Sticking to the text: a corpus linguist's view of language

Antoinette Renouf

Director of the Research and Development Unit for English Studies

University of Birmingham

Introduction

Corpus Linguistics is the study of large, computer-held bodies of text, or 'corpora'. In the last five years, this approach to language study has become increasingly popular among linguists, and developments in computing technology and software and in storage mechanisms like CD are making it possible even for the individual PC user. The aim of the linguist is to describe the language, and corpus linguistics reflects the shift in academic focus from the brain to the text as the appropriate source of information. A description derived by introspection will tend to be idiosyncratic and partial, since no individual has total awareness of how they or others use language. A description based on the observation of appropriate corpus data, on the other hand, can provide a broader view of language use, including statements about the relative typicality of individual features based on their frequency of occurrence in the corpus.

On the corpus-linguistic continuum, the study of raw ASCII text is situated at one end, and the study of heavily pre-coded text at the other. The work of the Unit sits very much at the 'raw' end, a philosophical position which is reflected in the title of this paper. Some corpus linguists are concerned with language description for its own sake; we study the patterns of language in order to see how the computer can be made to discover and exploit these automatically, in order to facilitate even larger-scale study, and to solve problems in associated areas, such as text retrieval.

In our approach, since the computer does not move away from the text, we must discover ways in which it can be made to work with what is available in the text itself. Accordingly, the basic units of information are words, singly and in combination, word frequencies, and the positions of words in relation to each other. 'Sticking to the text' in this way leads us to develop systems that do things differently from the way a human would, and I would like to explain how this approach works with reference to some examples of recent work in the Unit.

Recent Research

Using Word Frequency to Identify of Changes in the Lexicon

There is a need in several fields, among them language teaching, information technology and lexicography, for an automatic means of discovering facts about the vocabulary of the language and about how it is changing. In a large DTI-SERC funded project, entitled 'AVIATOR', one of our aims has been to develop systems that monitor such information, and we have now completed the work.

To establish changes in the language, it is necessary to treat text as a chronological flow, rather than as a static entity. Our data flow has chiefly been the Times newspaper, although other types of text, such as BBC World Service data, are also handled. The system consists

of four filters, each of which records different facts.
Filter 1 is set up to identify new words in the language. Since the computer cannot know a priori whether a word is new or not, new words are defined as being those that the computer has not encountered before. In fact, a human would ultimately have to adopt the same criterion. The system compares the contents of the text flowing across it with the words it has already recorded, and marks and dates first and latest occurrences. A sample of output from the 'ordinary words' category will show what is found by this method:

NEW WORDS: TIMES FEBRUARY 1991

Feb-06 rector of LWT, says his boardroom is now debugged before every meeting. "We had a bit of
Feb-14 is attractively handled, and enriched by the gloopy dollops of pastiche knight-speke that
Feb-17 down at any rate." The result is an allobiography (about other people rather than oneself)
Feb-17s his youngest son, Richard Brooks, a buppie (black urban professional) malcontent. The son h
Feb-23 in his tracks, force him to unplug the acoustiguide and stand and stare?
Feb-24 It is readily available from tax returns. Footering with past failure to face an election
Mar-02 rationalise, internationalise, and almost to common-marketise our birds. It is a task
Mar-02 Poets are not very useful, because they are not consumeful or very produceful
Mar-OS national Symbolism to feature the sort of boneless-wonder figure drawing that occurred
Mar-OS tation, Manchester's SOS State represents the dance-trance arm of latterday psychedelia.
Mar-09 pping over others. It is in these unobtrusive arpeggiations that the fire reveals itself.
Mar-23 s, a few crude paintings. A teeny bit dog's dinnerish. The service is occasionally patronising
Mar-24 light of chandeliers. "It's our attempt to de-yuppify the place," said Derek Statt
Mar-24 clash last weekend between England and France. Empurple the prose any way you like, and
Mar-31 ome brethren of the Holy Trinity were keen choirboy-spotters, Hopkins's diaries
Mar-31 publisher, Or Francis, and Thatcher not the ex-leaderene but a certain David. However

What emerges is a range of lexical phenomena. There are bona fide new words, built by the standard rules of word formation, such as the blend 'acoustiguide'. There are some productive items, built by analogy on already extant forms, such as 'buppie' and 'de-yuppify'(after the acronym 'yuppie') or 'choirboy-spotters' (after the compound 'train-spotters'). There are some new inflections, such as 'common-marketise'. However, some of the words that are identified are not actually new, although meeting the automatic criterion of not having appeared in previous text. One example is 'arpeggiations'. This word has not appeared before because it is a fairly specialised technical term, that pops up rarely; in other cases, an 'old' word may suddenly re-enter the language, as with 'poll-tax', or 'ecu'. Some words in the above sample would strike the human as being ephemeral or nonce formations, such as 'consumeful' or '(dog's) dinnerish'; our

system can also identify them as such, by noting when and to what degree they appear, flourish and disappear again: this diachronic monitoring is the function of Filter 4.

Using Word Collocation to Identify Changes in Word Use

Filters 2 and 3 can automatically identify changes in word use. The computer does not know about senses and meanings per se, of course, but we have developed a system which records the collocational environment of a word and compares this with the environment of subsequent occurrences of the word. In order to 'stick to the text', we use the criterion of collocational change to indicate a change in use. The theoretical assumption underlying this approach is that collocation is a type of meaning. This may all sound rather esoteric: an example of our output may make things clearer. The following lines are instances logged by the computer where the word 'charter' has changed its profile from collocating with 'company' or 'plane' and meaning 'specially-hired' or 'cheap', to a new profile with a more sociological flavour:

CHARTER - Profile of Collocational Changes

num	of	the	@social	charter	there would	be a	
the	prime minister's	>citizens'	charter	to	be unveiled next	*OUMMY*	
the	government's planned	>citizens'	charter	*OUMMY*	*OUMMY*		
	OUMMY						
the	proposed european		@social	charter	tx	although	shares
	are						
the	proposed	european	@social	charter	tx	sir	trevor
	holdsworth						
government on the		@social	charter	tx	this	is not	
commitment to the		@social	charter	tx	the mr	labour	party
the the european		@social	charter	when	norman		fowler

that	the	european	@social	charter	which seeks to	ensure	
nor	on	the	@social	charter	which both	went	through
elements	of	the	@social	charter	which risk	addin g	to
is	what	the	@social	charter	would do	to	this
that	as	the	@social	charter	would require	the	abolitio n

In the profile above, 'citizens" is shown as being a newly-occurring collocate, while 'social' is being shown as an established collocate which has suddenly increased significantly its frequency of occurrence with 'charter'. Presented to the researcher, they combine to suggest a shift in meaning. The 'new' sense of 'charter' here is actually a recurrence of the sense found as far back as the Magna Carta.

The computer cannot make the final decision about whether the new uses of the word 'charter' also signify new senses. Ultimately that is a matter of human judgement. However, an automated system can sift through vast amounts of data and create a feasible post-editing task for the human.

Using Word Repetition and Word Positioning in Automatic Abstracting

As ASLIB members hardly need reminding, the human being is capable of reading a text and summarising its message in the form of a new, shorter text. The computer cannot yet do this, and any abstract that

it creates at the moment has to be made up of words and sentences drawn exclusively from the original text. This type of abstract is better termed an 'abridgement'.
Various methods of abridgement are being experimented with by academics and companies such as software houses but results are not likely to be impressive because of the lack of knowledge about the relationship between the words and the ideas in a text. Sentences are extracted on the basis of frequency of occurrence of keywords, or because they contain pre-selected phrases thought to mark important stages in the structure of a text, such as 'the process'. Whilst there might well be some purpose in exploiting the metalinguistic aspect of text, these approaches are still too linguistically naive. The result is likely to be an odd selection of sentences, not all of which summarise a section of the text, and which are not easy to read because they are not related to each other. Keeping to the principle of 'sticking to the text', however, my Unit has developed systems of automatic abridgement that in most cases produce acceptable abridgements. Based on ideas of Dr. Michael Hoey, of Birmingham University, our systems variously trace the patterns of lexical repetition in a text and use this information to select key sentences. Sentences found to be most heavily cohesive are deemed to be core information bearers. The abridgements are not only adequate accounts; they preserve a lexical interrelationship between the key sentences which allows these to be read together as a text.
The following is a 6 sentence abridgement. The original article, of 22 sentences in length is provided for reference in the Appendix .

Article from the Times Newspaper, 31st December 1992 Bullish Lamont offers no early rate cuts

By Peter Riddell and Anatole Kaletsky

31 December 1992

BRITAIN'S economy will do much better next year than in 1992, but there will be no further reductions in interest rates unless growth falls below the Treasury's expectations, according to the Chancellor of the Exchequer, Norman Lamont, in an exclusive new year interview with The Times.

[8] ‘’Mr Lamont’s remarks may, however, disappoint the City, where many investors have been hoping for a further cut in interest rates early in the new year.

[9] The Chancellor said that interest rate reductions would be considered only' 'if monetary demand was manifestly too low".

[10] Asked whether he would expect to change interest rates if the economy performed in line with the Treasury's forecast of 1 per cent growth, the Chancellor replied with an emphatic "no".

[12] Mr Lamont said that Autumn Statement measures for industry and housing, and the big cuts in interest rates and the devaluation of sterling since Black Wednesday had' 'created the right conditions for confidence and growth".

[14] Mr Lamont estimated that as much as two-thirds of the impact of the recent three-point reduction in interest rates was "still in the pipeline" and added that the "very warm welcome" given by industry to his Autumn Statement measures meant that there was "every chance

that they will succeed".

This abridgement was generated using our default settings. A shorter, four-sentence, version of the abridgement contains the sentences, 1, 9, 12 and 14. Although our system does not apply a weighting to any particular section of the text, it tends to select initial sentences in journalistic articles because they are lexically rich and so achieve the required threshold in terms of repetition. This accurately reflects journalistic practice, where the essence of the text is typically summarised in the opening sentence or sentences.

Using Word Clusters in Automatic Text Retrieval

An essential part of helping a database user to select a relevant text is discovering a way of conveying information as to what the text is about. The abstract is an explicit statement of 'aboutness'; an index is a more implicit one. Indexing has long been automated, and ASLIB members will know best what the latest methods of word selection are, and how successful. Generally speaking, automatic indexing will be based on calculations of frequency and/or relative frequency of word occurrence.
Another aim of the Unit's AVIATOR project has been to investigate the relationship between the patterns of words in a text and its conceptual concerns. The research builds on a pilot study done in the early 1980's by Dr. Martin Phillips, a former postgraduate student at Birmingham.
By recording and monitoring the contexts of every word in a text, we can automatically identify lexical 'clusters' that echo the topic or topics of a text. We shall illustrate what is meant by a cluster
by presenting a set of clusters extracted from a book entitled "The Living Planet", for Chapter 1, "Furnaces of the Earth".
The book, by David Attenborough, is written to follow the episodes of a television series. It has an underlying theme of how nature survives and lives on in various hostile environments. Each chapter has a distinct topic. Chapter 1 describes the different processes by which volcanoes are formed, the kinds of devastation they cause, and the way in which natural life returns in the aftermath. The set of clusters looks as follows:

Set of Clusters for Chapter 1 of 'The Living Planet' lava
ash chamber
lava splashes vent thrown river
basalt flows volcanoes erupting ridge
gas smoke steam
kilometres long small krakatau anak
flow currents convection descending
line junction concealed
produced catastrophic explosion
sumatra java emit
krakatau reclaimed century
mallee fowl incubation

This is obviously an implicit statement of what the text is about, and at first sight, it looks rather strange. To highlight its particular features and to evaluate its success in expressing the concerns of the text, we present here a manual abstract and index of the same chapter. This work was commissioned by us from Information Unlimited, and the abstractor, who is not an indexer, nevertheless was kind enough to do both tasks for us because of time constraints. It is not usual for an abstractor to summarise the chapter as a textual unit; she therefore treated it as though it were an article in a journal. Her analysis is

as follows:

Manual Abstract

This heavily illustrated chapter is a dramatic description of the Earth's more spectacular volcanic phenomena, both on land and under water. The author explains the burning lava flows of Iceland, the colossal volcanic explosions of Krakatoa (Krakatau) and Mount St. Helen's, as well as the underwater hot vents and hot springs on land. The author suggests briefly that such phenomena were probably involved in the origins of life on earth; he also emphasizes how fast flora and fauna colonise the volcanic debris of the devastated landscape.
Indexing Terms

Major terms: Volcanoes
Lava
Minor terms: Eruptions, volcanic
Geysers Krakatoa/Krakatau
Mount St. Helen's
Compound terms: Volcanoes - Flora and Fauna - Origins and causes
Iceland - volcanoes

Reader's remarks
The passage is self contained, easy to read, with many fascinating facts about volcanic activity. The illustrations add significantly to its interest. Most intelligent readers would be encouraged to find out more on the subject, and to continue reading the book. No
external information is necessary to understand the chapter.

Whilst there is a degree of match between the words in the abstracts produced automatically and manually, the differences that exist are fairly clear. An obvious one is that the automated set of clusters consists exclusively of words in the text, whereas the manual abstractor does not feel constrained to express herself solely in the words of the original text. For instance, she uses the terms 'flora' and 'fauna' in summarising the 'Living Planet' chapter, although only the term 'flora' appears, once, in the text.
The abstractor's approach is to have a particular readership in mind for each abstract. She assumed in the case of our project books that these readers might read her abstract in a public library whilst in search of a book on the particular topic. She sees it as her job to interest the potential reader in the book. For this reason, she includes evaluative comment in the abstract, about the intrinsic level of interest generated by the topic, about the way the author expresses him/herself, and even about the layout of the text. For instance, she comments on the degree of illustration in 'The Living Planet'. The automated system may also capture some metalinguistc comment, but since its criterion for entering words into a cluster is statistical, the metalinguistic words will have had to have occurred significantly, and be rooted in the text itself, to be picked up. They will therefore refer more to the organisation of the text, words such as 'tables',
or 'graphs', or 'chapter', rather than to its interest value. Identification of topic is a problem both for the machine and the human. Pilot study done with MA students at Birmingham revealed that each one saw the main concerns of the 'Living Planet' chapter somewhat differently. Or, to be more precise, the wording they used to express what they saw as its conceptual concerns differed. The topic of topic complexity is an important one, but cannot be gone into in detail

here. Suffice it to say that the 'Living Planet' chapter is clearly about several things, but two main aspects or stages of the topic flow are volcanic eruption and the subsequent regeneration of natural life. The manual abstract reflects this, and the automatic clusters do too, but
do not give it as much prominence, because it is expressed in more varied vocabulary. However, it has been picked up in the clusters:

krakatau reclaimed century
mallee fowl incubation

Another difference between the two abstracts is a result of the human abstractor's ability to ignore or weight certain sections of the
original text. On the one hand, the human tends to ignore examples
given in the text, concentrating instead on the main arguments; on the other, she uses any metatextual or typographical information, such as sub-headings, as a guide to focus and weighting of the abstract. It
would be possible for us to modify the automatic system to take some
of this into account.
Comparing the index with the automatic cluster, it is clear that there are also differences here. In the selection of key terms, the human abstractor has selected those that she thought would occur in the
minds of potential readers, even if they did not necessarily occur in
the original text.
It is also customary for a professional indexer to index on nouns, and furthermore on plural nouns, such as 'Geysers' or 'Volcanoes'. Indexing on the singular form of a noun would only be done where it was abstract
or uncountable, i.e. where it didn't really have a plural form. Whilst clusters can be created where all inflections of a words are
counted together, or 'lemmatised', and the clusters made to present
just the base (or other preferred) form, the form chosen for the
cluster must have occurred in the original text.
The automatic' clusters are fairly satisfactory in that they reflect
the changing aboutness of the chapter, and have done this purely by
'sticking to the text'. They are less easily readable than the manual abstract, but they are potentially more informative than an index made
up of individual words.

Conclusion

Nobody can know for sure what the future will hold for textual information, but some trends are clear. Some texts will continue to exist in at least two mediums, particularly where people wish to read them for pleasure, or need to carry out close study of them, or observe some feature of their original layout. Many of these will also exist electronically, where people need to receive, extract, retrieve and transmit information quickly. In addition, reflecting the two types of reader need, some texts that have hitherto been available only in hardcopy, such as old manuscripts and facsimiles, will be increasingly converted to electronic form; while, on the other hand, texts that are only needed for their information content, such as operating manuals, will only have an electronic existence.
Abstractors and indexers may find themselves taking the original wording of the text more into account as the focus moves towards the electronic medium and away from hardcopy. There will be software of the kind described in this paper available to influence them in this. Automatically-generated products, whilst being fast and excellent for some purposes, are not yet all readable or reader-friendly. This is partly because the computer can only represent the writer's model of text, whereas the human agent, as abstractor or indexer, adopts

the reader's perspective. I think that the kind of software described in this paper, in addition to being used to present finished products to the user, will serve a very useful function as an intermediary in the information chain. For example, the automatic abridgements could be used to find other relevant texts in databases; the clusters as a feedback stage in a keyword search system, or as a first-stage indexer. Abstracts of the abridged variety may be first-drafted by computer, then smoothed into a more interpretable shape manually. This would suggest a change in the role of the human agent, from text creator to text editor.

Bibliography

Attenborough, D (1984): The Living Planet, Collins/BBC, London.
Blackwell, S A (forthcoming): 'From Dirty Data to Clean Language', in Proceedings of 13th ICAME Conference, eds. De Haan, P; Oostdijk, N and J Aarts, Rodopi, Amsterdam.
Collier, A J (forthcoming): 'Issues of Large-scale Collocation Analysis', in Proceedings of 13th ICAME Conference, eds. De Haan, P; Oostdijk, Nand J Aarts, Rodopi, Amsterdam.
Hoey, M P H (1991): Patterns of Lexis in Text, O.U.P., Oxford. Phillips, M (1985) 'Aspects of Text Structure: an investigation of the lexical organisation of text', North Holland, Amsterdam.
Renouf, A J (forthcoming): 'A Word in Time: first findings from the investigation of dynamic text', in Proceedings of 13th ICAME Conference, eds. De Haan, P; Oostdijk, Nand J Aarts, Rodopi, Amsterdam.

Acknowledgements

We acknowledge with thanks the support of the Department of Trade and Industry, the Science and Engineering Research Council, BRS Software and HarperCollins in the AVIATOR Project; and of British Telecom in some of the work relating to automated textual abridgement.

Appendix

Original Text of Automatic Abridgement Bullish Lamont offers no early rate cuts By Peter Riddell and Anatole Kaletsky 31 December 1992

BRITAIN'S economy will do much better next year than in 1992, but there will be no further reductions in interest rates unless growth falls below the Treasury's expectations, according to the Chancellor of the Exchequer, Norman Lamont, in an exclusive new year interview with The Times. He is bullish about the economic prospects and unrepentant about the government's performance in the past year.

Mr Lamont said: "Recent evidence has been encouraging. We have had very good car sales in December and reports of buoyant sales in the shops. Surveys of business confidence have improved. There is every reason to believe that 1993 will be much better than 1992. I would not be surprised if trends in the British economy were better than in some of our European competitors." Mr Lamont's remarks may, however, disappoint the City, where many investors have been hoping for a further cut in interest rates early in the new year. The Chancellor said that interest rate reductions would be considered only "if monetary demand was manifestly too low".

Asked whether he would expect to change interestrates if the economy 1 per performed in line with the Treasury's forecast ofcent growth, the Chancellor replied with an emphatic "no".

He repeatedly expressed confidence that he had done enough in his Autumn Statement to ensure that his forecasts of economic recovery would be fulfilled. Mr Lamont said that Autumn Statement measures for industry and housing, and the big cuts in interest rates and the devaluation of sterling since Black Wednesday had "created the right conditions for confidence and growth". Monetary policy had already been relaxed "very substantially" through the interest-rate cut and sterling's devaluation,

Mr Lamont estimated that as much as two-thirds of the impact of the recent three-point reduction in interest rates was "still in the pipeline" and added that the "very warm welcome" given by industry to his Autumn Statement measures meant that there was "every chance that they will succeed". The combination of monetary relaxation and carefully directed fiscal measures had created a climate of confidence.

Mr Lamont was unrepentant about sterling's withdrawal from the European exchange-rate mechanism. The ERM had brought "enormous benefits" to Europe and had helped Britain to defeat inflation during its membership, he said. However, if other countries now chose to tie their currencies even more closely in narrower ERM margins, the implications for Britain would be limited, he said.

Mr Lamont was unperturbed by the size of Britain's current account deficit, despite concern in the business community that the balance of payments gap will be a constraint on economic growth. "I'm obviously not indifferent to the current account, but I do not regard it as my major problem, " he said.

The one economic problem that did seem to worry Mr Lamont was the high level of the public sector borrowing requirement.