What the linguist has to say to the information scientist.

Antoinette Renonf, *Research and Development Unit for English Studies, School of English, University of Birmingham, Westmere,* 50 *Edgbaston Park Road, Birmingham B15* 2RX

REUNITED

John Major, the dog who went missing for three weeks has been reunited with owner Eric Templeton from Kenelm Road, Oldbury, thanks to Solihull-based Petsearch.
(Birmingham Evening Mail, Dec. 1992).

I. INTRODUCTION

The term **linguist** is broad in application, but in this paper it will be used to refer to the **corpus linguist**. **Corpus linguistics** is a particular branch of linguistic study that is concerned with the analysis of patterns of words and word occurrences in a large, computer-held store of textual data. Such a store is referred to as a corpus. Corpus linguistic work continues within the Research and Development Unit for English Studies, where our current focus is on monitoring hundreds of millions of words of text, using computational and statistically-based techniques as part of a large, government-funded, collaborative project with industry.

Having spent over twelve years in the field, I see the academic purpose of the corpus linguist as being to produce better descriptions of the language, that will contribute to a greater general understanding, but at the same time being alert to developments and changing needs in associated fields, where linguistic knowledge could help to solve problems. In this paper, I should like to report on some of my obsetVations about the nature of language that are of relevance to the process of automatic text retrieval, with particular reference to the problem of keyword search. I shall address the following areas:

- Multicontextuality: the reliability or otherwise of words as search terms;
- Word Boundary: the word as the basic unit in text;
- Thesaurus: the nature of semantic relations in text;
- Topic: the relationship between word and topic in text.

2. MULTICONTEXTUAUTY, THE REUABIUTY OR OTHERWISE OF WORDS AS SEARCH TERMS

The success of textual search methods based on keywords depends on the ability of a word to retrieve a particular text. However, observation shows that relatively few words in English belong, or refer exclusively, to a single context. The majority are therefore unreliable, to varying degrees, as search terms. The text in Fig. 1 will serve to illustrate this point:

FT 07 Sep 90 Burmah profits edge up by 3% midway:
By Richard Gourlay

Burmah Castrol, the specialist lubricants and chemicals group, reported a 3 per cent rise in pre-tax interim profits to Pounds 79.2m yesterday after what it described as its toughest six month period this decade.
Group sales rose 4 per cent to Pounds 831. 7m and the company said gross margins were maintained. Earnings per share increased 3 per cent to 24.7p and the company raised its interim dividend by 0.5p to 8.5p.
However, the City continued to believe the company would be hit by sterling's strength, and the slowing of business activity in its main markets.
Burmah's shares closed down 9p at 506p.
Mr Lawrence Urquhart, the chairman, warned that market conditions for a number of its sectors would remain tough throughout the year and currency gains in the first half would turn to losses if sterling remained at current levels.
Mr Jonathan Fry, Burmah Castrol's managing director, said the Gulf oil crisis was beginning to force up prices of base oils, the raw material for the main lubrication division which supplied 66 per cent of trading profits in the first half.
As medium-term supply contracts began to expire during the autumn the effect of higher base oil prices would become more acute. Mr Fry said he was confident Castrol would be able to pass on the higher prices to its customers as it had always succeeded in doing during earlier oil price crises although in some countries this might prove more difficult, notably in West Germany.
The sale of its 29.7 per cent stake in Premier oil in August for Pounds 138m has reduced Burmah Castrol's gearing from 30 per cent at the interim stage in 1989 to less than 10 per cent. The loss of consolidated earnings from Premier would be more than made up by interest on the proceeds of the sale not used to pay down debt.
First-half trading profits in lubricants rose 10 per cent to Pounds 60, with Pounds 3.1m of that coming from currency gains on translation from currencies that strengthened in the company's main markets, notably Germany. The fuels business rose 23 per cent to Pounds 11.1 m partly through acquisition of properties in Australia, Chile and Sweden.
Chemicals trading profits fell 18 per cent to Pounds 7.5m, hit by the downturn in the US and UK construction and consumer markets while losses in the US coatings sector offset profit growth in Europe.
Burmah Castrol also announced yesterday it had sold two ultra large crude carriers to Concordia Maritime, part of the Stena group, for Dollars 47 *Am.* Stena is already operating the ships under a seven year charter.

FIG 1: Article *from* the *Financial* Times *of 7th September, 1990*

Fig. 1 contains 240 different words, most if not all of which could also retrieve text types and topics different from this particular one. This is because text is largely made up the following types of words:

2.1 Very common words

A good part of text consists of a few hundred words that are used again and again, for different purposes. Those that spring to mind are the grammatical items, like **the, of** and **up**, which contribute to the grammatical structure of a text, but also combine with each other and with common, more lexical, words to create the phraseology that contextualises the propositional content of a text. The common lexical words consist of nouns like **year** and **time**; verbs like **give, take, put** and **make**; adjectives like **big, white** and **good**. Phraseology made up of common words in Fig. 1 would include:

throughout the year
in the first half
at the same time
would turn to

Such words occupy about 60% of the text in Fig. 1. As common items, they are not useful as search terms, because they are by definition multicontextual.

2.2 Discourse organising words

Words like thing, kind, level and effect are also very common in text, where they are typically used to link the text together, almost like grammatical words. They are not usually topic-bearing in themselves, because they do not have a full or specific meaning: **this problem was soon resolved** requires the reader to look elsewhere in the text to discover what

precisely the word **problem** is referring to in the particular context. These words are frequently used to create frameworks to carry technical or topic words, as in the following examples from Fig. 1:

the specialist lubricants and chemicals **group**

its toughest six-month **period**
the fuels **business**
at the interim **stage**

It seems likely that some of these may tend to be found in some textual domains rather than others: for example, **character** is commonly used in biology, whereas the noun **characteristic** might be more common in other domains. However, the precise extent and nature of any correspondence between these words and text type is not yet known to the linguist.

2.3 Homonyms

Homonyms are words with more than one meaning, such as **rose** or **down**, where the two meanings are unrelated. That is, **rose** can mean **flower**, or **got up**. In the Burmah Castrol text, the following words are homonyms: **pounds, main, rose, base, down, page** and **acute**. Their ambiguity means that they are likely to retrieve text on at least two different, and probably unrelated, topics. It is rare that both meanings will occur in the same text; if they do, this would skew any software weightings based on keyword frequency.

2.4 Polysemes

Polysemous words are those which have one basic meaning when considered in the abstract, but which can take on an additional range of nuances from the contexts in which they occur. From the retrieval point of view, they are therefore likely to be found in many different kinds of text, and so not be useful as search terms. Fig. 2 presents a selection of different contexts containing the word thin, taken from a corpus of general texts. The word can be seen to gravitate towards different topic and technical areas.

Science and natural history: objects or creatures thin by nature

Our Sun emits X rays from its thin outer atmosphere.
The thin capillary tube in the thermometer
The lightest touch peeled off thin slivers of wood.
The top was covered with a thin layer of soil
The plate is made from thin aluminium sheet
The cheetah has a thin elongated body

Architecture, design, cookery: objects or substances made thin for a purpose

They were all made of very **thin** bricks
A rather **thin** building in an Italian Renaissance style

Her water-colours and a **thin** camel-hair brush
The **thin** silk of the dress
Warm the syrup until **thin** and runny.

Description of people who are thin in the sense of elegant or refined

He was a tall, **thin** man with a lean, ascetic face
A man of letters, a **thin**, blond, refined man in a bow-tie.
He stretched out a long **thin** elegant leg
delicately put together, with a **thin** face and large, warm eyes.

Writing, food, clothing: objects or substances that are undesirably thin

The volume was **thin** and a trifle soiled
The writing was **thin** and hurried.
The coffee was **thin** and tepid
Split shoes, and cement yards, **thin** coats and mealless days
The light entered through the **thin** cheap floating curtains

Description of people or features that are undesirably thin

They were **thin** and hungry, dressed in rags.
His nurse, a skeleton **thin**, faded girl
The **thin** children's voices
There was a **thin** cynical smile on his mouth.

Idiomatic Phrases

The joke had begun to wear **thin**, after a nightly airing.
They had disappeared into **thin** air.
The predictability argument is **thin**, but it is often raised.
The Arab Legion, under the **thin** disguise of being British troops
I knew that I was on **thin** ice.
You were so **thin** skinned.
That poor girl gets a pretty **thin** time of it.

FIG 2: *Examples of contexts associated with the word* thin

2.5 Semi-technical words

Some words have both a technical and a more general, lay meaning. This means that they may retrieve texts related to their technical meaning, but they may also retrieve quite different ones. Some words from Fig. 1 reveal this feature. The first, technical, context for each is taken from the text in Fig. 1:

company

the **company** said gross margins were maintained

I shall enjoy the **company** and conversation of young people

sale

The **sale** of its 29.7 per cent stake in Premier Oil

He had picked it up at bargain price in a **sale**

business

The fuels **business** rose 10 per cent to Pounds 60

I'd be obliged if you'd mind your own **business**

losses

Losses in the US coatings sector offset profit growth in Europe

Cut your **losses** and start again

It is a fact of the language that the split between technical and non-technical meaning often coincides with grammatical word-class. For example, the words **fuels, gains** and **profits** are all associated with a technical context as nouns, but as verbs they are more general in reference.

Retrieval software can exploit this knowledge to some extent if all words in the database have been assigned a grammatical code.

2.6 Words with *several technical* senses

A subset of polysemous words are those that have more than one technical sense; that is, they occur in more than one technical context. Some words taken from Fig. 1 illustrate this. The first context for each comes from Fig. 1; the others are extracted from the general Birmingham Corpus:

oil

the effect of higher base oil prices would become more acute oil,
free milk, free cod liver free schools, free medicine
rub in linseed oil to darken

group

Concordia Maritime, part of the Stena group
This pupil has been placed in the wrong group

charter

Stena is operating the ships under a seven year He charter
is travelling on a charter flight

lubrication

raw material for the main lubrication division to
Neglecting to blink prevents the necessary lubrication the eye

reduced

The sale of its stake has reduced Burmah Castrol's gearing
The side toes were When reduced to internal vestiges
the liquid has reduced by half, pour over the mackerel

2.7 Metaphor

Overlaying the above types of words that traditionally make up text, there is the textual phenomenon of metaphor. This is the process of transposing words that typically belong in one text type or discipline into another, for example to explain a technical concept by analogy with a familiar, everyday one. In technical writing, there is a tendency for terms from one technical area to be adopted gradually by another, partly motivated by the abovementioned need to explain. An example from Fig. 1 is the word **gearings,** which originated in the domain of engineering, as did the similar term **leverage. Mezzanine** is an instance of transfer to finance from architecture.

Metaphor is particularly prevalent in journalism, where the explanatory metaphor is routinely supplemented by metaphorical conceits, intended to divert and entertain readers of what journalists feel to be dreary or inaccessible material, and to score points with colleagues. Some metaphors are intrinsic to our society and language, such as the conventional **time is money or the election is a horse race,** while others are more ad hoc. Any kind can be problematic in text retrieval, because the actual words used to express them are unpredictable.

Fuels and **currency** are tWo example of words in Fig. 1 that commonly occur metaphorically in contexts not exclusively related to business and finance. The metaphor equating turbulent activity with fire underpins such phrases as **his sudden departure fuels rumour,** whereas metaphor characterising someone's behaviour as his way of paying for something leads to phrases such as **lying was his currency.**

2.8 Typographical ambiguity

There are some cases of ambiguity which come into being when typographical distinctions, such as upper and lower case marking and

punctuation, are ignored. Proper Names, like Fry and Times, in Fig. 1, are a case in point. Initialised abbreviations, such as U.S., in Fig. 1, are another. The occurrence of acronyms in text is increasing and their treatment is quite variable: in UK texts they are rarely punctuated, while in US texts punctuation is the norm. Some of them, such as ARMS, are confusable with relatively common bonafide words. Some recent acronyms found in UK newspaper data are shown in Fig. 3.

Sep-25: BAT Industries saw earlier gains halved with a rise of 7p to 645p
Sep-29: Top men at BET to go; In Today's Other Papers
Sep-24: or by a member of the CAB or by a member of a law centre. Sep-23: ministers are to approve a CAP reform to jump start the Gatt talks
Sep-07: 10,000 in the (90-day) investment account compounds up to a CAR of 7.89
Sep-14: Based on the French appellation controlee system, Italy's superior DOe
Sep-29: "Without the seal it isn't real" is the slogan FACT is using to
Sep-14: Orton, managing director of HIT Communications, which funded the series
Sep-04: NUT seeks collective bargaining; TUC conference
Sep-28: have been misused even in the absence of the PIN number. Sep-Ol: players from their own raw courage, the WHO points out that the ruling
Sep-Ol: "Given the WHO report, and our own medical association findings

FIG 3: *Acronyms drawn from Times newspaper data of September 1991*

2.9 Preferred contexts

The eight categories of words outlined above, which all in one way or another move between textual domains, are indeed a problem for text retrieval. Language does, however, have certain redeeming features, that could make it easier to deal with if understood fully. One is that, in the totality of text, a word tends to have one predominant sense and one or more much rarer ones. The word fell can mean hill, but it is far more likely to carry the semantics of falling, except in a specialised, and fairly predictable, text domain. Similar words to fell in the Fig. 1 text are: rose, down, page, current, stage, stake and offset. The database searcher can probably guess that rose refers more often to a verb than to the flower, down to direction and not plumage, page to a sheet of paper rather than a servant, and so when these words are safe to select as keywords.

Related to this general semantic bias, there is the contextual one. A restricted textual domain, such as finance or medicine, is likely to feature a word in one sense only. For example, the word head is very common and has several senses, but in the city pages of the Independent Newspaper, its primary sense has been found to be director. It is therefore potentially useful as a search item in the business section of a newspaper, whereas it would not be in the newspaper as a whole. Generally speaking, the criteria for success for an individual keyword will be rarity and technicality: the rarer or more technical a word is in the language, the more likely it is to have a unique relationship with a textual domain. In addition, if several keywords are submitted simultaneously, they will of course serve to disambiguate each other, and achieve greater success in so doing.

3. WORD BOUNDARY: THE WORD AS THE BASIC UNIT IN TEXT

The keyword approach to text retrieval is based on the traditional metaphor of a word being a boat that carries an individual and discrete meaning. This seems intuitively true, since we can ask of ourselves what meaning an individual word has and usually find an answer. However, it is not adequate to account for the status and behaviour of words when they are combined to form text; when we look at text, it is clear that words are coselective, and dependent on each other to realise a relevant sense in the text. A second, and apparently conflicting metaphor, is also required, on the lines of a word being a plant, linked to roots and suckers. This metaphor is appropriate because it reflects several characteristics of words in their interaction with their neighbours in text. One of these is their habit of combining with each other to create a unit of shared meaning, or a phrase.

The consequence for text retrieval is that the selection of a single key word may be inadequate.

3.1 Multi-word items

3.1.1 Idioms

Idiomatic phrases are a prime example of linked words forming a unit from within which it would be nonsense to select a single word. This is demonstrated by the phrases relating to the word thin in Fig. 2, including on thin ice and into thin air.

3.1.2 Combinations of common *words*

As said earlier, the commonest words of the language are common because

they combine again and again, especially with each other, to form the essence of our everyday phraseology. Phrasal verbs are one case of two or more words creating a single unit of meaning. In the verb formations **put off, put out, put one over on** and **put up with,** the verb **put** does not stand alone, in the way that the fully lexical equivalents of its verb phrases do: compare it with **postpone, inconvenience, cheat** and **tolerate. Put** and other common words are, for this reason as well as for their previously mentioned tendency to polysemy, generally not good search terms.

3.1.3 Technical terms

Technical terms very often consist of long, descriptive noun phrases; the following come from a text on thermodynamics:

air separation plant
steam power plant
electric generator
vapor-compression refrigeration-cycle
thermodynamic equilibrium

and these from the short text in Fig. 1:

pre-tax profits
gross margins
earnings per share
interim dividend
managing director
oil crisis
raw material
trading profits

Previously, these multi-word terms proliferated mainly in technical and scientific texts, where around 70% of the text could be made up of noun phrases, but this feature now appears to be on the increase in all types of text. A recent spoken example, not necessarily typical, was the humorous formulation **a catalyst of change in our** society, used in place of the alternative **accountant.** The move towards nouns is accompanied by a move away from verbs: many nouns actually do the job of a verb in expressing the concept of process or action. **Privatisation** has, for example, subsumed the verb sell.

3.1.4 Proper names

Proper names include the names of people, places, objects and companies, and they are frequently multi-word items. It is not sufficient to search on Robert, or Burmah, or Consolidated, since these individual words are not unique identifiers of meaning or reference. (Though even word pairs, like John Major, cannot guarantee the retrieval of a relevant text, as the opening quote of this paper demonstrates.).

3.1.5 Hyphenated word strings

The general trend towards creating multi-word units reflects a desire to compress a lot of information into few words and a little grammar. Many hyphenated strings reduce a clause to a word. Business text regularly spawns hyphenated adjectives, for example:

export-led
product-driven
cash-starved
asset-backed
debt-ridden

Technical terms and proper names are also frequently hyphenated. There is a trend in journalism towards ever-longer strings of hyphenated words, such as those in Fig. 4.

pre-tax profits
short-term credit
year-on-year deficit
one-for-one share exchange
a **one-child.per-family** policy
state-of-the-art printing presses
a **once-in-a-lifetime** payoff
this **sport-is-good-for-business** concept
breakthroughs in **telecoms-allied-to-the-computer**
a **take-it-or-leave-it** package
the **pile-it-high-sell-it-cheap** way
her **screw-them-before-they-screw-you philosophy**
the **remote-control-telly-change-over-panel_thingy Clapped-out-old-windbag-of-a-backbench-constituency_MP**

FIG. 4: *Hyphenated word-strings from* the *Times newspaper for December 1991*

The longer strings tend to be more idiosyncratic, and are therefore less significant in text search. The more conventional strings are fairly short, containing just two or three words, and can often be important technical terms; these are the type that need to be accommodated in a keyword search facility.

3.3 Separable phraseology

The linkage or dependency between words is not only adjacent. The metaphor of the plant also accommodates the phenomenon of more long distance networks of connections. Returning to the phrasal verbs, put and off, in the sense of **postpone**, still create a shared meaning when they are separated: He **put** the whole question **off** until later. Multi-word technical terms can also be split on occasion. Take the example from Fig 1 *of:* **pre-tax** interim **profits**.

4. THESAURUS, THE NATURE OF SEMANTIC RELATIONS IN TEXT

The lines of context provided in Fig. 2 for the word thin raise the question of thesaural relations in text retrieval. There are several different kinds of thesaural relations, but we shall focus on a central one here: synonymy, which is the relation of similarity of meaning that exists between two words. The thesaurus is theoretically a useful tool, since it offers the opportunity for automatic expanded search, whereby the machine will look to a store of associated words once it has an indication that a first word has proved useful. The nature of language is such, however, that it is very difficult to find or create thesauri that are actually useful. One important reason for this is that there can be said to be three main different kinds of synonymy abstract synonymy, contextualised synonymy and instantial synonymy defined as follows:

Abstract Synonymy

The abstract level of synonymy is that which is accessible to us inside our heads. When asked what a synonym of thin would be, we would say skinny, perhaps, or slender, providing a single word synonym for what we perceive to be the meaning of thin in the abstract. This is perfectly normal and acceptable, and a system which we operate with every day. It is accounted for by the boat model of language.

Contextualised Synonymy

The contextualised level of synonymy is primarily accessible to us through observation of text. In context, the synonyms of thin become more particular: unconvincing, in the context of argument, reedy in the context of children's voices, and underfed, when used
to refer to poor or neglected people. This follows the plant model of language.

Instantial Synonymy

The instantial level of synonymy occurs in individual texts and only has momentary or passing validity. An example is where, in a particular text or context, such as where the house pet is a cat called Sherlock: two otherwise unrelated words - **cat** and **Sherlock** _ become synonymous (or co-referential). This pairing clearly does not have sufficient generalisability to warrant its being included in a thesaurus. In Fig. 1, the writer uses the phrase **pay down** of debts, which is presumably synonymous with **reduce** or **payoff**; it is questionable whether the equivalence is worth recording. A more familiar problem for database users is the kind of ephemeral synonymy represented by the temporary synonyms **President** and **Clinton**. This word pair is synonymous (or co-referential) only for a limited period, but needs to be recognised for that duration.

For text retrieval purposes, the second of these synonym types is generally speaking the most useful, the third useful if thesauri can be updated regularly; but most users will not be aware of these various distinctions as a matter of course.

4.1 Scope of reference

Even established synonyms are not co-extensive in reference. **Dropped** and fell both occur in financial contexts, but **dropped** also occurs in rugby contexts: **dropped kick**, and so on.

4.2 Synonymous phrases

Synonyms are generally assumed to be pairs of individual words, that occupy the same grammatical word-class, and belong to the grammatical word-class of nouns. These are wrong assumptions. Synonymy exists between words other than nouns, and between words of different grammatical classes. The meaning relation is also frequently sustained above word level: that is to say, that phrases can share meaning with an individual word or another phrase. Fig. 5 demonstrates both these features.

Sentence no.			
1	o *FT* 07 Sep 90 Burmah profits BURMAH Castrol reported a 3%	**edge up** **rise rose**	by 3% midway (453) in pre-tax interim profit 4 per cent to Pounds 831.7m
2	Group sales		
3	the company said margins were Earnings per share and the company	**maintained.** **increased** **raised**	3 per cent to 24. 7p its interim dividend by
4	The company would be and the	**hit slowing down**	sterling's strength, of business activity 9p at 506p.
5	Burmah's shares closed		

6	market conditions would	**remain tough**	throughout the year
	and currency	**gains**	in the first half
	would	**turn**	to
		losses	
	if sterling	**remained**	at current levels.
7	The crisis was beginning to	**force up**	prices of base oils
	which supplied 66% of ttading	**profits**	in the first half
8	As supply conttacts began to	**expire**	during the autumn
	the **effect** of	**higher**	base oil prices would
		become acute.	
9	Casttol would pass on the	**higher**	prices to its customers
10	The sale has	**reduced**	Burmah Castrol' s ~g
11	The	**loss of**	consolidated earnings
	from Premier would be more than	**made up**	by interest on the proceeds
12	trading profits in lubricants	**rose**	10% to Pounds 60.3m,
	currencies that	**strengthened**	in the main markets
13	The fuels business	**rose**	23% to Pounds 11.1 m
14	Chemicals trading profits	**fell**	18% to Pounds 7.5m,
		hit	by
	the	**downturn**	in the consumer markets
	while	**losses**	in the US coatings sector
		offset	
		profit	
		growth	in Europe.

FIG. 5: *Synonymous words and phrases expressing the notion of* **change**

The tendency to use multi-word synonyms is particularly apparent in the mention of people and companies in text, where it is chiefly a matter of coreference. In referring to a person, it is normal to select from a range of three co-referential phrases if they are not famous:

Mr Peter Smith, Mr Smith, Peter Smith

four or so, if they hold a position of importance or relevance (example from Fig. I):

Mr Jonathan Fry, Jonathan Fry, Mr Fry, Burmah Castrol's managing director

and for the number of co-referential terms to proliferate if they begin to attract media attention, as shown in Fig. 6:

A: in referring to the person
Inscrutable Man
John Birt
the new Director-General
Director General
Birt
the new DG
Birt the man at the top of the 24,OOO-strong monolith of the BBC this young outsider to the BBC
John

B: in characterising the person
Stalin
Cromwell
Robespierre
journalism's Jeremiah
a hidden God
a Martian
one of us
a Sphinx without a riddle
Lenin
a Trappist monk
a missionary
an evangelist
a hidden God who broods in his office
the Bruiser
a gauche provincial

FIG. 6: *Article* on *the appointment of John* Birt *as Director-General of the BBC, Sunday Observer, 03 January 1993*

Awareness of these multi-word co-references is vital to successful text retrieval, but they pose obvious problems both to user and to sofrware developers.

5. TOPIC: THE RELATIONSHIP BETWEEN WORD AND TOPIC IN TEXT

5.1 Topic complexity

The user of a textual database has to make a connection in his or her mind between the kind of text that he or she wants to retrieve and the words that would seem to identify it. This is a difficult thing to do, because topic, the **aboutness** of a text, is a complex phenomenon. To take an example, Chapter 1 of a book by David Attenborough entitled The Living Planer1 can be seen to be **about** something on several levels. The following is a simple representation:

General topic area:	Natural History
Specific topic focus:	1. Volcanoes 2. The natural world
Point being made:	Volcanoes erupt in various ways, and devastate areas, but plant and animal life soon regenerates itself
Writer's evaluation:	This is a wondrous thing
Larger significance:	All part of life's rich pattern

5.2 Topic flow

A text is not always uniformly about one topic. The type of text it is, whether argumentative or narrative, and so on, will affect the number of aspects of aboutness contained in it. The Living Planet text mentioned above is expository/narrative: that is, it describes and narrates, and it has two main foci. The first is the process of volcanic eruption, and the second is the regeneration of living organisms in the aftermath. The flow from one topic to another is characteristic of many text types, and means that some parts of retrieved texts will be relevant, while others are not. In journalism, the treatment of topic can often be quite cavalier, and a journalist will patch together several tenuously related items of information under one heading. In the text in Fig. 1, the last two sentences are tacked on, and only linked by implication to the rest. A degree of sophistication is required by the software to accommodate these phenomena, and by the user to deal with it.

5.3 Relationship between words and topic

What might be considered to be the identifying topic words of a text by a database user may not occur at all. The text in Fig. 1 is about the financial performance of Burmah Castrol, but the words performance and results are not used. It may help to adopt an interactive model of text to explain why this is. The writer could be said to be creating text in answer to an imagined question, this in the case of Fig. 1 being: What has been the financial performance of Burmah Castrol? The response is expressed in a series of verbs and phrases to do with rising, falling, strengthening, etc., on the one hand; and with warning, believing, announcing and other types of reporting and predicting, on the other. In many texts, topic words do occur, but sometimes only once. In the Maastricht treaty, a major theme is subsidiarity, yet this appears just once. Topics are obviously also expressed and developed through other words.

5.3.1 Reiteration through re-phrasing

Part of the reason for this is that stylistic rules place a restriction on the number of times a particular word may be repeated, even if it expresses a core idea in a text. The writer might substitute a synonym or a paraphrase: for instance, **devolution of power** for **subsidiarity**. These are features that would not automatically be taken into account by software that weights a text for relevance on the basis of the number of mentions of a keyword.

5.3.2 Grammatical reiteration

A topic theme can also be carried by grammatical words, like pronouns and pro-forms. The text on dyslexia in Fig. 7 illustrates the use of such words: .

> Dyslexics can be perfectly intelligent but need expert attention. Reva Klein explains what can be done.
>
> Paul Levy is a bright 16-year-old, sociable and generally well adjusted. But ask him to write in longhand, and it becomes obvious that something is wrong. His uniformly misspelt and ungrammatical writing is incongruous with his sophisticated verbal expression.
>
> Paul's problem is dyslexia, or 'specific learning difficulty', and he is in good company. Once denigrated as a peculiar condition of children of middle-class parents, it is now known to affect one in 10 people of every background. It appears to be hereditary and can range from a moderate, short-lived problem to a severe one that lasts a life-time. One in 25 people has a problem serious enough to require specialist tuition.
>
> Dyslexia is defined by Harry Chasty, Britain's leading expert on the condition, as 'a congenital organising disability that inhibits the development of a child's literary skills - particularly reading. In its effects, this condition can range from slight reading difficulty to complete illiteracy.'
>
> Paul was nine when his difficulty was diagnosed. Before that, he was thought to have emotional problems. Then someone mentioned dyslexia, he was tested, the condition was confirmed and he was sent to a specialist school.

FIG. 7: *Extract from* Article on *Dyslexia, Independent*, 13 *February, 1992*

In the text above, proforms include the pronouns he, his, him, it, its and the dummy noun one, all words that are not recognisable as topic-bearing, but which refer to and stand for the topic words, reiterating the central theme.

5.3.3 Reiteration through discourse organising words

Discourse organising words have been mentioned earlier. They operate as linkers, or cohesive elements, in text, in much the same way as pronouns do, although they do not look grammatical. In the text on dyslexia, the recurrent words of this kind are **problem(s), condition** and **difficulty**. They all stand in for the word **dyslexia**.

Like pronouns, these discourse-organisers are not exclusively related to a particular topic, but can occur in various domains. They are therefore not particularly useful as search words in themselves. However, they should really be taken into account in text retrieval systems that weight key words on the basis of frequency.

6. CONCLUSION

I have tried to give some insights into the nature of language that have some bearing on text retrieval. The features identified have largely been those that cause problems for both the database user and for the software writer. However, in the area of corpus linguistics, progress is being made towards finding solutions2. Some obvious ones have been implied at each stage of this paper, which may be summarised in the following terms. With reference to the selection of useful keywords, it is clear that rare words, and words with a primary meaning associated with the technical area required, are good candidates. Secondly, the submission of several keywords (e.g. Gulf, oil and price) is likely to achieve greater precision than that of a single term (e.g. oil). Thirdly, in relation to word boundary, an effective search term may need to consist of several words. Finally, with reference to the use of the thesaurus, an awareness is required of what constitutes a synonym as realised in text. Database users and software developers have a degree of intuitive awareness of such features of the language, and already exploit them to differing degrees. Dialogue between the fields of information retrieval and corpus linguistics would, I feel, lead to benefits on a larger scale.

REFERENCES

1. ATTENBOROUGH, D. The Living Planet. London: Collins, 1984.
2. RENOUF, *A.l.* Sticking to the text: a corpus linguist's view oflanguage. Aslib Proceedings, 45(5), 1993, 131-136.