DRAFT

Antoinette Renouf English
Language Research University of
Birmingham

THE EXPLOITATION OF A COMPUTERISED CORPUS OF ENGLISH TEXT

A collection of texts, or 'corpus' can be processed by computer to produce information, both statistical and linguistic, which is of use to the language teacher, the materials writer, the lexicographer and the linguistic researcher, in so far as these are distinct.

Simply by identifying the linguistic elements which occur in large amounts of text, singly and in combination, and by indicating the relative degree of frequency of each phenomenon the computer can make apparent important facts about the language which are often unavailable to the naked eye.

A computer is not restricted to the selective and arbitrary perceptions of the human, and this makes it such a valuable counterpart to the intuitive analyst. It can be made to process a text exhaustively, and thereby focus the attention of the analyst on features which would otherwise have passed him or her by.

This paper will present some type of computerised output which can be the extracted from a corpus. accompanied by the occasional explanation and brief comments on possible and actual applications.

The Birmingham Collection of English Text (Renouf. 1984) will inevitably be a focal point in this paper. The corpora referred to will be the Birmingham 7.3 million-word 'Main' Corpus, of written and spoken language: the one million-word 'Subcorpus'. which can be accessed online for interactive use: the approximately 13 million-word 'Reserve Corpu.'. of written language: and the one million-word 'T.E.F.L.
Corpus'. of language teaching course books. But reference will also be made to other major work in the area where appropriate.

1. Character and Word (or 'Token') Counts

Beginning with the smaller units of textual organisation, the computer can provide basic statistics on the composition of text in texts of character (letter). morpheme, word or sentence. It operates on the basis of units or 'bytes' of electronic memory, each of which stores a character of text or a 'terminator', which is an element of punctua-

tion. An allowed set of characters and terminators can be specified, and the larger units of text have to be defined in relation to them: words as character strings bounded by spaces or punctuation, and so on.

Sample output is here edited from a demonstration package of programs for textual analysis on the microcomputer which was developed at Birmingham in 1984 by Professor Yang, of Shanghai University. The statistics relate to a short narrative text:

Extract 1

TEXT ANALYSIS STATISTICS

Total no. of words (tokens)	K 196	
Total no. of words (types)	K 122	
Total no. of characters	= 935	
Average word length	= 4.77	
The largest word length	= 14	
Total no.of sentences	= 8	
Average sentence length	=24.50	
The largest sentence length	= 30	
No. of sentences with fewer than 10 words	= 0	0.00 %
No. of sentences with 11 to 20 words	= 1	12.50 %
No. of sentences with 21 to 30 words	= 7	87.50 %
No. of sentence. with 31 to 40 words	= 0	0.00 %
No. of sentences with more than 41 words	= 0	0.00 %

The first two statistics above distinguish between the number of 'tokens' and 'types' in the text. 'Tokens' are the running words which constitute a text, while 'types' are the different word. in a text, in the sense that THE is a different word from AND. In the text referred to above, there are 196 running words, and 122 different words. Thus, the 'type: token' ratio is 122:196.

Another terminological distinction is made in this area of linguistics. The term 'word' refer., in its precise use, to a particular lexical 'form', such as SIT, TABLES or RUNNING. In Extract 1, this is the case, where Professor Yang is concerned with the size of the individual word forms which make up the text.

Sometimes, however. the term 'word' is used loosely, to refer to whichever lexical entity is being dealt with. This might be a lexical inflexion, SINGS, or a base form. SING; or an abstraction, SING, which is understood to subsume all the associated form of the word, namely SING, SINGS, SINGING. SANG. SUNG, and optionally also the derived forms SINGER. SINGERS. (Incidentally. a more precise term is used in computational linguistics for the lexical abstraction just mentioned: it is 'lemma'. The process of grouping associated lexical forms is. correspondingly referred to as 'lemmatisation', and will be demonstrated later.)

It is also the case that the term 'word' in this field of study refers only to a physical string of characters, and not to the various possible meanings or senses associated with them. There exists, as yet. no automatic means of semantic disambiguation.

Statistics such as those in Extract 1 above are clearly of relevance in the description, evaluation or comparison of texts, in providing an objective method of measurement. They also offer a starting point for various types of lexical research, for example into the nature and role of very short or very long words in text.

2. Word (or 'Word Type') Lists:

2.1. Word Form Lists:

In addition to identifying the word types which make up a text or corpus, the computer can also list them. Such lists are annotated with information are usually about the number of occurrences of each type.

Word listings are routinely organised in two ways. In the first case, word forms/types are ordered or ranked according to the frequency of their occurrence, as shown in the following extract relating to the 7.3 million-word Birmingham Corpus:

Extract 2

WORD LISTING ORDERED ACCORDING TO FREQUENCY OF OCCURRENCE IN THE BIRMINGHAM MAIN CORPUS

309497	the	49186	is	29512	they
155044	of	42057	he	28958	at
153801	and	40857	for	26491	his
137056	to	37477	you	26113	have
129928	a	35951	on	25419	not
100138	in	35844	with	25185	this
67042	that	34 755	as	23372	are
64849	I	29799	be	22445	or
61379	it	29592	had	21916	by
54722	wa.	29572	but	20964	we

or alternatively:

DRAFT

Extract 3

WORD LISTING RANKED ACCORDING TO FREQUENCY OF OCCURRENCE IN THE BIRMINGHAM MAIN CORPUS

1	the	11	is	21	they
2	of	12	he	22	.t
3	and	13	for	23	his
4	to	14	you	24	have
S	a	15	on	25	not
6	in	16	with	26	this
7	that	17	a.	27	are
8	I	18	be	28	or
9	it	19	had	29	by
10	was	20	but	30	we

Word frequency lists reveal certain important facts about the language. They identify the commonest word forms, thereby providing an objective criterion for lexical selection for dictionaries, lexicons, and so on. A current beneficiary of this type of information is the COBUILD English Course (Willis, forthcoming). Frequency lists reveal patterns in the relative frequency of lexical occurrence which have theoretical and practical implications. Zipf's work (1936) on the hierarchy among associated word forms is seminal in this area.

Ranked frequency lists also identify the rarer word forms in a teat, which probably have a series of specific functions of which we are still largely unaware. Mich6le Thomas, a language ‘assistante’ at Birmingham University, is currently studying the textual functions of low frequency and single occurrence words, or 'hapax legemena'. The Rev A Q Morton (1986) carried out a similar study. The COBUILD lexicographic team at Birmingham University has already been able to discover the unexpectedly interesting linguistic behaviour of many of these words (see Hanks, forthcoming).

The second standard method of organisation is an alphabetical one, again with frequency information attached. An alphabetical word list is usually used in conjunction with a frequency listing, as an index.

Less commonly, a word list can be organised according to the order in which word forms first occur in the text. Frequency statistics may still accompany each type, and the effect is as follows (Renouf, ibid):

Extract 4

WORD LISTING ACCORDING TO ORDER OF FIRST OCCURRENCE (based on the first 100 word. of text in 'First Things First', by L G A1eaander, Longman)

excuse	2	you	4
me	2	very	3
yes	4	much	3
is	17	my	4
this	12	coat	2
your	12	and	2
handbag	3	umbrella	4
it	5	here	2
thank	4	sir	2

One obvious use for this type of listing would be in the creation of a language course book, such as that mentioned above, in indicating the point at which a new lexical item is first introduced, and the number of times it is reinforced subsequently.

The extracts shown s far have all been drawn from word 1ists of entire texts or corpora. It is also possible to produce word lists from different part. of the same teat or corpus, as in the following example, taken from the Birmingham TEFL Corpus:

Extract 5

RANK LISTING OF THE INSTRUCTIONAL AND NON-INSTRUCTIONAL LANGUAGE IN THE TEFL CORPUS

	instructional	non-instr.		instructional	non-instr.
the	1	1	write	11	279
.nd	2	6	this	12	23
in	3	7	these	13	144
to	4	3	for	14	14
you	5	4	are	15	19
a	6	2	questions	16	311
of	7	8	or	17	48
about	8	3	with	18	37
ask	9	177	what	19	13
your	10	40	words	20	364

The immediate use for such a presentation is in the comparison of the different elements of the teat. In the case above it serves to point up a difference between the controlled teaching teat and the generally less controlled metalanguage in a series of course books. It has also been used to compare features of the spoken and written components of the Birmingham Corpus.

DRAFT

The statistics of word frequency which emerge from a single text give a picture of the 'coverage' of each word type; that is, of the proportion of the text accounted for by the presence of each type. This profile is usefully supplemented by information on the proportion of text which is contributed by each type across a series of texts; that is, of its 'distributional' frequency. The resultant comparison indicates which types are common to most texts in a given corpus, and which only occur significantly in one or two texts. In the Birmingham Corpus for example, the type TRADE is very fairly widely distributed, whereas many instances of BABY are shown to come from the single book 'Baby and Childcare' by Dr B Spock. Types which have a wide 'distribution' are clearly stronger contenders for inclusion in any language teaching programme.

Distributional frequency information is also being used in the research of Professor Yang (1986) in his efforts to develop automatic techniques for the identification of technical terminology. His starting point is that word types which are frequent in a particular texts but rare in the corpus as a whole, are likely to be the 'technical' terms in that text.

2.2. Lemmatised Word Lists

Lemmatisation. or the grouping of the base and inflected forms of a word under one 'lemma'. can take a number of forms, depending on the particular research purpose. Sandy Harris, a former researcher at Birmingham, employed a particular combination of morphological and semantic criteria in lemmatising text, as part of a procedure he was developing for automatic text compression. a lemmatisation program based on his work, which was written by Ainon Raja Noor, a visiting lecturer in 1984, produces output of the following kind:

Extract 6

LEMMATISED FREQUENCY LIST WITH BREAKDOWN OF FORMS

absolute	804	
		235
	absolutely	563
	absolutes	6
accept	1160	
		48
		9
	acceptance	11
		4
	accepted	44
		4
	accepting	77
	accepts	36

And a rank listing of lemmata in the Birmingham Corpus. with forms grouped according to the same principles. looks as follows:

Extract 7

RANK LISTING OF TOP LEMMATA IN THE BIRMINGHAM CORPUS

a	456871	they	60978	one	26830
be	195263	you	43891	I	26274
both	156768	she	41077	do	26115
of	155044	for	40857	not	25419
to	137056	on	35951	this	25185
in	100138	with	35844	say	24827
he	82025	as	34755	either	23953
it	72080	we	32588	will	23801
have	68799	but	29572	man	23538
that	67042	at	28958	by	21916

It will be noticed in the listing that A has been promoted above THE, normally the commonest word in text. This is because the lemma A is here allowed to subsume the forms A, AN and THE. The grouping principles will vary according to the research purpose.

3. Words in Combination

Thanks to the increasingly sophisticated text processing software which is being evolved, the phenomenon of word combination in text is now able to be put under closer scrutiny, and is becoming a central area of interest. Linguists differ in their view of the relative importance of aspects of lexical co-occurrence; of lemma VB word forms of 'grammatical' vs 'lexical' words, of the degrees of proximity. Whichever ap-

DRAFT

proach is taken, however, the required type and organisation of data can nowadays generally be supplied by fairly standard automatic means.

Corresponding with the various approaches in this field of study, there is a range of terminology. In Rivas (1987), for instance, the term 'collocation' is used to refer to the 'coexistence d'unites consecutivés', the 'contiguite d'une cooccurrence immédiate', and the term 'cooccurrence , to denote 'voisinages plus eloignés apparaissant dans des...contextes plus larges'. We at Birmingham, on the other hand, tend to use the term 'collocation' of word combination in general, but to modify it according to the particular relationship involved. In our case immediately adjacent items form 'contiguous' or 'immediate collocation', and non-adjacent items 'discontinuous collocation'.

The collocational patternings associated with a given word form, or 'node word', can be automatically identified, counted and printed out as follows:

Extract 8

COLLOCATIONAL PROFILE FOR THE WORD FORM .PRETTY' FROM THE BIRMINGHAM MAIN CORPUS (total no. of cases of this word form is 669)

CONTIGUOUS COLLOCATES:

	LEFT-HAND			RIGHT-HAND	
96	cases of match on	A	50	cases of match on	WELL
37	on	WAS	40		GOOD
29		VERY	19		AND
28		IS	18		SURB
27		ARB	16		MUC
24		THE	15		H
17		AND	13		GIRL
7		BE	9		SOON
16		SOME	7		LITTLE
12		I'M	7		FAIR
					YOUNG

Data of this kind has several uses. It provides, for example, a criterion for semantic and syntactic disambiguation. In the extract above, the collocates indicate that in at least 200 out of the total 669 cases of PRETTY the word is functioning as an intensifier, and meaning something like 'FAIRLY'. The profile a1so reveals the commoner collocational partners of PRETTY, information which is of relevance in the selection of linguistic features for pedagogic purposes. The computerised study of contiguous collocation in text is notab1y carried out by Dr Goran Kjellmer, of Gothenburg University. and by Professor Sinclair (1974), who began in the days when it was a significantly more cumbersome task than it is now.

The term 'discontinuous collocation' is generally understood to refer to items which occur at a given distance away from each other in text. They may or may not be functionally related, and opinions differ as to how wide the span between such items can be before their cooccurence is likely to be accidental. At Birmingham, we operate with a maximum span of eight words in the environment of the node word - that is, in the range +4 and -4 to either side.

It is possible to produce a total collocational profile across this range (Sinclair. 1969). Or a partial account might be required; in Extract 9 below, for instance, the computer presents an analysis o every instance of the combination A and OF, divided by a single space:

Extract 9

"A...OF" IN THE BIRMINGHAM MAIN CORPUS
(Items occurring within this framework are presented in descending order of frequency)

1288	cases of match on	LOT
515		KIND
458		NUMBER
383		SORT
353		COUPLE
297		MATTER
281		BIT
217		SERIES
201		PIECE
190		MEMBER

This profile highlight. the highly productive nature of a combination of grammatical or 'function' words like A and OF. and undermines the view that such words are 'not worth teaching'.

In Extract 9. the discontinuous items and the range were prespecified. For other purposes, only two of the three variables might be specified.

It is uncommon to consider collocational patterning without. Positional restriction on the elements in question. This is probably a matter of tradition. However, corpus evidence shows that there is in fact much to be learned from studying 'non-positional' discontinuous collocation; that is to say. the pattern of co-occurrence of the nodeword with any word in its proximity, within a given range but irrespective of position. Lexical attraction often accommodates syntactic or grammatical variation. Whereas, for example, a computerised profile of positional collocation will help to identity fixed phrases containing the nodeword, a statement of non-positional colloca-

DRAFT

tion will also draw in variable phrases, separable phrasal verbs, and so on.

Research into non-positional collocational patterning in a large corpus is currently in progress at Birmingham, with the aim of discovering facts which could, among the other things, inform speech recognition programs.

4. Word. in Context

At the moment. any analysis of the meaning and use of words in a corpus is still largely manual. The computer can help by extracting lines of context, or 'concordance' lines, from the text for each word under investigation, and print these out in a specified order. A set of such lines for a particular word in a text or corpus is known as a 'concordance'.

The standard format for concordance print-out is a series of one line 'KWIC' (key word in context) concordances, organised alphabetically by the first letter of the word to the right of the key-word, or node-word. See the following example:

Extract 10

CONCORDANCE LINES POR 'TURN' IN A ONE-MILLION WORD CORPUS. SORTED BY RIGHT CONTEXT

ghten the gland nut. you needn't	turn off the mains water. just remove
asons. Whatever the cause, first	turn off the water supply to the house
eenagers, when it comes to their	turn. 1: Oh yes 3: We have got another
n last night, but they appear to	turn on some agreement between the US
ou unblock it, replace screw and	turn on water to check it's tight bef
are the real women? Every time I	turn on the television there's a woma
pp. It's terribly good of you to	turn out on a night like this." "You
our brassieres and girdles, it's	turn out that Rudy was downright comp

For a common word like TURN, there will be too many occurrences in a large corpus to cope with manually. In this case, it is sometimes useful to produce a sample of lines, selected from the corpus according to the principles of simple or logarithmic frequency, or randomly. In the sample below, it will be seen that the right-hand patternings show up less clearly than in the previous, full, extract. This can be offset by using the sample in conjunction with a collocational profile like the one in Extract 8, which has been run on all instances of the node word.

Extract 11

RANDOM SELECTION OP CONCORDANCE LINES FOR 'TURN' PROM A ONE-MILLION WORD CORPUS. RIGHT-HAND SORTED

r way flashing showing the right turn and they hold it in just a littl
lming College has been forced to turn away 300 prospective students.
s that breed pride, and pride in turn builds the kind of discipline th
window if a room gets too warm turn down the heat. Don't draw curtai
much of it could take off in its turn for the sunbelt. Average hourly
l exercise which suggests a down turn into an automobile. Ferry the fo
amese civilians and who by night turn into Vietcong, tossing hand gren
e three forces are all trying to turn it round aren't they? S: Yes T:
d be unthinkable for the west to turn its back on the welfare state

Right-hand sorted concordances highlight patterns to the right of the node word, which. as said, is the standard format. This is probably a matter of tradition. bound up with notions of the predominance of the verb and the directionality of language. However. it is useful to complement it with a left-hand sorted concordance. which can reveal other patterns. This can be sorted alphabetically by the first letter of the word prior to the node word. or by the last letter, as shown in Extract 12:

Extract 12

e tail is?... Right kay. We can	turn that backwards, and we put that
ly infuriated by the way men can	turn a trivial point into a real tabl
s that breed pride, and pride in	turn builds the kind of discipline th
eful to discuss each of these in	turn, recognising however that there
l exercise which suggests a down	turn in the secular trend of normal
lming College has been forced to	turn away 300 prospective students.
o, the students will be asked to	turn over their papers and read their
too involved in the business here to	turn against Sadat. They will just st
the middle, so you don't have to	turn the water off at the mains to in
e three forces are all trying to	turn it round aren't they? S: Yes T:

A one-line context is sufficient in many cases, and very convenient to work with, in that it fits a screen, a fiche, or a page. But in the study of certain linguistic features, notably of lexis which has a larger-scale discourse function. it is desirable to work with multipleline contexts. This is quite possible, as shown in Extract 13. The disadvantages are that fewer instances can be juxtaposed for comparative purposes, and that the node word is not quite so conveniently placed:

DRAFT

A SELECTION OF EXTENDED CONCORDANCE LINES FOR 'REASON'. SORTED BY RIGHT-HAND CONTEXT

adaptors or the cast or the producer, but simply that I don't know that anything from Wodehouse would adapt really very well, for a simple reason. which is that if you do a radio play you've got direct speech, you've got actors addressing each other. Much of t

king-size beds with sable quilts, their hand-made automobiles, their sleek girl in every port. Money has got to be the reason, a primary reason anyway, why the insulted umpire sent his officials to beg the tennis star to return to the court and go on

In the study of certain lexis, it is necessary to have access not just to larger contexts but also to large amounts of corpus data. The data shown in the Extracts above is all derived from Birmingham-held corpora of 1 million or 7.3 million words of text. Since about 80% of the word forms even in the 7.3 million-word corpus occur fewer than ten times, it will be understood that still more corpus evidence is required if one is to say anything authoritative about them. In a corpus of almost twice the size, i.e. of about 13 million words. (Renouf. 1987), each word form occurs about three times as often, on average, and this additional evidence is important in endorsing or modifying the initial view of a particular word.

Concordanced data generally carries coded information of some sort. The next Extract illustrates two kinds of linear coding, interlinear and intralinear.

Extract 14

CONCORDANCE EXTRACT FROM THE T.E.F.L. CORPUS

e004 nly fell down. 0> you say: 3> I was	walking down the road or 3>
e010 t, mrs riley, an elderly widow, was	walking along a dark london
e017 were just ordinary postmen, fond of	walking, and dogs and chri
e012 rds 3> smoking typing writing eating	walking reading sitting down
e019 3 > go fishing; go water-skiing; go	walking, etc. 0> Unit 1> whe
e013 0> page three pictures 3> he's been	walking for twenty minutes n
e024 e present. 3 > he's at school. after	walking for several hours
e003 answered. 1 > "I hope so! we've been	walking for twenty minutes n
e003 at the flowers. a young couple were	walking hand in hand. some 3
e013 swimming. he saw them while he was	walking. he ran back 3> to h

This extract is foreshortened to the right, so that the left-hand interlinear coding can be seen. The coding, for example e004, indicates the source of the line which it precedes. The letter 'e' stands for 'English Course Books', those used in the teaching of English and making up the specialised corpus at Birmingham which was referred to previously. The next three digits indicate the particular book from which the line came. In the main Birmingham Corpus, there is a more detailed interlinear coding system, indicating also the page of the source book, the language variety of the author and that of the publishing house. Such coding is undoubtedly useful to the concordance user, but it exacts a space penalty, reducing the amount of context that can be contained in the remainder of the line.

The intralinear coding seen in Extract 14 indicates the language types found in the course books. 'O>' marks the start of metalanguage, whether rubric or page heading. '1>' introduces 'concocted' speech in transcription: invented conversations, and son, as opposed to '2>', which would mark authentic speech in transcription. '3>' precedes 'concocted' writing: text constructed for purposes of exemplification rather than communication; '4>' would mark 'authentic' written text. The purpose of these codings is to facilitate a comparison of the various language components of the books.

A corpus can also be grammatically coded, or 'tagged', by having word-class codes attached to each word form. The BROWN and LOB corpora have undergone this process, and a brief extract of tagging program devised by Jeremy Clear at Birmingham, is offered below by way of illustration:

Extract 15

TAGGED TEXT 'HEART OF DARKNESS' BY J CONRAD

prep	In	adv prep	together
d	the	d	without
ing	offing		a
d	the	n	joint
n	sea	£	and
lnk	and	lnk adprep	in
d	the	d	the
n	sky	adj	luminous
aux	**were**	n	space
pap	welded		

The next stage on from tagging is to parse a corpus, as is happening in Nijmegen; or it can be coded for phonetics and prosody, as in the Lund 'Corpus of English Conservation' (Svartik, 1982).

This paper represents an attempt to give a brief introductory aCCount of current procedures and applications in the field of corpus exploitation. There remains much to be said. about the types of listing and Coding mentioned. and concerning other recent events and applica-

DRAFT

tions. Major changes in the field can also be expected. Very soon, for example, the finite corpus which is a given element throughout this paper will be supplemented by a non-finite 'monitor' corpus, with its attendant new processing and codification strategies.

References

Alexander, L.G 1967. First Things First. Longman: Harlow.

Hanks, P.W (ed.) Forthcoming. Collins COBUILD Enlgish Language Dictionary. Collins: London

Morton, A.Q. 1986. “Once. A Test of Authorship Based on Words which are not Repeated in the Sample” In Literary and Linguistic Computing, Journal of the ALLC. OUP: Oxford.

Renouf, A.J. 1987 “Corpus Development at Birmingham University” In Aaerts, J. and Meijs, W. (eds.) Corpus Linguistics in the Use of Computer Corpora in English Language Research. Adolpi: Amsterdam

Renouf, A.J. 1987. “Lexical Revolution” In Meijs, W. (ed.) Corpus Linguistics and Beyond: Proceedings of the 7th International Conference on English Language Research on Computerized Corpora. Rodopi: Amsterdam

Rivas, M. 1987. “Quelques Aspects Ponctuels Relatifs aux Collocations” In Rivas, M. (ed.) Actes du VIIIème Colloque GERAS. Université de Paris –Dauphine.

Sinclair, J. McH., Jones, S. and Daley, R. 1969. English Lexical Studies, University of Birmingham for the office for Scientific and Technical Information

Sinclair, J. Mch. 1974. “English Lexical Collocations” In Cahiers de Lexicologie. Institut des Professeurs de Francais a l’Etranger: Paris.

Spock, B. 1979. Baby and Childcare. The Bodley Head: London.

Svarvik, J. and Quirk, R. 1979. “A Corpus of Spoken Conversation” In Svartvik, J. and Quirk, R. (eds.) Lund Studies in English 56. CWK Gleerup: Lund.

Willis, J. and D. (forthcoming) Collins COBUILD English Course. Collins: London

Yang, H. 1986. “A New Technique for Identifying Scientific/Technical Terms and Describing Science Texts” In Literary and Linguistic Computing. OUP: Oxford.

Zipf, G.K. 1936. The Psychobiology of Language. Routledge: London.

DRAFT