The ACRONYM Project: Discovering the textual thesaurus

Antoinette Renouf

University of Liverpool

1. Introduction

The ACRONYM Project began in April 1994, funded by the British government and industry for three years. It is a collaborative project, its partners being the University of Liverpool, FT Info (the archiving branch of the *Financial Times)*, and the Press Association (a news agency supplying news to the regional press). The project has as its objective the development of an automated system for identifying thesaural items in electronic text. ACRONYM is an acronym, standing for 'The Automatic Collocational Retrieval Of Nyms'.

2. Definition of a 'Nym'

'Nyms' are pairs of words (or lexical items) which occur in text in similar collocational environments, and which we have found to be in some semantic or other non-random relationship to each other. We call these word pairs nyms because they do not map exactly onto conventional semantic categories. The nyms which pair with the word *luxury,* for example, range from being synonym-like *(five-star, s-class)* to antonym-like *(no-frills, three-wheeled)* to hyponym-like *(jaguar, lexus);* to more unconventional pairings. Our use of the term 'nym' also reflects the fact that, for some purposes, it does not matter to us what the precise nature of the relationship is.

2.1 Applications of the system

There are two applications for the nyms, once discovered. The first is in expanded search routines for large textual databases. Nyms will constitute potentially useful alternative search terms for text retrieval systems. It does not matter what the category of 'nymness' is, so much as that nym pairs will, almost by definition, retrieve similar textual passages.

In automated text retrieval software a thesaural facility is often an empty framework, into which users are invited to insert their own synonyms and alternative references for a given word. For instance, they might put the abbreviation

Synchronic corpus linguistics. Papers from the sixteenth International Conference on English Language Research on Computerized Corpora. Toronto 1995. Amsterdam: Rodopi, 1996.

for their company or other companies or organisations, so that every time a user submits the full name as a search term, the thesaural facility will be activated and a 'secret' search will additionally be undertaken for the abbreviated form.

Attempts to attach existing thesauri have proved unsatisfactory. A general thesaurus is likely to offer alternatives which do not match the particular text closely enough. The entry for the word *luxury* in the *Col/ins thesaurus* (HarperCollins, 1995) is fairly typical. The synonyms which it offers are:

> *Luxury* 1. affluence, hedonism, opulence, richness, splendour, sumptuousness, voluptuousness 2. bliss, comfort, delight, enjoyment, gratification, indulgence, pleasure, satisfaction, wellbeing 3. extra, extravagance, frill, indulgence, nonessential, treat.

These alternative terms are not those used in journalism, particularly in business or financial text. They are, on the whole, too literary. *Frill* is alone in echoing journalistic style, and does appear in our databases, though only in the form *no-frills.* Another problem is that traditional thesauri are grammatically-oriented. That is, they assign a word class to an item, and only offer alternatives within the word class. In fact, sense equivalence is often conveyed across word class in text. A related problem, illustrated by the word *luxury*, is that nouns, rather than adjectives, are frequently used to modify other nouns. So *luxury*, rather than *luxurious*, is the conventional modifier in journalistic text, and its thesaural alternatives thus include adjectives.

Some specialised thesauri have been created manually, but this process is lengthy and expensive, and the results are patchy and impressionistic.

Existing thesauri are also not readily updatable, and so cannot cater for the continual change in the textual lexicon and thesaurus. Neologisms and new usages occur in text (Renouf, 1993a); coreference is the fastest and most predominant area of change as the text mirrors real world events, and in particular the changing identity of people in public roles.

It is envisaged that our system will improve the situation in several ways. Our thesaurus will be a representation of textual reality rather than of the mental lexicon, because it will be extracted directly from the particular database on which it is operating. The thesaural information will be specific, reflecting the particular domain and style of the database. Since it is automatically generated, this thesaurus will also be readily updatable.

The second application for the nyms, once identified, is in linguistic description. Nyms constitute elements of the textual thesaurus, which we hope to describe over the next few years.

2.2 Underlying hypothesis

The underlying hypothesis in ACRONYM is that two words with similar collocational profiles will share meaning, reference, or use. This notion has evolved out of the Firthian (1957) view that collocation is a type of meaning, which has been fundamental to our research approach for the last 15 years. It coincides with our experience that the automation of procedures relating to text handling can be achieved through direct contact with raw text, and gives us a means of accessing the semantics of text at surface level.

2.3 Pilot test of hypothesis

Tests in 1992, in which simple collocational profiles were extracted from *The Times* newspaper text, supported the hypothesis to a certain degree. Our observations were as shown in Table 1.

Table I shows that the words *dropped* and *fell,* which we intuitively accept as being semantically related, as synonyms, do indeed share many col locates. They seem to overlap in their reference to financial matters. However, as with all synonyms, they have areas of semantic (or referential) dissimilarity, and hence unmatched collocates also. For instance, *dropped* seems much more closely associated with bombs, temperatures, the sports of rugby and golf, whereas *fell* is linked with financial market prices, currency values, and the sport of cricket.

Table 1: Collocational profiles for *dropped* and *fell* in the Times 1991

dropped and *fell*		*dropped* only		*fell* only	
bombs	36,972	goal	122.474	turnover	24.363
sharply	10,704	goals	67.823	wickets	22.652
points	8.907	penalty	60.579	foul	21.464
dramatically	7,633	penalties	40.305	love	15.296
fifth	6.489	shots	29,033	victim	14.390
below	6.225	tries	20,930	rain	13.878
per	5.740	conversion	18.045	nikkei	13.724
ball	5.720	catches	14.560	operating	12,644
cent	5.456	conversions	14.485	end-september	12,144
catch	5.316	pears	14.322	snow	11,974
cents	5.233	shot	14.147	dm	11.616
consumption	4.959	try	13.904	fence	11.570
kerry	4.361	mullen	13.538	dow	11.273
quarter	3.859	thorburn	13.459	imports	10.916

(conL)

Table 1: Collocational profiles for *dropped* and *fell* (continued)

dropped and *fell*		*dropped* only		*fell* only	
figure	3.841	leaflets	11.331	apart	10.874
half	3.607	davies	10.095	trap	9.761
third	3.582	stephens	9.358	lowest	9.725
index	3.540	slip	9.130	sterling	9.514
profits	3.534	steele	8.880	dollar	9.335
prices	3.468	evans	8.528	million	9.308
immediately	3.443	bomb	8.295	asleep	9.143
output	3.355	scorers	8.182	gdp	8.980
floor	3.217	montlaur	7.927	curtain	8.462
averag e	3.205	planes	7.884	darkness	8.428
shares	3.112	pilot	7.778	swoop	7.781
friday	2.933	tons	7.748	ft-se	7.686
low	2.919	camberabero	7.579	pound	7.603
behind	2.889	hints	7.375	lIoyds	7.541
favour	2.888	neat	6.967	wicket	7.499
off	2.887	barnes	6.838	diluted	7.448
while	2.737	hobbs	6.490	assets	7.347
after	2.680	temperature	6.194	pounds	7.289
earnings	2.628	temperatures	6.147	barrel	7.282
second	2.492	contention	6.072	portfolio	7.254
seven	2.413	turner	6.000	ounce	7.229
short	2.391	zoing	5.965	manufacturing	7.173
levels	2.377	strokes	5.851	deaf	7.066
sales	2.359	eighth	5.828	volumes	6.968
net	2.349	newport	5.716	wayside	6.847
nearly	2.249	anchor	5.707	arrears	6.788
rate	2.235	schofield	5.649	asset	6.647
price	2.226	chalmers	5.551	ill	6.599
ft	2.189	richmond	5.536	brent	6.567
industrial	2.175	Ib	5.534	roof	6.567
month	2.160	olazabal	5.513	ears	6.516
billion	2.097	parry	5.480	category	6.394
pre-tax	2.051	pass	5.360	slightly	6.340
share	2.039	squad	5.309	runs	6.297
before	2.034	hodgkinson	5.305	steadily	6.141

Table 2 represents the results of a search for all words which have collocates in common with the word *head:* these results indicate that *head,* normally a heavily polysemous word and therefore not a good search term, in fact does attract several nyms which are synonymous with its sense of 'being in charge', such as *director, chairman, executive, chief* There do not appear to be any nyms relating to the 'visage' sense of *head.* The lack of ambiguity in the use of the word *head* is probably attributable to the nature of the textual domain, and is a fact of the language which works in our favour.

Table 2: Nyms for *head* from the *Times* 'City Page'

81	head	12	left	9	firm	7	part
33	former	12	research	9	j apanese	7	partner
32	director	12	senior	9	joined	6	ash worth
24	has	11	john	9	two	6	broker
22	chairman	11	new	9	desk	6	capel
20	securities	11	trading	8	european	6	carol
18	team	10	company	8	govett	6	david
17	corporate	10	division	8	jon	6	department
16	executive	10	equities	8	morgan	6	financial
15	chief	10	equity	8	sir	6	fund
15	group	10	international	8	years	6	grenfell
15	managing	10	says	7	arm	6	house
14	aged	9	after	7	cstb	6	june
13	management	9	bank	7	de	6	london
13	sales	9	benson	7	gilts	6	manager
13	uk	9	county	7	investment	6	run
12	analyst	9	diary	7	jarnes	6	trust
12	its	9	financees	7	leonard		

Table 3: Nyms for node *establish*

497	establish	107	development	98	require	91	creating
171	create	106	ec	98	peace	91	based
151	develop	105	retain	98	international	91	agreement
150	established	104	un	98	authorities	90	statutory
147	maintain	104	economic	98	apply	90	operate
146	ensure	103	law	97	planning	90	help
145	provide	103	government's	97	finplement	90	european
140	build	103	give	97	governments	89	preserve
135	establishing	102	security	96	seek	89	including
126	secure	102	providing	96	companies	&9	decide
126	protect	102	management	95	required	89	authority
125	introduce	102	encourage	95	legal	89	any
123	restore	102	community	94	setting	88	creation
122	proposed	102	co-operation	94	produce	87	review
122	improve	101	protection	94	monetary	87	necessary
117	achieve	100	proper	93	separate	87	meet
114	impose	100	keep	93	policy	87	information
III	promote	100	agreed	93	determine	87	financial
III	control	99	its	92	towards	87	defend
110	prevent	99	government	92	regional	86	raise
109	reform	99	funding	92	powers	86	pursue
109	reduce	99	extend	92	legislation	86	iraq
109	make	99	environmental	92	avoid	86	developing
109	existing	99	bring	91	standards	86	committee
108	support	99	allow	91	social	86	carry

Table 3 shows us an impressive array of nyms sharing collocates with the word *establish.* The results are interesting, since *establish* is a word whose role is to structure text rather than convey meaning and as such might not have had any conventional synonyms. Table 4 presents the nymic output for the word *rowing.* Here we see that there are no synonyms-in fact only the near synonyms *sculling* or *punting* could have been hoped for. However, we do find a different semantic group here: co-taxonyms, such as *athletics, football, golf, cricket, boxing,* and *cycling.* We also see words in looser lexical association with *rowing,* such as *Oxford, Cambridge, boat, beat, college, crews.*

Table 4: Nyms for node *rowing*

85	rowing	24	cycling	20	club	19	memorial
36	oxford	24	evans	20	crews	19	middlesex
32	athletics	24	keith	20	cup	19	min
32	men's	24	league	20	davis	18	australian
31	cambridge	23	kingston	20	division	18	badminton
31	mike	22	andrew	20	dr	18	bath
29	david	22	birmingham	20	event	18	brian
29	junior	22	colin	20	gary	18	bristol
28	football	22	correspondent	20	international	18	captain
27	britain's	22	elliott	19	brighton	18	champion
27	england's	22	ian	19	bryan	18	clark
27	golf	22	leicester	19	cardiff	18	commonwealth
27	hall	22	manchester	19	christopher	18	don
27	john	21	beat	19	class	18	england
26	cricket	21	college	19	coach	18	final
26	former	21	fourth	19	coventry	18	fletcher
26	jones	21	hope	19	craig	18	gloucester
25	aged	21	james	19	crew	18	grand
25	boxing	21	kent	19	edinburgh	18	having
25	championship	21	leading	19	hill	18	huge
25	championships	21	lewis	19	indoor	18	june
25	chris	21	michael	19	irish	18	leeds
25	george	20	allan	19	liverpool	18	lengths
24	alan	20	amateur	19	mark	18	lightweight
24	barry	20	british	19	martin	18	london
24	boat	20	chess				

2.4 Project objectives

The pilot study encouraged us to go ahead with the project proper. This has a series of more detailed objectives, as follows:

automatic identification of nyms;
automatic identification of multi-word nyms (a multi-word nym is a lexical unit consisting of more than one word; for example, *Prime Minister* or *John Major);*
a system to monitor nyms in dynamic text; application of the system to specialised text domains; a pilot study to apply the system to texts in French; design of a machine-aided system to characterise pairs of nyms.

2.5 Computing resources

The software is being developed on a Unix system, with 200 mgb of memory, 2 CPUs and a 20 gigabyte disk. A client-server model is used, whereby large processes sit on the system, so that indexes, for example, are permanently in memory.

2.6 Data resources

The textual data is supplied by our industrial partners. We have about 200 million words of the *Financial Times* and the *Independent* newspapers, from 1988-1994, 5 million words of *McCarthy Business Reports,* and satellite access to P A news data services of various kinds. All data types are still being accumulated.

In the remainder of this paper, we shall outline the findings on the nature of the nym and of the thesaurus, which are emerging from the first two stages of the new prototype system.

3. Stage 1: Automatic identification of nyms

The procedure for identifying nyms is as follows. Each word type in a given corpus of text is taken as a focus, or 'node', and a cumulative collocational profile is built of all the words which occur within a span of, say, four places to its left and right. (In fact, some word types and collocates are ignored, or treated as 'stopwords'; these are very common, chiefly grammatical, items thought to be less useful as nyms by virtue of their tendency to collocate frequently with most other words.) The collocational profiles are stored in a 'Collocational Database'. The database records information of the following kinds:

length of span;
positioning in span;
ordering in span;
lemmatisation;
punctuation;
typography;
statistics of combinatorial possibility.

The profiles are compared in order to find nodes with collocational patternings which are similar to the patternings of a target word, which could be a search term supplied by a user. These nodes, or 'candidate nyms'. of the target word, are then ranked according to various characteristics, with similar ity scores based on a measure of relative frequency; that is, the ratio of observed to expected occurrence with reference to the frequency of the node in the corpus as a whole, and to its behaviour in collocation with the target word.

3.1 First results

3.1.1 Single word search terms

The project is committed in the first instance to discovering nymic pairs. In terms of information technology (IT), this would mean that a database user would submit a single word as a search term, and our system would supply a series of alternatives.

Table 5 presents the ranked nymic output for the word *director* in a corpus of business text extracted from the *Independent* newspaper. One has to ask oneself what one could expect to be presented in the way of thesaural information or alternative search terms, for the word. There is not actually a true synonym, at least in business text. We are offered *chief* and *head*, which could be near synonyms; also *chairman* and *director-general*, which can be assumed to be contrastive with *director*. There are also many proper names which are not disambiguated because, at this stage, multi-word items are not recognised unless already hyphenated. If they were, *chief executive* would also be a nym.

Interestingly, we find a series of hyponymic elements created by the addition of a modifier to the superordinate, *director.* These are *managing, finance, non-executive, deputy,* and *marketing.* This type of compounding is not surprising in the case of *director,* because the items are in a sense metonymic; aspects of the concept of *director*. But we have also observed it in relation to other words. The word *bid,* for instance, has the hyponyms *hostile bid, take-over bid, union-led bid.* We take these compound hyponymic or metonymic items to be a feature of the textual thesaurus. They are perhaps prevalent in business text, where superordinate nouns may often be metaphors of action. We shall investigate this phenomenon later on in our project.

Table 5: Nyms for node *director*

managing	john	chosen	succeed
finance	peter	elected	geoffrey
chairman	deputy	become	director-general
former	replaces	ian	officer
non-executive	general	bob	ward
chief	becomes	securities	stephen
executive	robert	announced	keith
resigned	gordon	appointment	graham
group	michael	head	barry
sir	borrie	richard	senior
appointed	marketing	acting	management
mr	david	financial	president

Table 6: Nyms for node *luxury*

goods	secondhand	luxuries	posh
car	three-wheeled	car-maker	resource-based
cars	electrohome	products	ricard
f-type	lexus	xj	scooters
up-tidy	wacoal	saloon	silkience
hotels	gleneagles	mass-market	wella
sub-compact	infiniti	fivestar	philips's
foreign-made	all-suite	s-cIass	uralmash
hennessy-Iouis	co-responsibility	maker	worryingly
two-seat	crude-oil	french	reputedly
hotel	dacha	accor's	luxurious
powerboat	full-size	all-new	mrb
Ivmh	houseware	badgemore	circle's
acura	lancashire-based	fus	kao
steigenberger	petro-chemical	fmm	lurgi
periquito	stahlverformung	frills	seibu
brinkhaus	tuborg	henlys	yugo
nukh	peking-backed	hino	bmw
sawn	formule	hypocritical	bentalls
pullman	goods'	inter-city	jan-sept
three-star	legends	juggernaut	middle-sized
drinks	loudspeaker	liqueurs	perfumes
jaguar	distilling	militarism	up-market
british-made			

Table 6 presents the nymic output for the word *luxury* from the *Financial Times.* This base form seems to be used instead of its inflexion *luxurious* to modify nouns. This is probably a strategic matter, where something different is implied by the base form, and it is probably quite common. It is not something reflected in a traditional thesaurus, however.

It can been seen from the list of items that *luxury* is ambiguous or rather multi-referential. We are offered a variety ofnyms which, in different contexts, are associated with the notion of luxury and/or might substitute in text for the word *luxury.* For instance, *f-type, jaguar, lexus, infiniti, xj, s-class, bmw.* In the context of hotels, we find *Steigenberger, Periquito, Gleneagles, five-star, posh.* A *dacha* is a luxury home in Russia, a *pullman* a luxury train, *Hennessy-Louis* a luxury brandy. In addition, reassuringly, we find inflexions of the word: *luxuries, luxurious.*

3.1.2 Two search terms

It is, however, more typical of a database search that more than one search term is submitted. We have therefore extended the system to identify nyms for two search terms. This happens to favour our approach, since two search terms are mutually disambiguating, in contextualising each other, and it allows our output to be more focussed on relevant aspects of the words, as indicated by that shared context.

Table 7: Nyms for *luxury* and *hotel*

inter-continental	gleneagles	bungalows	lodges
three-star	travelodge	bedroom	catering
steigenberger	mediterranee	luxurious	yue skyscraper
periquito five-star	sofitel	high-rise	three-bedroom
savoy	beefeater	novotel	antalya plush
houseware	westport	restaurant	regal two-bedroom
petro-chemical	sheraton	grosvenor	occupier
formule peking-backed posh	hotels	voyager	ritz
badgemore	marriott	no-frills	
	yaohan	rooms soft-loan hyatt	
	caterers		
	hilton		

Table 7 shows a much more referentially-restricted list of nyms for the search terms *luxury* and *hotel.* We find here synonyms of *luxury* in the context of hotel: *five-star, posh, luxurious, plush;* antonyms: *three-star, no-frills.* We also find a taxonomy of hotels: *Intercontinental, Steigenberger, Savoy, Mediterranee, Sofitel, Sheraton, Marriott, Grosvenor, Hyatt, Ritz.* This output is interesting from the IT point of view in indicating the central role of proper names in the textual thesaurus; from the linguistic point of view, the reduced set of synonyms and antonyms alerts us to the fact that the textual thesaurus as realised in any specialised domain is restricted, and does not make use of the full inventory of possibilities that are stored in the general mental thesaurus. The actual synonyms and antonyms used are also unexpected, if nice.

In our study so far, proper names are seen to feature centrally in the textual thesaurus. The picture emerging is one of a fairly flat semantic hierarchy: a *superordinate-hotel-and* below this, very often a taxonomy of actual proper names.

3.1.3 Proper names as search terms

Users of databases will often submit proper names as search terms. Table 8 presents the nyms for the word *conference* combined with the proper name *Marriott.* The Marriott Hotel is the venue for many conferences, and the idea here would be that a user might like to search for additional texts to do with conferences and conference hotels. The nyms fall into three recognisable referential groups: names of conferences, such as *Siggraph, Interforest, Helfex, Firex, Supercomputing;* other conference venues: *QEII, Plaisterers, Wallop, Accra, G-Mex;* and hotels: *Novotel, Gleneagles, Inter-Continental, Ramada.* This output is useful from the IT point of view because many nyms are reasonable search alternatives.

Table 8: Nyms for *marriott* and *conference*

financial	undersecretary	symposium	rejoining
siggrc:ph	accra	conferences	press
interforest	montreux	hancox	inter-continental
helfex	novotel	cardo	prev
qeii	bilspedition	newgate	birendra
plaisterers	supercomputing	gleneagles	potsdam
firex	jermyn	nf	ramada
wallop	11	tibbett	holdsworth
hotel	swithin's	qe	castle's
conf	g-mex	torquay	dematerialisation
haileybury	ennex	un's	sodexho
az	farringdon	inter-parliamentary	benard

3.1.4 Automatic identification of polysemy

The final task within the first objective has been to find an automatic means of identifying polysemy. This non-trivial task is considered desirable in order to achieve a closer match between the search term and the nymic response. If the various senses (or rather references or uses) of a search term can be established, then the user may decide which he or she had in mind in selecting the search term. Of course, whilst not every word is polysemous as such, virtually every word has the potential to be multi-contextual or multi-referen tial in text. Metaphor is a major reason for this phenomenon, in that it lifts a term from its literal context and inserts it incongruously in another for stylistic effect.

Collocation has once again been the means whereby we have identified polysemy. Taking the list of collocates for a word, we have in turn observed their collocational profiles, and on the basis of the degree to which those profiles are similar, have clustered the collocates. A public domain piece of clustering software, known as pam, has been modified for our use. Table 9 shows the clusters created out of the collocational profile for the word *air*.

Table 9: *air:* 5 collocate clusters

= Run 4 Cluster I =		= Run 4 Cluster 3 =	
air	0.19	aircraft	0.41
water	0.06	jets	0.20
hot	0.03	fighter	0.06
pollution	0.02	charter	0.02
breath	0.02	carriers	0.01
warm	0.01		
breathe	0.01	= Run 4 Cluster 4 =	
conditioners	0.01	rail	0.52
filters	0.01	freight	0.36
compressed	0.01	road	0.26
		express	0.21
= Run 4 Cluster 2 =		services	0.13
missile	0.44	links	0.11
Iraq	0.11	passengers	0.04
surface	0.08	strike	0.02
Iraqi	0.06		
attack	0.06	= Run 4 Cluster 5 =	
launched	0.04	Bosnian	0.73
		Bosnia	0.58
		Serb	0.56
		Serbs	0.55
		embargo	0.29
		forces	0.25

These could be said to correspond to the following senses:

air as atmosphere;
air as theatre of war;
air as medium associated with air-based transport; *air* as one of the several media of civil transport; *air* in association with war in the former Yugoslavia.

This is early output, and the system is still being improved, but we are very happy to discover that, on this simple basis, we are able not just to distinguish between two senses of a word-itself an achievement but to identify several. The clusters are more or less recognisable in conventional semantic terms, although it must be understood that there are some instances, as in category 5 above, where a dominant topic imposes a particular set of references on the text, which does not have a corresponding discrete sense.

Another complication in attempting to reflect the senses of a word is the fact that the commoner words of the language play a major role in creating the phraseology of text, not all of which contributes to its propositional content (Renouf, 1992, 1993b). Phrases such as *on the face ofU, at the end of the day* paraphrase lexical items such as *superficially, ultimately*, and the lexical words in them are not functioning independently as 'signifiants' (Saussure, 1964), Le., as referents to objects and concepts. This notwithstanding, we have some promising output even for a common word like *face*, as shown in Table 10:

Table 10: *face:* 5 collocate clusters

= Run 4 Cluster I =		= Run 4 Cluster 3 =	
slap	0.03	criminal	0.77
cliff	0.03	charges	0.68
sampling	0.03	trial	0.50
flies	0.01	prosecution	0.48
quota	0.01	disciplinary	0.40
= Run 4 Cluster 2 =		= Run 4 Cluster 4 =	
severe	0.56	daunting	0.66
disruption	0.12	task	0.29
losses	0.11	challenge	0.20
loss	0.04	prospect	0.20
consequences	0.04	must	0.10
difficulties	0.04		
criticism	0.03	= Run 4 Cluster 5 =	
problems	0.02	eyes	0.81
		hair	0.37
		tears	0.37
		nose	0.36
		staring	0.26

Here, the categories are as follows:

face in a variety of phrases, such as *slap in the face, the cliff face, fly in the face of, face* meaning 'be confronted by', in the context of such negative things
such as *loss, consequences, criticisms, problems; face* meaning 'be confronted by', in the context of legal matters; *face* meaning 'to have in prospect', in the context of challenging but not necessarily negative things; *face* meaning 'part of the head'.

4. Stage 2: Automatic identification of multi-word nyms

It is well known that lexical items often consist of more than a single word. It has therefore been necessary to find a way of identifying multi-word items automatically in text, with a view to improving our system. The aim would be to allow *John Major* to be entered as a single-item search term and for the system to match it to, for example, *Prime Minister*. We have just begun to work on this stage of the project, and two-word terms have been focussed on in the first instance.

By matching the collocational profiles of words, we have been able to isolate those which share collocates and occur significantly often next to each other.

In Table I I, the word pairs relate to the words *Blair, Heseltine* and *Euro* respectively. The word pairs are all bonafide, except for *Blair Labour* and *Heseltine Secretary.* These cases arise through the lack of punctuation in the newspaper text between a proper name and an appositional *designation-Tony BIair Labour Party Spokesman of Employment,* and *Michael Heseltine Secretary of State for Defence/the Environment.* Some items look strange because they are in fact part of a larger multi-word nymic unit. *Tory Euro,* for instance, is part of *Tory Euro Rebel/Sceptic.* The initial output is felt to be hopeful.

Table 12 lists items which pair with the word *party,* and again the output is promising: *advance party, garden party, party line, party dress, dinner party, guilty party, interested party, working party* are all good candidates. The rather extensive output does raise the question of which multi-word units we want to retain and which discard: and beyond that, what a bonafide compound is: must it be metaphorical, must it mean more than the sum of its parts, and so on. Generations of linguists have investigated this area in detail, establishing a whole battery of criteria for the class of nominal compounds, and

Table 11: Ranked list of words pairing with the nodes *Blair, Heseltine,* and *Euro*

B/air	*Hese/tine*	*Euro*
Neil Mr	Mr	Atlantic
Les	Secretary	Brokers
Hunter	Michael	sceptic
RN	plan	Mix
Labour	Politics	Disney
Barley		scepti cs
Mrs		Africa
Tony		federalism
Colonel		Disneyland
Stewart		enthusiasts
bt		Tory bank
		elections
		MP MPs

Table 12: Ranked list of words pairing with the node *party*

talks funds	cells dress	secondary	separate
advance	dinner	Inkatha	thrown
source	wins	lines	holds
managers	House	Christmas	workers
garden	inter	cell	rule
Tory	guilty tour	interested	tea politics
democracy	cross	chairman	office
rules	political	Senior	featuring
line	systems	working	boating
non drinks	consensus	multi	agreement
		anniversary	

we shall have to at a later stage, in our description of this phenomenon in the textual thesaurus. For IT purposes, however, it probably suffices for us to adopt a loose definition, since the user will be regularly submitting multi-word units of varying kinds and degrees of fixedness.

5. Conclusion

We continue to discover facts about the thesaurus in journalistic text, and in particular financial text. To summarise some of our findings so far, they are as follows.

The range of lexical realisations found in text for each type of semantic relationship is different and more extensive than that found in a traditional thesaurus, partly because it crosses word boundaries. Within a single textual domain or text type, however, there will be a specific, more restricted range of lexical choices.

Related to this, we have observed that, in anyone text, the semantic hierarchy, the system whereby related things and concepts are referred to at progressive levels of generality, as in *poodle, dog, animal, mammal,* is fairly flat. So far, it appears in our data to be primarily two-level, as in *hotel: Sheraton.*

This means that hyponymy is a dominant nymic relation in our data. It manifests itself partly in a preponderance of proper names, these being the end points in a hierarchical system of reference. Proper names seem in fact to be central to nymic relationships in journalistic text. Given this centrality, proper names deserve greater attention than they have had hitherto in linguistic description.

Nyms are also often multi-word items; that is to say, nymic and collocational relations cannot be described independently of each other. Many multi-word nyms are hyponyms which have been formed by modification of a superordinate term, as in *managing director, talre-over bid.*

References

Firth, J.R. (1957). *Papers in linguistics,* 1934-1951. London: Oxford University Press.

HarperCollins (ed.) (1995). *Col/ins thesaurus: The ultimate word finder.* London: HarperCollins.

Renouf, AJ. (1992). 'What do you think of that: A pilot study of the phraseology of the core words of English', in Leitner, G. (ed.), *New directions in English language corpora: Methodology, results, software developments.* Berlin and New York: Mouton de Gruyter. 301-318.

Renouf, A.J. (I 993a). 'A word in time: First findings ITom dynamic corpus investigation', in Aarts, 1., P. de Haan and N. Oostdijk (eds.), *English language corpora: Design, analysis and exploitation.* Amsterdam: Rodopi. 279-288.

Renouf, A.J. (1993b). 'What the linguist has to say to the information scientist', *The Journal of Document and Text Retrieval,* 1.2. 173-190.

Saussure, F. de (1964). *Cours de linguistique generale.* 4th ed. 1st ed. 1916. Paris: Payot.