1
Nails	1
Nails’	1
name	1
nearly	1
negation	1
never	1
no	1
notebooks	1
nothing	1
novel	1
oats	1
obliged	1
occasional	1
occasions	1
office	1
often	1
old	1
once	1
one’s	1
opinions	1
or	1
other	1
own	1
pace	1
paper	1
parents	1
parish	1
part	1
past	1
people	1
Plas	1
pocket	1
poet	1
poetry	1
posted	1
preliminary	1
premises	1
prodding	1
profile	1
prying	1
pulling	1
put	1
questions	1
Ray	1
regret	1
regular	1
reins	1
required	1
restriction	1
revert	1
riding	1
road	1
room	1
round	1
salaries	1
Sam	1
Scores	1
Sear’s	1
seems	1
seldom	1
several	1
should	1
show	1
shut	1
sighing	1
simply	1
sketches	1
laved	1
sleeping	1
social	1
someone	1
somewhat	1
son	1
special	1
spirits	1
staff	1
strips	1
stuck	1
tank	1
teeth	1
these	1
thing	1
things	1
thinking	1
tie	1
together	1
told	1
touch	1
treasure	1
two	1
typical	1
under	1
unpublished	1
unvarying	1
used	1
veritable	1
want	1
wanted	1
watch	1
whenever	1
while	1
whole	1
wife	1
Wilfred	1
will	1
Wilmot	1
wont	1
workers	1
yew	1
You’d	1

APPENDIX 2: Statistically significant collocations of KEEP (*z-scores*)

Types	*z-scores*	Types	*z-scores*
pony	43.938489535	are	9.1620841143
hands	40.519300438	but	9.103835908
themselves	35.334195956	we	8.8348414766
feet	34.611056082	were	8.7933182281
going	28.713144661	with	8.7078778874
your	22.241106218	she	8.7819039837
off	21.11049718	all	8.3567243173
up	20.281175809	her	8.2737624536
always	19.035088921	on	8.1044624203
house	17.547255459	you	8.1569757642
because	16.892582283	one	8.1072638116
down	15.912511247	to	7.5131904835
could	15.532705362	by	7.9933712839
him	15.046437833	that	7.7026241986
well	14.617538824	not	7.6211188056
our	14.125181914	an	7.4307469612
if	14.086680646	which	7.2159862265
some	12.619157651	for	7.0078354785
out	11.745960466	It	6.814018187
would	11.553197983	was	6.4643741642
me	11.553296279	is	6.1068585324
their	9.8398604942	in	5.3022303972
my	9.6589305918	a	4.6564099588
from	9.4379164748	and	4.3028186585
he	9.2571259703	I	4.5055296849
his	9.2347221047	of	3.6182220552
they	9.2025100949	the	2.0621643093

APPENDIX 3: Concordance lines for the lemma KEEP

```
 ...eraser to use for what purpose, how to [[keep]] a sheet clean by masking or tacking dow...
...cle Knacker's to see the new horses. To [[keep]] a pony in the middle of a town, and ke...
...answering their questions and trying to [[keep]] a conversation going while cooking a c...
 ...views about the company her son should [[keep]] and used to embarrass him by vetting hi...
              ...o tears. `Not Bones!" `To [[keep]] Bones, would you rather go on with the ...
       ...ive. `Might be a bit of trouble. [[Keep]] clear. Bit nervous, this one. Bin in ba...
... a stringy-looking tie that would never [[keep]] closed up to the collar itself. His ve...
... been doing throughout the action. They [[keep]] coming, like immigrants, or refugees, l...
...r other writers he published. I did not [[keep]] copies of my letters to him _ or to an...
... they had to stay put if they wanted to [[keep]] designing. There were almost no ...
 ... neck, pushing him down. `Bloody hell! [[Keep]] down! It's the fuzz!" Keeping be...
 ...my girl. You owe it to your parents to [[keep]] him off the premises." `Is he th...
       ...ed an old brown photograph. `You [[keep]] him here, and say your prayers, and all...
 ...nd she babysat nearly every evening to [[keep]] him in good oats (not Uncle Knacker's) ...
 ...ep a pony in the middle of a town, and [[keep]] him fit, was hard work, as he had to be...
 ...e willed him to, because she slaved to [[keep]] him up to it. When she looked at poor l...
... to deteriorate, although nothing could [[keep]] him from his garden and the prize bird...
...t the subject of his letters, he cannot [[keep]] his mind off poetry for more than a fe...
... guessed that some instinct told Sam to [[keep]] his mouth shut about the knacker-yard ...
...here the author's genitals were wont to [[keep]] house. Gregory Woods...
...he home, the more labour is required to [[keep]] it fit== ; the more labour required, t...
...treasure for this purpose, if one could [[keep]] it. Otherwise a spoon sufficed. The pe...
       ... ambition in life?" `You joking? [[Keep]] lot of the old man's hair mostly." ...
... the kind of parson I go to. Some would [[keep]] me trotting round the parish all day: ...
...nto the bed; his chattering teeth would [[keep]] me awake after he himself had fallen in...
 ...e rather narrow gate and was trying to [[keep]] Midnight from eating a lavender bush by...
...ncouraged by our drawing instructors to [[keep]] notebooks and to draw whenever we coul...
...rd pencil, but one had to be careful to [[keep]] one's hands off the sheet to avoid spo...
       ... much better." `You'd be able to [[keep]] order, sir. I mean, it's not enough to ...
...cks his head, to hold it up. But if you [[keep]] pulling on the reins he'll get cross."...
...ur nags then?" Sebastian asked. `Better [[keep]] that one away from the others" _ he no...
...er, were not such that he could hope to [[keep]] the beautiful daughter of the former m...
        ... Mysterious ministering spirits [[keep]] the room in order, for surprising thing...
     ...ister ) but, poor fellow, he can't [[keep]] the name of Brawne out of it! ...
 ... `I had an idea," she said slowly. `To [[keep]] the whole thing going. Save the horses ...
... had run off with an Irish bricklayer. `[[Keep]] the social workers in full employment,...
...are and how to get the correct angle to [[keep]] the edge of the instrument clean by pr...
 ...ing or tacking down strips of paper to [[keep]] the instruments riding somewhat above t...
...anadian context. I regret that I didn't [[keep]] the letters Ray wrote us during the wa...
 ... submission that for so long helped to [[keep]] the Jews of Eastern Europe in medieval ...
...he checked with her uncle that he would [[keep]] them in horse-feed, and he agreed, amu...
... raised her eyebrows, impressed. `Well, [[keep]] them all on your list. It might be use...
 ... the street, it was insisted that they [[keep]] themselves to themselves. The situation...
...ils working together and not obliged to [[keep]] to their preliminary sketches, arrive a...
 ... an hour each and two shorter ones. To [[keep]] up with this pace, that is, just to car...
...natural swimmer and loved it, and could [[keep]] up with Nails for several lengths, wit...
 ... beating us, she's dangerous. You just [[keep]] us posted, let us know how they're gett...
...reak into houses. The most he can do is [[keep]] watch. He manages to acquire a little m...
 ...ce in every two circles, but could not [[keep]] with it for more than two or three stri...
       ...to stay squarely in the saddle. `[[Keep]] your hands down. Don't pull on his mout...
       ...t better. `Where are we going to [[keep'em]]?" `Here, Dad said." `Here...
 ...to walk backwards away from my father, [[keeping]] an unvarying distance. It seemed as tho...
       ...hell! Keep down! It's the fuzz!" [[Keeping]] bent double, Hoomey wriggled in a frenz...
 ... to avoiding Bones's enormous feet and [[keeping]] his hands away from the ever-questing l...
 ...ite acceptable characteristics. He was [[keeping]] his feet well out of the way, following...
...sociations that any change seems out of [[keeping]]. Such conservatism _ at times stodgines...
       ...e horses _" `My uncle will go on [[keeping]] the horses if we want them. I can teach...
...all directions, but Nutty had no mercy, [[keeping]] the pony going. After a bit she ...
...y had known in Birkenhead, the habit of [[keeping]] themselves to themselves was not easil...
 ...mation on buying, building, and simply [[keeping]] up a house. Alongside the standard sani...
 ... his head clear of the water. He swam, [[keeping]] up with Nails' feet, gulping, coughing....
 ...t of Edward's gambling and the cost of [[keeping]] up Plas Wilmot) left Alderman Shaw with...
...o be the same. He ran all the way home, [[keeping]] well off the road in case Nails came i...
              ... `You're joking!" `No. He [[keeps]] a sleeping bag up there, stuffed behind...
... and neither provides it; but where one [[keeps]] faith with it by negation and suffering...
...bility to which an admirer of the novel [[keeps]] having to revert. Times have ...
```

```
...fifteen miles. It won't show. He always [[keeps]] his tank full. Sometimes I take the br...
 ...e to be someone other than himself who [[keeps]] murdering people, he does a tremendous ...
 ...his civil servant toils and spins; She [[keeps]] the Government. (`Daumiers", N...
 ...preferences of her clients. Her office [[kept]] a veritable Sear's catalogue of histori...
... , Maybeck's, it is in part because she [[kept]] a low profile in designing with her cl...
 ...used to `visit the sick in their beds; [[kept]] a regular list of the industrious poor ...
... tangible as well as intangible ways. I [[kept]] an occasional journal during the war y...
 ...n Bombay. His certificate, however, he [[kept]] and was to treasure all his life. His s...
...join together two things that are often [[kept]] apart: lyrical effusions and public poe...
...re and more elaborate and my mother was [[kept]] busy making altar cloths, stoles, and ...
 ... in eight acres of land, on which were [[kept]] cattle, pigs, and poultry, and became a...
 ...d contentedly And ever with his dreams [[kept]] company. To-day, the music of the slow,...
 ... particular favourites, but these were [[kept]] for special occasions. Both my father a...
 ...Jew with a pencil stuck behind his ear [[kept]] gesticulating with his hands and talkin...
...chingly done _ a frail unpublished poet [[kept]] going by his wife and son (the wife is ...
 ...editor who seldom wrote editorials and [[kept]] his own literary opinions in the backgr...
 ... a suggestion for sugarlumps, which he [[kept]] in his pocket. When he got two out on t...
 ...he work. It was typical of him that he [[kept]] in touch with the boys themselves and v...
 ... probably made some mistakes because I [[kept]] no journal to which I could refer. In s...
...g ones. Scores were always meticulously [[kept]], not for the evening only but from wee...
...was where a delivery pony had once been [[kept]]. Now Midnight lived a life of luxury, ...
 ...to Star. Our friends and acquaintances [[kept]] on multiplying so there was always a st...
 ...g artist colleagues would go. Mr. Cret [[kept]] prodding me in this direction, unsucces...
...nched away: _ and still she comb'd, and [[kept]] Sighing all day _ and still she kiss'd,...
... on account of the dark yew hedges that [[kept]] the garden from prying eyes, and a fin...
...ry, don't they? But they're the same. I [[kept]] the past alive out of a desire for rev...
     ...al du Moulin de la Galette Lautrec [[kept]] the dance, with its implications of un...
... restraint in form or materials, always [[kept]] the house a backdrop _ rather than a c...
...eft the Technical School, but its staff [[kept]] their interest in him. When he joined ...
 ...tionalized educational discrimination, [[kept]] their salaries low, and further restric...
...t what it felt like to ride a horse. He [[kept]] thinking of Lucky Lady Firelight and h...
... would have been better if our dear boy [[kept]] to pencil and charcoal [drawings] whic...
...uch formulations. Commentary on form is [[kept]] to a minimum and is almost entirely su...
...t least he was supposed to walk, but he [[kept]] trotting. Nails gritted his teeth and ...
 ...y happy but I am not wild. We are both [[kept]] under great restriction. We got up a la...
...oe, English schoolboys. The jargon they [[kept]] up was delicious for mc to hear. The g...
...unts of witnesses confirm, show that he [[kept]] up with the developments in the work o...
 ...of the pines or the image of Henriette [[kept]] Wilfred awake. On Easter Monday he was ...
```