Linguistic Approaches to Literature:

The Macrocosm and the Microcosm: The Corpus and The Text

Ramesh Krishnamurthy (COBUILD, University of Birmingham) 23.12.93

1. INTRODUCTION

My aim in this paper is to see what happens when a large language corpus is compared with a single `unknown' text and what features of the text are revealed. It is hoped that this approach may suggest ways of introducing a greater degree of objectivity into the field of literary analysis.

2. THE CORPUS

Cobuild, a joint project run by Birmingham University and HarperCollins Publishers, has been collecting modern English written and spoken texts for several years, and making extensive use of computers to analyse the language and produce a range of reference books for learners of English. In the process, various computer programs and computational strategies have been developed, which have been instrumental in this study.

Two corpora are referred to: the first (henceforth CORPUS 1) is a 16.78 million word corpus of written texts from the 1970s and 1980s. It was specially designed for EFL purposes, and took approximately six years to build (for further details, see *Looking Up*, ed John Sinclair, Collins, 1987).

The second corpus (CORPUS 2) consists of 121 million words of mainly post-1985 texts, broadly similar to the first corpus, but with spoken material. The selection procedure has been less rigorous because of the amount of data and the speed of collection, but a wide variety of texts has been included, so no one text or text-type should skew the general linguistic picture too drastically.

3. THE TEXT

The text to be scrutinized is the poem `Spring' by Philip Larkin, written in 1950. Using a corpus of texts from the same period as the poem might yield different results, as might a corpus consisting exclusively of poetry. However, such corpora were not available to me.

I would like to thank my colleague Alex Collier of Birmingham University for his assistance in some of the corpus analyses used in this paper.

Spring

Green-shadowed people sit, or walk in rings,
Their children finger the awakened grass,
Calmly a cloud stands, calmly a bird sings,
And, flashing like a dangled looking-glass,
Sun lights the balls that bounce, the dogs that bark,
The branch-arrested mist of leaf, and me,
Threading my pursed-up way across the park,
An indigestible sterility.

Spring, of all seasons most gratuitous,
Is fold of untaught flower, is race of water,
Is earth's most multiple, excited daughter;

And those she has least use for see her best,
Their paths grown craven and circuitous,
Their visions mountain-clear, their needs immodest.

4. COUNTING LETTERS

A fairly simple computer program can count the characters in a corpus or in a text. As far as the computer is concerned, the corpus is just a large text. A standard computer `sort' program can put the list of characters in frequency order, that is with the most frequent character first and the least frequent last.

In CORPUS 1, the most frequent letter of the alphabet is `e', which occurs over 10 million times, then `t' (nearly 8 million), `a' (nearly 7 million), and so on down to `z' which only occurs about 60,000 times.

```
MOST FREQUENT   >--------------->   LEAST FREQUENT

e t a o i n s r h l d u c m f w g y p b v k x j q z  (CORPUS 1)
```

Using the same programs for the poem, we find that the text (including the title) consists of 619 characters. There are 52 occurrences of `e', 47 of `s', and so on.

```
MOST FREQUENT   >--------------->   LEAST FREQUENT

e s a t i r n l o h d g u c m f p w b k y v x        (`Spring')
```

Comparing the two lists, several differences become apparent. I will only deal with the most obvious one here: the letter `s' has moved up from 7th most frequent letter in CORPUS 1 to 2nd most frequent in the poem (other major differences that could be explored are: `g' moves up from 17th in CORPUS 1 to 12th in the poem, `o' moves down from 4th to 9th, and `j', `q' and `z' are absent from the poem).

39 words out of the 99 words in the poem contain `s'. The recurrence of consonants, termed alliteration in literary analysis, usually focusses on word-initial consonants. 10 words in the poem have initial `s':
see
seasons
she
sings
sit
spring (x 2)
stands
sterility
sun

However, the letter `s' can have various phonetic realizations, and the Cobuild corpora have not been phonetically transcribed, so any further analysis of alliteration would currently have to be

done manually. I will therefore merely note that 9 out of the 10 words above have an initial /s/ sound, only `she' does not. If one were also to look at the words with `s' in non-initial position, one would find additional reinforcement of the /s/ sound (for example, `most', `across', `best', circuitous', and so on).

A colleague has pointed out that, since the poem is written in the third person singular of the present tense, a high incidence of word-final `s' is inevitable. On one level, this is a chicken-and-egg question. Could one not equally argue that the poet may have chosen to write in the third person present singular precisely in order to enhance the alliterative effect? On the phonetic level, this objection is invalidated by the fact that the final `s' of the relevant words in the poem are mostly pronounced /z/: `stands', `sings', `is', `has'. Only `lights' adds to the /s/ count.

If one were to count the letters in adjacent pairs, triplets, and so on, one would no doubt discover more of the poem's alliterative and assonantal effects (for example, note the number of words beginning with `th' and `gr' in the next section).

5. ALPHABETICALLY SORTED WORD-LISTS

A different perspective on word-initial letters is obtained by asking the computer to produce an alphabetically-sorted list of the words in the poem. Treating the hyphenated words as separate items, and again ignoring phonetic variation, we discover new information. Although the letter `t' actually moved down from 2nd place in CORPUS 1 to 4th in the poem in terms of frequency (see above), 13 words in the poem begin with `t', and in all of them `t' is followed by `h'. The repetition of words in the poem (which will be taken up later) is also highlighted, as is the fact that most of the `th-' words are grammatical (or function) words - and a cursory inspection shows that the only non-function word, `threading', is also the only word to show a different pronunciation of `th'.

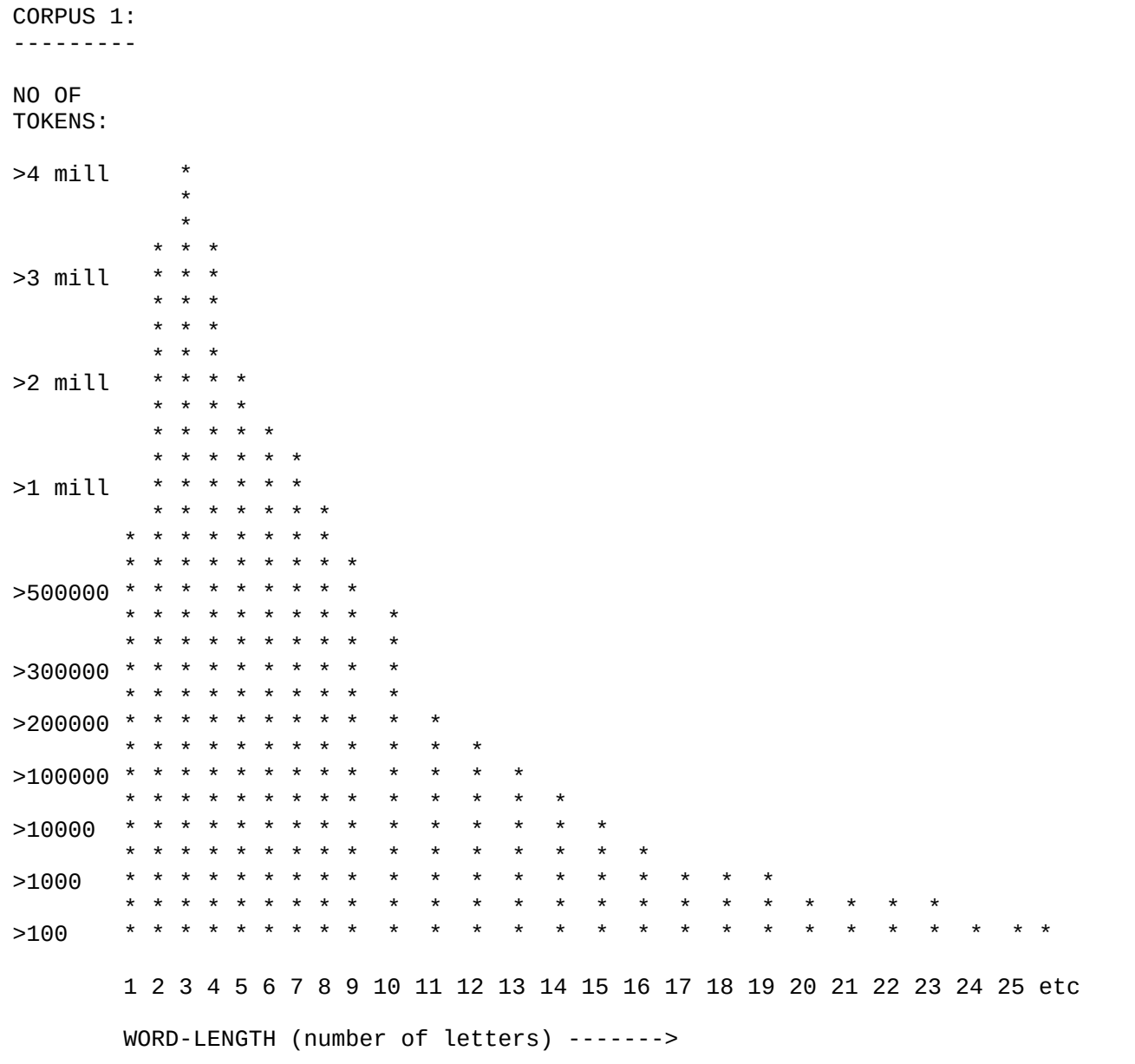
The full list is as follows:

```
13 that that the the the the the their their their their those threading
12 a a a across all an and and and and arrested awakened
11 seasons see shadowed she sings sit spring spring stands sterility sun
7 calmly calmly children circuitous clear cloud craven
7 me mist most most mountain multiple my
6 balls bark best bird bounce branch
6 immodest in indigestible is is is
5 finger flashing flower fold for
5 glass grass gratuitous green grown
5 leaf least lights like looking
5 of of of of or
4 park paths people pursed
3 dangled daughter dogs
3 untaught up use
3 walk water way
2 earth excited
2 has her
2 race rings
1 needs
1 visions
```

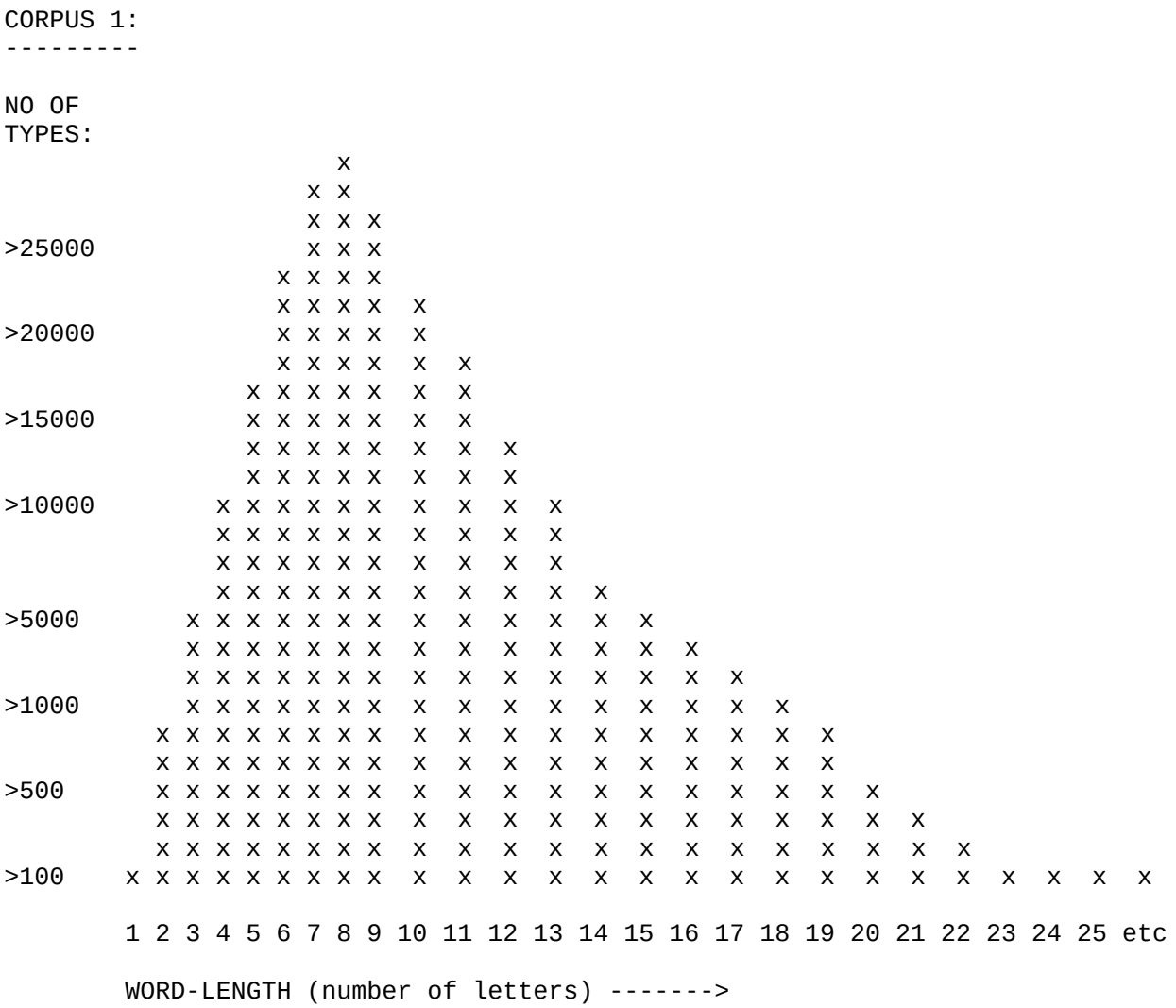
6. WORD LENGTH

It is necessary to introduce the technical terms `tokens and `types' at this point. `Token' corresponds to our normal idea of `word' when counting. A 10,000-word essay contains 10,000 tokens, the poem `Spring' contains 99 tokens, and CORPUS 1 contains 16.78 million tokens. `Type' refers to what we might more casually call `different words' or `different word-forms': am, then, never, table. Thus the sentence `The cat sat on the mat' contains 6 `tokens', but only 5 `types', because there are two tokens for the type `the'. Corpora therefore naturally contain vastly more tokens than types. In CORPUS 1, for example, there are 16.78 million tokens but only about 220,000 types.

Now, these 16.78 million tokens contain 82,689,780 alphabetical characters. By simple division, the average length of an English word is therefore 4.93 letters. However, this is too crude a picture of the language. A closer look shows that nearly 4 million out of the 16.78 million tokens, or approximately 24%, are three-letter words. Over 3,160,000 (19%) are four-letter words, about 3,150,000 (19%) are two-letter words, just over 2 million (12%) are five-letter words, and so on.

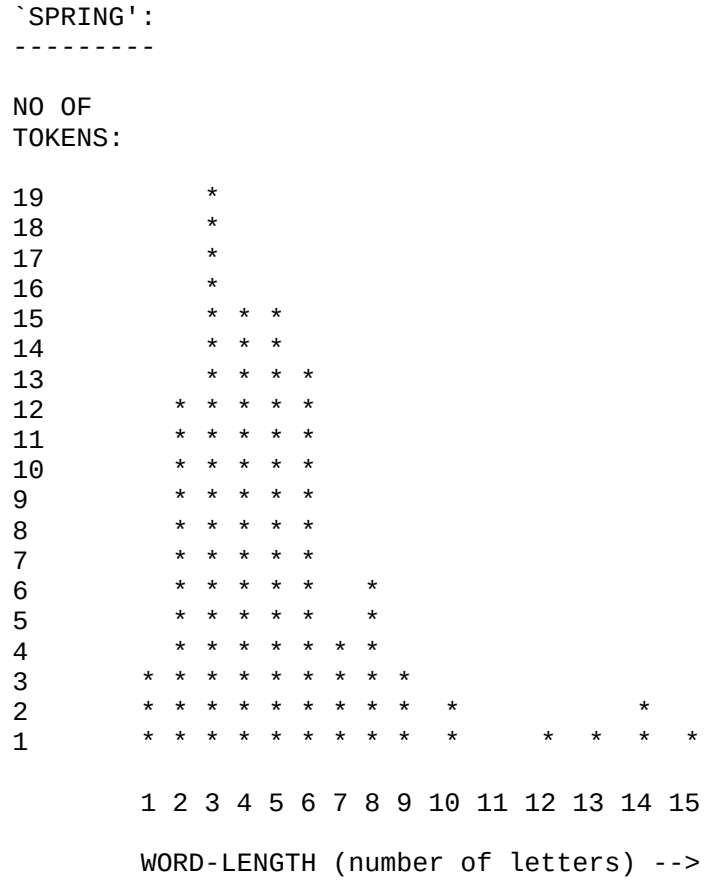


If we look at it in a different way, of the 220,000 types in the corpus (or, if we accept CORPUS 1 as representative of the language, in the `core' general vocabulary of the English language) about 28,000 (13%) are 8-letter words, 27,000 (12.5%) are 7-letter words, 26,000 (12%) are 9-letter words, 24,000 (11%) are 6-letter words, and so on.

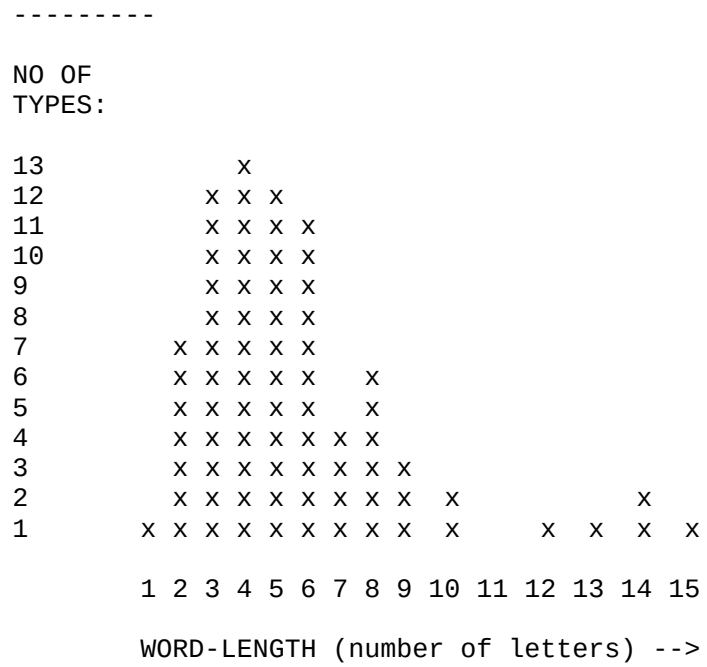


Another interesting pattern also emerges from this analysis: the number of tokens rises from one-letter words up to three-letter words (the `token-peak' if you like) then falls regularly. There are obviously only 26 one-letter types (`a', `b', etc), but then the number of types keeps increasing - 752 two-letter types (`of', `to', `in', etc), 4452 three-letter types, and so on - until the `type-peak' is reached with 8-letter types. Thereafter, the numbers fall again regularly.

A similar analysis of the poem looks as follows. Note that the `token-peak' is at three-letter words, as for CORPUS 1. But the longest word is only 15 letters (`branch-arrested'), and the rest of the distribution is understandably patchier than the corpus, given the minuscule size of the poem in comparison.



The situation with types is more anomalous. The type-peak is at 4-letter words; 5-letter and 3-letter words follow, with a mini-peak at 8-letter words as a slight echo of the distribution in CORPUS 1.



The poem is thus seen to have mostly shorter words than CORPUS 1. In general, experience and intuition suggest that texts containing short words are likely to be typical of informal spoken language and those with long words of formal written language. This is borne out by calculations

on CORPUS 2. The spoken data has the shortest average word-length, newspaper words are slightly longer, and books have the longest average word-length. So one might typify this poem as being `informal' and closer to speech than writing in respect of word-length.

A syllabic count might be more accurate than a merely orthographic one, but would need to be done manually, and has therefore not been attempted.

7. COMMON AND UNCOMMON WORDS

Let us next look at word frequency, that is which words (`types') occur least frequently and most frequently in the corpus, and compare the situation in the text.

As we have already mentioned, CORPUS 1 has about 220,000 types. Of these, about 117,000 occur only once (the term `hapax legomenon' – or more simply `hapax' - is used to refer to words occurring only once in a text or a corpus). CORPUS 2 has about 475,000 types (and it is interesting to note in passing that a 6-fold increase in corpus size yields only just over a 2-fold increase in the number of types), and again roughly half of these (slightly fewer in fact, about 214,000) occur only once.

The poem contains 82 types, 72 of which occur only once. The proportion of hapaxes is understandably greater in a smaller text, with less opportunity for repetition.

The top ten items in the frequency lists for CORPUS 1 and CORPUS 2 (i.e. the commonest words in the corpora) are:

CORPUS 1: no of occurrences (approx) (out of 16.78 million words)		CORPUS 2: no of occurrences (approx) (out of 121 million words)	
the	1,000,000	the	7,000,000
of	500,000	of	3,300,000
and	480,000	to	3,100,000
to	450,000	and	2,900,000
a	390,000	a	2,600,000
in	310,000	in	2,300,000
that	190,000	that	1,300,000
was	190,000	s	1,200,000
it	180,000	it	1,100,000
i	170,000	is	1,100,000

(Capital letters are treated as lowercase for frequency counts; `s' appears in the CORPUS 2 list because anything after an apostrophe is now counted as a separate word).

It is interesting to note that 8 out of the ten words in the lists are the same. However much we increase the size of the corpus, the commonest words seem to stay in roughly the same relative position (`was' and `I', though absent from the top ten in CORPUS 2, are actually 12th and 13th respectively, so they have not fallen far).

In the poem, only 10 words occur more than once. In frequency order, they are:

POEM: no of occurrences (out of 99 words)	
the	5

and	4
of	4
their	4
a	3
is	3
calmly	2
most	2
spring	2
that	2

So, even given their vastly disparate sizes, the corpora and the poem have five items in common in their `top tens': `the', `of', `and', `a', `that'. However short the text, it is unlikely to be without these function words.

The other five items in the poem's multiple-occurrence list occur in the Cobuild corpora as follows. Frequency figures are not the only way of indicating how common or uncommon words are, so I will give the positions of these items in the corpus frequency lists instead:

position in	CORPUS 1:	CORPUS 2:
is	12th	10th
their	40th	44th
most	87th	91st
spring	1340th	1607th
calmly	5524th	9281st

Note that these words remain in the same relative positions in the frequency lists for the two corpora.

The poem's 72 other types (i.e. the hapaxes in the poem) all occur in both the corpora, but it is not worth looking at all of their ranks or frequencies here. However, it is interesting to look at words in the poem which have the lowest corpus frequencies, i.e. the least common words that the poet has chosen to use.

CORPUS 1:	number of occurrences	position (approx)	CORPUS 2:	number of occurrences	position (approx)
sings	43	18,000th	awakened	412	16,000th
shadowed	39	19,000th	pursed	185	25,000th
sterility	34	21,000th	dangled	143	29,000th
craven	26	24,000th	sterility	127	31,000th
indigestible	19	29,000th	gratuitous	116	32,000th
threading	19	29,000th	indigestible	83	38,000th
gratuitous	16	32,000th	circuitous	83	38,000th
circuitous	15	33,000th	threading	79	40,000th
immodest	13	36,000th	immodest	30	62,000th
untaught	8	47,000th	untaught	9	110,000th

The preponderance of longer words at the bottom of the list is evident, and bears out the comments made above about word-length: rarer words also tend to be longer.

But although these are the `rarest' words that the poet has chosen to use in this poem, most of them actually belong to the `core' vocabulary of English. For example, all of the items from CORPUS 1 except `untaught' are headwords in the Collins Cobuild English Language Dictionary, a dictionary for learners of English which was based on that corpus.

There are some variations between the corpora: `sings', `shadowed', and `craven' are not as rare in CORPUS 2 as in CORPUS 1. This is not surprising, perhaps, in the case of `sings' and `shadowed'. But I must admit I was astonished to find that `craven' was more common in the more recent corpus. However, this turns out to be somewhat of a red herring.
The greater proportion of journalistic material in CORPUS 2 has increased the number of proper names, and many of the citations are in fact for a journalist named `Nick Craven', the South African Rugby President `Danie Craven', the pop singer `Beverley Craven', and `Craven Cottage' (the home of Fulham Football Club), not to mention a few `Craven Roads' and pubs called `The Craven Arms'.

The newcomers in CORPUS 2 to the `rarest-words-in-the-poem' category are `awakened', `pursed', and `dangled', none of which seem too surprising. The failure of `untaught' to make any impact despite the six-fold expansion in corpus size justifies the decision to omit it from the Cobuild dictionary, and earns it the title of `rarest' word in the poem. Yet few learners, let alone native-speakers, would have any difficulty in understanding its literal meaning. Its actual meaning in the context of the poem is another matter, and I will not enter into a semantic debate here.

However, it is perhaps worth pointing out at this stage that none of Cobuild's programs is yet able to take account of meaning, so `rarest' does not necessarily entail `most-difficult-to-understand'. Meaning arises out of context, and as we shall see in the next section, it is perhaps in terms of contextual environment that the poem most sharply differs from the corpora.

8. WORDS THAT GO TOGETHER: COLLOCATION

Language is not a random system. In English, some words tend to go together, others do not. To use an oft-quoted example, although `strong' and `powerful' are sometimes interchangeable, we say `strong tea' and `a powerful car' but not usually `powerful tea' and `a strong car'. This feature of language is called collocation, and in corpus analyses is usually defined as `the significant co-occurrence of two words'. That is, if one of the words occurs, the other is likely to be found in its proximity. For example, if `kith' occurs, `kin' will not be far away, so `kin' is termed a `collocate' of `kith' (note, however, that the reverse is not true to the same extent: `kin' is a more independent word, and can occur in contexts without `kith'; the co-occurrence is thus `more significant' for `kith' than for `kin').

An obvious indication of a close relationship between two words is hyphenation. After all, many hyphenated words can also be written as a single word. There are five hyphenated items in the poem (`green-shadowed', `looking-glass', `branch-arrested', `pursed-up', and mountain-clear), so let us start by looking at these.

The simplest procedure is to check if they occur in the corpus wordlists. A quick glance tells us that none of the four occur as one word in either corpus. For example, here are the relevant sections of the wordlists for words beginning with `greensh':

```
CORPUS 1:                    CORPUS 2:

greensboro 3                 greensboro 73
greensh 1                    greensborough 2
greenshaded 1                greensburg 11
greenshank 1                 greenscam 1
greenside 2                  greenset 1
                             greenshank 24
                             greenshanks 2
```

```
                              greenshields 4
                              greenshirts 1
                              greensick 1
                              greenside 16
```

Next, we can look for hyphenated forms. Here again are the are the relevant sections for words beginning with `green-sh':

```
CORPUS 1:                     CORPUS 2:

green-ruffed      1           green-salad   1
green-shaded      5           green-seal    1
green-shirted     1           green-shaded  4
                              green-shoe    1
```

Of the five hyphenated items in the poem, only `looking-glass' is well-attested in the corpora: 13 occurrences in CORPUS 1, 46 in CORPUS 2. `Pursed-up' occurs once in CORPUS 1, but not at all in CORPUS 2, and `branch-arrested' and `mountain-clear' (like `green-shadowed') are not attested in either corpus.

But perhaps the items occur as adjacent words, without a hyphen. Cobuild can and has generated lists of all pairs of adjacent words in a corpus, but as can be imagined, the output is enormous and is dominated by function words. For example, in a 5-million-word corpus, the first few items were as follows:

```
of+the      40683
in+the      30826
to+the      13805
for+the     11810
on+the      11109
to+be        9766
at+the       9529
```

So at this stage, we cease to use the wordlists. Another basic tool of corpus analysts is the concordancing program. This usually presents all occurrences of a selected word (the `keyword' or `nodeword'), with some surrounding context (therefore usually called KWIC concordances, for `Key-Word-In-Context'). Most concordancing programs have the facility to sort the concordance lines by the word to the left or right of the keyword, so adjacent co-occurrences are easy to spot: for example, there are 13 occurrences of `looking glass' in CORPUS 1, and 117 in CORPUS 2; 5 occurrences for `pursed up' in CORPUS 1, and 3 in CORPUS 2. Here are the relevant concordance lines for `pursed up', obtained by sorting concordance lines for `pursed' by the word immediately to the right:

```
CORPUS 1:
  sofa beside her. Her little mouth was pursed up tight and there was a whitene
 gh there had ever been any doubt. Soso pursed up her face into the shape of a
 ody was waiting for them. Instead, she pursed up her bloody mouth, and narrowe
  coming to hear me play." Her face was pursed up in a gross vegetable shape; s
 ll your poem be about?' Rhoda at first pursed up her mouth; then she said: "Ab

CORPUS 2:
ES> about something; his small mouth pursed up and slightly open, his nostrils d
n Mrs Zuckerman's will?" <LTH>?" She pursed up her mouth, and looked very Sunday
FCH> if <FCH> I did -- <CQ0> Here he pursed up his lips, and looked so solemn an
```

There is no occurrence of `mountain clear' in CORPUS 1, and only one in CORPUS 2:

```
There was a warm wind from the ocean and it would soon blow the mountain clear
of cloud.
```

However, there are still no lines for `branch arrested' or `green shadowed' in CORPUS 1 or CORPUS 2. This made me wonder whether the requirement that the two elements be adjacent to each other in the general language of the corpora was too exigent. Perhaps the poem merely reflects altered proximity, rather than altered collocation.

The Cobuild concordancing program allows us to increase the amount of context to 512 characters (about 100 words). In such a large context, many words will co-occur, but their relationship is likely to be more than somewhat tenuous. In fact, even by selecting a 512-character-context, the 39 concordance lines for `shadowed' in CORPUS 1 do not throw up a single occurrence of `green'. And in CORPUS 2, among the 224 lines for `shadowed', we find `green' co-occurring in several passages such as the following, which prove that the concern about over-extending the context was justified: (my capitals)

```
...decay and over-sweet. <LTH>. Behind him lay the park of Purslem Manor, the
grass already GREENer for yesterday's rain, the smooth slopes bright and quiet
in the early sunlight. He felt an intense reluctance to shut this out, and to
step into the small SHADOWED room, to be alone with Paul, towards whom, in the
sense that he himself was a part, perhaps the most important part, of Paul's
tragedy, he had felt a horrified sense of guilt. <LTH>. But if this was murder,
these were not feelings which should be...
```

For normal purposes, Cobuild adopts the criterion (based on previous research and our own experience) that co-occurrence is only worthy of note if it takes place within 4 words to the left or right of the keyword. Looking again at the 224 lines for `shadowed' in CORPUS 2, 3 instances of `green' are found within a 9-word-span. The `faces shadowed green' in the second line is the nearest we have got to evidence for the poem's `green-shadowed people'.

```
 friends... Every man's door being shadowed with green birch, long fennel, St.
 reds, startling yellows and faces shadowed green; it is too easy to be seduced
pinned to the front. Her eyes were shadowed in the same green as the turban. Sh
```

Using the same method, out of 8154 lines for `arrested' in CORPUS, 5 contain `branch' in close proximity to it, but both words are being used with different meanings to those in the poem:

```
the anti terrorist branch he hadn't arrested a single terrorist and that he had
<  August alone, Wells Fargo guards arrested three gangs outside branches      >
former Sumitomo branch manager, was arrested last week on suspicion of         >
<at one of its Tokyo branches, were arrested for tax evasion. Tokyo public     >
Toyo Shinkin branch manager who was arrested along with Ms Onoue for issuing to
```

So `green-shadowed' and `branch-arrested' are certainly major deviations from the collocational norms of large corpora. Further evidence of collocational deviance is that the two lexical items which occur in the same line of the poem as `looking-glass' (`flash' and `dangle') almost never co-occur with it in the corpora (there is in fact one line in CORPUS 1: `he saw an image flash across the looking-glass of the wardrobe door'). Similarly, not only does `pursed-up' not occur in the corpora, and very few of the lines for `purse/purses/pursing/pursed' co-occur with `up', but the main collocates of `pursed' (`lips' and `mouth', evident in the lines cited earlier, which occur 167 and 15 times respectively in the 185 lines in CORPUS 2) are absent in the poem, which instead gives us `threading my pursed-up WAY'.

Reviewing the wordlists near the beginning of this section, I noticed that although there is no evidence for `greenshadowed' or `green-shadowed', there was one occurrence of `greenshaded'

and several occurrences of `green-shaded'. I conjectured that the poet may have substituted `shadowed' for `shaded'. But in both corpora, `green-shaded' is used only of lamps, lights, and rooms, never of people:

```
CORPUS 1:
 ella stand and an antique desk under a green-shaded Victorian lamp. Across one
  than that. Stein leaned to direct the green-shaded desk light into the safe's
 n'. (( 44 )) Tuesday, November 5th THE green-shaded lamp in Dawlish's office i
 They brought the baby to be fed, under green-shaded light, a virtually weightl
 nd regained consciousness. He was in a green-shaded room. He was under water.

CORPUS 2:

   desk with a silver inkstand and a green-shaded lamp. He had sat at the
The library was lit only by a single green-shaded lamp. Gerald Lovell took
    across from Perigord's desk. The green-shaded lamp cast an amber circl
      and piles of unsorted books. A green-shaded lamp provided the only
```

There is one further approach I would like to introduce. Recently, Cobuild has developed computer programs which calculate the `significance' of co-occurrence by statistical methods. Word X occurs so many times in the corpus, word Y occurs so many times, so we can calculate how many times X is `expected' to occur within 4 words of Y on a purely random basis. If X actually occurs more often than expected in the vicinity of Y, we can say that X is a `collocate' of Y. And the ratio of `actual' occurrences to `expected' occurrences can be taken as a measure of X's `significance' as a collocate.

The poem itself constitutes too small a body of data to generate any reliable text-internal collocational statistics. However, we can take the words of the poem, look at the collocates those words have in the corpora, and then see whether the corpus collocates are present in the poem.

As the title of the poem is `Spring', let us see what collocates this word has in the Cobuild data. It occurs 1216 times in CORPUS 1. The collocation program's full output consists of over 150 collocates in order of significance. Let us confine ourselves to the top 25:

<u>Main collocates of `spring' in CORPUS 1:</u>

```
1963
1968
afternoon
autumn
back
barley
blossom
bubbled
buds
championships
cleaning
comes
day
drier
during
early
every
flower
flowers
following
grass
```

```
green
hares
hot
its
```

The first two collocates, `1963' and `1968', remind us that in the general language `spring' is often used in time references. For example, CORPUS 1 has many lines like these:

```
By the spring of 1963 he was treasurer of the OAS.

Wendy quickly became pregnant again and was due to give birth in the
early spring of 1968.
```

The poem is not located in historical time, so no year is mentioned. Neither is the time of day specified, so `afternoon' and `day' from the corpus collocate list do not occur in the poem (we are left to deduce from `shadow', `cloud', `sun', etc that it is daytime).

So are any collocates of `spring' in the corpus represented in the poem? The ones evident above are `flower', `grass', and `green'. Others in the full list of 150 are `sun', `water', and `leaf'. So the poet has in fact selected 6 of the most significant collocates of `spring' in CORPUS 1. This analysis does not imply that the other words in the poem never occur in the vicinity of `spring' in CORPUS 1, merely that they do not occur sufficiently often for the program to consider them to be `significant'. The point made earlier that none of the Cobuild programs mentioned in this paper are as yet able to take any account of meaning is relevant here, because the collocation program, for example, will have included in its calculations instances of `spring' used with the meaning `to leap' and `a natural source of water' as well as `the name of a season'.

9. CONCLUSION

Let me briefly review the main findings of this paper. Comparing character frequencies in CORPUS 1 and the poem revealed the significance of the letter `s' in the poem, and its alliterative function. An alphabetically-sorted wordlist of the poem further indicated the prevalence of function words beginning with `th'. A comparison of word-length showed that although both corpus and poem had a `token-peak' at 3-letter words, the poem had its `type-peak' at 5-letter words as against 8-letter words. Or, to put it another way, the poem generally uses shorter words, which makes it resemble spoken data rather than written. Word-frequency comparisons showed that both poem and corpus share the most frequent items: function words such as `the' and `and'. However, even the `rarest' words used in the poem were not uncommon in the corpus. The examination of collocation probably revealed the greatest differences between data and poem: four of the five hyphenated forms in the poem had almost no support from the corpus. Yet 6 collocates of `spring' in the corpus were found in the poem.

Thus we have seen that every corpus tool or methodology used was able to give us some insights into the language of the poem, whether they served to confirm or cast doubt on our intuitions. I reiterate the general caution expressed at the outset that, given a corpus of a different period or composition, different results might have been obtained. However, in time, corpus sizes will increase even further, text-selection procedures will become more sophisticated and sensitive, and methods such as the ones outlined above will be improved and new ones developed. In the process, our awareness of the nature of the relationship between language and literature will undoubtedly be greatly enhanced.