Refining the Automatic Identification of Conceptual Relations in Large-scale Corpora

Alex Collier, Mike Pacey and Antoinette Renouf
Research and Development Unit for English Studies
University of Liverpool

Abstract

In the ACRONYM Project, we have taken the Firthian view (e.g. Firth 1957) that context is part of the meaning of the word, and measured similarity of meaning between words through second-order collocation. Using large-scale, free text corpora of UK journalism, we have generated collocational data for all words except for high-frequency grammatical words, and have found that semantically related word pairings can be identified, whilst syntactic relations are disfavoured. We have then moved on to refine this system, to deal with multi-word terms and identify changing conceptual relationships across time. The system, conceived in the late 80's and developed in 1994-97, differs from others of the 90's in purpose, scope, methodology and results, and comparisons will be drawn in the course of the paper.

Introduction

The team at Liverpool has created over the years a series of automated systems for handling and extracting information from large textual corpora. These systems consist of software and knowledge bases derived from those same textual resources. Most recently, the system known as ACRONYM (Automated Collocational Retrieval of 'Nyms') has involved the identification of conceptually related items. These are referred to as 'nyms', by analogy with 'synonyms' and other sense related items, as a reflection of the fact that it is conceptual similarity that is being discovered through collocation.

Like its predecessors, the ACRONYM system has a dual purpose. On the one hand, it is intended to generate pairs or clusters of items which can function as alternative search terms in a diachronic text retrieval environment. On the other, it is intended to support a description of the thesaurus in text. In the latter application, the precise nature of the nyms generated by a given target word is important.

The basic system for identifying conceptual relations

The starting point for the nymic identification system is the raw text from which thesaural relations are to be derived. This corpus currently contains over 300 million words from The Independent newspaper from 1988 to 1997. As with similar work (e.g. Brown et al 1992), the size of the corpus makes preprocessing such as lemmatization, POS tagging or partial parsing, too costly. The sole preprocessing performed on the corpus is thus the relabelling of numeric tokens into general categories.

During this preprocessing stage, the corpus is also integerised and indexed. This increases efficiency during later processing stages, in particular the creation of the collocate database. Within the system, collocates are by default defined as the four words to the left and right of every word. The only exceptions to this rule are that collocates are not recorded for a set of 253 stopwords (high frequency terms; grammatical words, numeric labels and some verbs), and these are also not recorded as collocates of any other word. The raw frequencies of left and right span collocates for a given word-pair are merged, and their significance measured, using a Z-score statistic. Statistically significant collocates are then stored as a sparse matrix in the collocate database. A corpus of 3×10^8 words and 1.5×10^6 word types produces just over 4.8×10^6 statistically significant collocates (using a liberal threshold for significance).

We refer to the set of statistically significant collocates for a given word as a 'collocational profile'. Similarity between any

two given words is then measured through comparison of their profiles; the measure itself is based on the size of the profiles of both words and the number of collocates they share (i.e., second order collocation), and also on the collocation between those two words (first order collocation).

The definition of what constitutes first-order collocation differs for different researchers. For example, the corpus described in Grefensttette (1992) is comparatively small, allowing for extensive preprocessing, including POS tagging and partial parsing. Only modifiers and their modified nouns are recorded as collocating pairs, allowing for a more fine-grained analysis, of just a subset of word classes. In contrast, Schütze and Pedersen (1995) treat the set of collocates for a word as a vector containing the frequencies of collocation with other words occurring within a 40-word window. Futrelle and Gauch (1993) use a similar approach but preserve positional information (i.e., the number of words to the left and right of the target word). Positional information is also retained by Brown et al (1992), who store collocation information as word n-grams.

For very large corpora, a lot of collocational information will be generated, making any form of collocational similarity measurement computationally expensive. Again, researchers have adopted differing approaches to this problem. Schütze and Pedersen (1995) build their collocate vectors using a bootstrap method involving increasingly larger sets of the lexicon, finally constructing a low (20) dimensional word vector space by Singular Value Decomposition. A simpler method is employed by Futrelle and Gauch (1993), whereby collocate vectors are recorded for all word forms, but for each word form, only its frequency of co-occurrence with the top 150 most frequent word forms is recorded. The researchers use a 2-word window to the left and right of the target word, but they also preserve the positional information of the collocates, resulting in a 600-dimension word-vector of mutual information measurements.

While there are many different ways to record collocates, the primary difference between the ACRONYM collocate database and those detailed above is the omission both of collocation involving grammatical function words, and of positional information. By omitting these two elements, the resulting collocate database focuses less on the syntagmatic similarity between words and more on their paradigmatic relations.

Generating conceptually-related word pairs

The definition and purpose of word similarity measures in the above systems also differ. Vector-based models define similarity as the cosine measure between two words, and the purpose of Schütze and Brown's vector-based work is to cluster grammatical word classes automatically. Mutual information-based approaches, such as those of Brown et al (1992) and Futrelle and Gauch (1993), measure word similarity in the context of a set of words to be clustered, typically with the aim of clustering for general similarity. The Jaccard coefficient measurement offered by Grefenstette (1992) is the nearest to our own, defining similarity simply in terms of the number of shared 'attributes' between two words against the number of attributes of both.

While Schütze and Pedersen (1993), Brown et al (1992) and Futrelle and Gauch (1993) all demonstrate the ability of their systems to identify word similarity using clustering on the most frequently occurring words in their corpus, only Grefenstette (1992) demonstrates his system by generating word similarities with respect to a set of target words. His purpose is to allow a user to specify a target word, and have the system return an ordered list of related words. To this extent, the purpose of the basic ACRONYM system is echoed in Grefenstette's work.

Given the liberal thresholds currently used in the ACRONYM system, such a list of conceptually related words may contain several tens of thousands of entries. As the size of the lexicon renders it computationally infeasible to calculate all word-pair similarities in advance, the system generates word similarity measures for a given word on the fly, using the collocate database described above.

Examples of conceptually-related, or nymic, output are given in Table 1 for the node words *key*, *medicine*, *pretty* and *testing*.

Node	Nyms
key	factor role element issues areas issue elements figure players component
medicine	complementary alternative food herbal preventive genito-urinary modern conventional clinical science
pretty	good sight looks look awful girl silly boring looked stupid
testing	nuclear random positive curriculum DNA drug genetic HIV psychometric tests

Table 1: Ten top nyms for nodes *key, pretty, medicine, testing*

Modification to software to increase semantic nature of nymic output

The nymic output in Table 1 contains collocates and other related items, which are relevant in principle for IT purposes, but which for linguistic purposes may be usefully separated out. This is achieved by using only second order collocation, which boosts semantically (and morphologically) related nymic output, as can be seen in Table 2.

Node	Nyms
key	crucial important vital significant essential main fundamental major strategic specific
medicine	medical medicines sciences mathematics biology science chemistry psychology physics clinical
pretty	fairly quite incredibly extremely terribly really nice extraordinarily lovely sexy
testing	tests test tested assessment monitoring screening research rigorous clinical curriculum

Table 2: Nyms for nodes *key, pretty, medicine, testing* suppressing first order collocates

The Deese Antonyms

The focussing effect achieved by suppressing first order collocational information may be further demonstrated with reference to the work of Deese (1964), cited in Grefenstette (1992), and specifically to a set of conceptually-related antonymic pairs which Deese had identified by a series of psycholinguistic tests. Grefenstette hypothesised that any system identifying shared collocation between words would pick up the Deese antonyms as being strongly related; he experimented by feeding the 'primer' word for each of those antonyms into his SEXTANT system, and listing the 10 most closely related words produced.

The same 'primer' words were fed into the refined ACRONYM system, i.e. with first-order collocates suppressed, and the results are displayed in Table 3 on the next page. (Deese antonyms, where they occur, are capitalised).

A basic comparison reveals that the ACRONYM system yields a similar level of results to Grefenstette's system. 15/33 of the Deese antonyms occur in SEXTANT output as first or second most-related word, compared with 13 generated by ACRONYM; whilst 16 of the Deese antonyms appear within the top 10 most related words of SEXTANT output, compared with 18 in output from ACRONYM. Similarly, as with Grefenstette's findings, there are cases where ACRONYM yields a non-Deese antonym which is nevertheless close: see for instance: *big-small; dark-pale; deep-surface; happy-unhappy; new-existing; old-modern*. A scrutiny of the actual contents of each ACRONYM list further reveals that, like Futrelle and Gauch, ACRONYM when using the particular non collocate upweighting method discovers "... (entire) graded fields, rather than just pairs of opposites". Of particular interest in this respect are the results for *big, easy, fast* and *strong*.

For some words in Table 3, failure to relate closely to their Deese antonym can in part be explained by textual domain. For example, in the listing for *empty*, the strongest nyms reflect the sense of an empty building or structure, suggesting that such a context predominates throughout the corpus. Likewise, with the word *pretty*, we see that intensifiers predominate as nyms, with the synonym *lovely* appearing only in 9th place. Analysis of a random sample of the corpus reveals that *pretty* is indeed predominantly adverbial, and only rarely adjectival.

Multi-Word Nyms

As described, the basic ACRONYM system generates information on both first and second order collocates within its single-word nymic output. First order collocation can be suppressed to enrich the semantic information, as

Primer	Nyms
active	actively organisations groups involved activities effective activity developing vigorous encourage
alive	DEAD loved dying die mum frightened loves buried forever loving
back	down away ball straight FRONT again foot around yards feet
bad	GOOD worse dreadful awful poor terrible nasty stupid silly appalling
big	bigger huge large biggest major larger smaller small massive enormous
black	WHITE red brown blue wearing pink yellow green grey leather
bottom	TOP relegated side relegation feet foot floor inches table straight
clean	wash cool dry cleaning smooth kitchen warm water washing shiny
cold	HOT warm dry cool boiling wet salt boiled cooked damp
dark	grey pale brown green bright white blue red thick purple
deep	deeper profound dark intense sand depth surface feelings thick mixture
dry	dried hot warm soft brown creamy crisp salt lemon cold
easy	easier difficult impossible harder HARD simple able quick enough unable
empty	deserted filled crowded derelict crammed windows floor surrounded cramped nearby
fast	faster bowler pace speed bowlers bowling SLOW slowly slower quick
happy	pleased unhappy mum nice happily enjoy relaxed OK loving cheerful
hard	harder difficult impossible EASY easier unable trying able enough tough
heavy	artillery thick huge massive metal large high rain vehicles low
high	LOW higher lower levels rising highest level increased increase falling
large	SMALL huge larger vast smaller big tiny substantial mainly plastic
left	leaving ball leave yards corner back injured pulled minutes shot
long	SHORT longer hair dark slow white wearing black down wide
narrow	steep WIDE broad stretch paths hill lined tiny path brick
new	existing technology proposed latest development plans current commercial systems design
old	ancient Victorian traditional white houses buildings around man black modern
pretty	fairly quite incredibly extremely terribly really nice extraordinarily lovely looks
rich	spicy delicious flavours wealthy sweet fruit flavour ripe soft texture
right	wrong LEFT want freedom rights able back necessary wanted law
rough	muddy sand wet dirt grass tricky dusty mud trees damp
short	LONG length brief straight tight balls wide wicket quick ball
sour	SWEET salty soured spiced delicious tomato soy creamy chilli pungent
strong	stronger strongest powerful WEAK strength sharp steady solid underlying strongly
thin	THICK brown strips pale slices orange fat white creamy soft

Table 3: Deese antonyms in ACRONYM nymic output

demonstrated in Tables 1 and 2. Often, however, these collocates are really part of multi-word units, combining with other words to form hyponyms. They often combine with the target word itself, thereby forming hyponyms of the type 'ordinate plus modifier'. These items are of prime importance in linguistic description, since they represent hitherto undocumented differences between the textual thesaurus and the mental lexicon.

Another refinement to the basic ACRONYM system is therefore achieved when the nyms of the single-word nymic output are re-combined into multi-word units which better represent the target concept. This procedure is carried out in two stages. First, the list of nyms is processed by a software module which attempts to identify the most likely word pairs that could be created by combining the individual nyms, making use of a variety of measures including collocational as well as contextual clues from the corpus database. The resulting list of word combinations, which need not necessarily be adjacent, is passed to a second-stage module, which checks which of these candidate word pairs have collocational environments similar to the original node word. The benefits of this approach are that no a priori word-pair list needs to be established, this being decided by the contents of the nym list and by the corpus, and that no collocational profiles need to be stored for word pairs.

Tables 4 and 5 display multi-word nyms for *therapy* and *weapons*.

gene therapy	drug group
shock therapy	hormone treatment
replacement therapy	drug AZT
group therapy	group sex
speech therapy	counselling sessions
occupational therapy	intensive treatment
therapy sessions	shock treatment
hormone replacement	medical treatment
cancer drug	drugs group
drug tamoxifen	drug treatment
sex therapist	cancer patients

Table 4: Multi-word nymic output for *therapy*

nuclear weapons	nuclear capability
chemical weapons	nuclear non-proliferation
biological weapons	arms cache
nuclear arms	sub-machine guns
anti-tank weapons	short-range nuclear
nuclear warheads	short-range missiles
nuclear arsenal	chemical warheads
weapons capability	nuclear missiles
nuclear arsenals	semi-automatic weapons
land-based nuclear	anti-tank missiles
land-based missiles	ballistic missile
chemical arms	ballistic missiles
air-launched nuclear	missile launchers

Table 5: Multi-word nymic output for *weapons*

In Tables 4 and 5, a series of multi-word hyponyms have emerged, several consisting of adjectives or nouns modifying the node or synonyms of it.

Multi-word nodes are a further refinement of the system. As with multi-word nymic output, the collocate profiles for multi-word nodes are generated on the fly. Table 6 presents an example of the multi-word nymic output for the multi-word node *Soviet Union*.

Soviet Union	Soviet bloc
eastern Europe	Algirdas Brazauskas
Soviet republics	Communist countries
Soviet republic	Soviet leaders
Soviet forces	East Germany
Eastern bloc	Nato leaders
Soviet nuclear	Pact countries
Communist leaders	former Eastern
Communist government	political independence

Table 6: Multi-word nymic output for *Soviet Union*

Semantic Clustering

As well as generating flat lists of semantically related words, the ACRONYM system can perform clustering upon a set of nyms, in order to reveal their semantic inter-relationships. In ACRONYM, the set of words to be clustered is usually one of the flat lists of nyms of the kind displayed above. This is in contrast to work by researchers such as Schütze and Pedersen (1992), Brown et al (1992) and Futrelle and Gauch (1995), where it is often the most frequent words in the lexicon which are clustered, predominantly with the purpose of determining their grammatical classes.

ACRONYM uses two publicly available clustering tools, PAM and AGNES, described in Kaufman and Rousseeuw (1990). The first, PAM (Partitioning Around Medoids), is a k-medoid partitioning method, while AGNES is a variant on agglomerative nesting. Both algorithms allow object-relations to be represented by a similarity measure, which we take as the collocational profile similarity measure described earlier. An example of PAM output is shown in Table 7.

=== Cluster 1 ===	=== Cluster 5 ===
electricity 0.21	litres 0.64
privatised 0.15	gallons 0.48
sewerage 0.09	pints 0.38
supply 0.08	=== Cluster 6 ===
mains 0.08	pan 0.27
companies 0.07	heavy-based 0.20
newly-privatised 0.06	heavy-bottomed 0.20
=== Cluster 2 ===	lidded 0.11
hot 0.09	=== Cluster 7 ===
aquifers 0.05	salt 0.14
cavern 0.05	Dissolve 0.10
=== Cluster 3 ===	Soak 0.06
bottled 0.19	=== Cluster 8 ===
ice-cold 0.07	Add 0.08
carbonated 0.05	winched 0.06
=== Cluster 4 ===	=== Cluster 9 ===
oz 0.30	pollution 0.19
tablespoons 0.17	sediments 0.08
ml 0.16	fast-flowing 0.07
tablespoon 0.14	polluted 0.07
fl 0.09	

Table 7: PAM Clustering for *water*

The PAM-generated clusters in Table 7 are created from the top nyms for *water*, and reflect several different meanings (senses, uses or references) that are associated with the node

word, namely: (1) 'water utility', (2) 'body of water', (3) 'type of drinking water', (4) 'fluid measurement', (5) 'unit of water', (6) 'liquid used in cooking', (7) 'medium for certain domestic processes', and (9) 'water in various more or less pure states'. Some of these senses are fairly conventional, others are more contextually determined. Not every cluster is adequate; here, Cluster 8 is weak and uninterpretable. Taken overall, it seems that this type of clustering does sharpen the picture for the user of the system.

While previous researchers have used agglomerative nesting clustering (e.g. Brown et al (1992), Futrelle and Gauch (1993)), comparisons with our work are difficult to draw, due to their use of the 1,000 commonest words from their respective corpora.

In Brown et al (1992), the authors provide some sample subtrees resulting from such a 1,000-word clustering. The sets of words from each subtree have been fed into the ACRONYM clustering system, and the results from AGNES are shown below. This is not strictly a fair comparison, as the clustering of a superset of these words would doubtless create a different structure. Nevertheless, it appears that ACRONYM organises these subsets into a more satisfactory taxonomy, in contrast with a tendency in Brown et al's system to produce right-heavy taxonomies.

In Fig. 1, the last example highlights a distinction between syntax and semantics. While Brown et al's system splits the four words along the singular/plural divide (i.e., *rep-representative* and *reps-representatives*), ACRONYM splits them semantically; the abbreviated versions refer to sales-people or those representing travel companies, whilst the full versions are used in a political context.

Identifying Change in Conceptual Relations

The ACRONYM database has also been designed in such a way that it can be accessed diachronically. This facility was incorporated in order to ensure that the system remains up to date in its application to text retrieval and linguistic description, and it has already enabled the Unit to establish a new words service within its web site (http://www.rdues.liv.ac.uk/newwds.html).

The team first explored the way in which language changes over time in the AVIATOR Project (Renouf 1993, Collier 1993, Blackwell 1993), where they investigated the dynamic aspects not only of single words, but also of the collocational behaviour of those words, with the goal of identifying new collocations or changes in meaning. In ACRONYM (Renouf 1996, Collier & Pacey 1996), the collocational and diachronic concepts have been developed considerably, taking advantage of improvements in technology and the greater availability of electronic text. The result is an integrated system of databases and indexes which can be accessed as one virtual entity or divided into any desired configuration of its constituent parts. In the current database, the smallest accessible component, which we refer to as a 'segment', consists of three months' of text from national UK newspapers, containing on average eight million running words (tokens).

Each segment is composed of an integerised corpus database, providing all the usual corpus- and text-retrieval facilities from simple frequency information for a single word to the full KWIC, sentence or article context for boolean (multi-word) searches. The frequency data is readily extractable, allowing a word or phrase to be 'tracked' over time. In addition, each segment has one or more collocate databases which store profiles for each word in the corpus. By comparing the output from two collocate databases, the change in collocational behaviour of any node can be identified in a similar fashion to a change in frequency-of-occurrence. As established in the AVIATOR Project, an alteration in a word's profile signals a change in its meaning, with a consequent change in the set of words which can be regarded as its semantic equivalents. If we require time scales longer than three months, the software needs to perform a comparison across several collocate databases rather than just two. In order to accelerate this process, we have added a facility for creating merged databases, for example combining all four databases for 1994 into one. By using this in conjunction with a similarly merged database

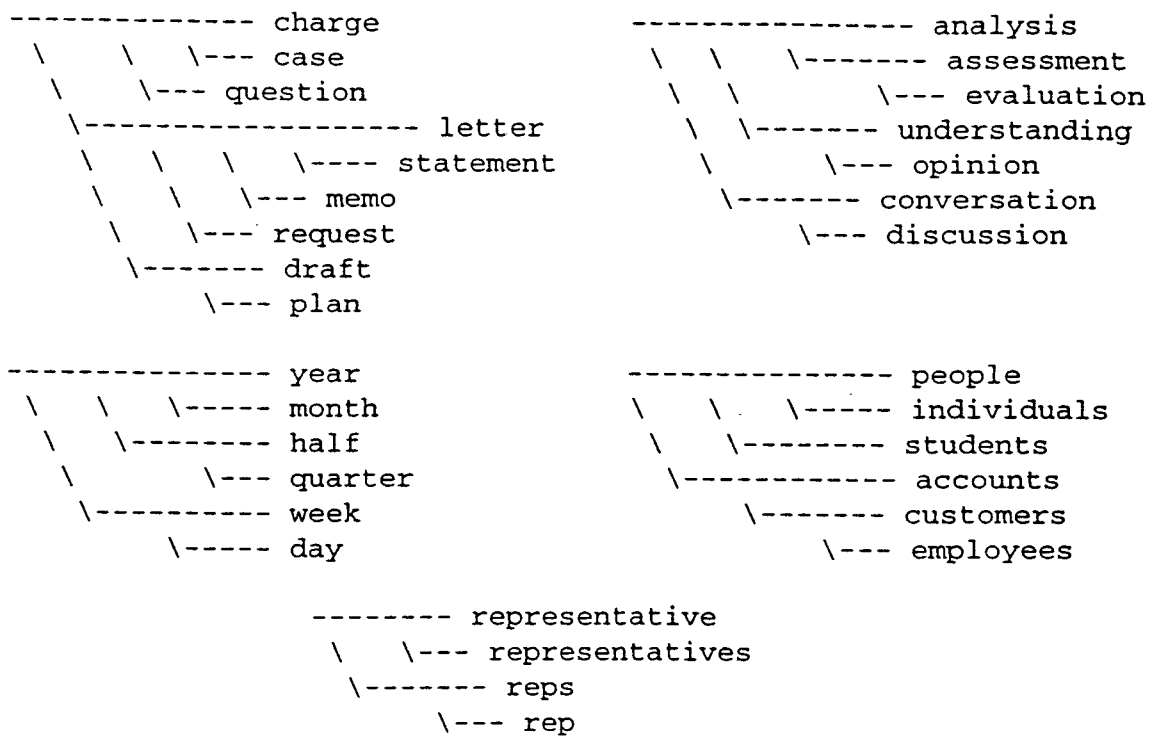

Figure 1: AGNES clusterings of selected subtrees

based on other individual segments, the collocate comparison process can be carried out more efficiently.

When identifying nymic change, we generally increase the time period to a whole year, to avoid recording any seasonal fluctuations in real world events. The usual procedure is to create two merged collocate databases, one for the year in question (the target corpus) and another of all segment databases prior to that year (the baseline corpus). These two databases are then compared and any significant change in collocate profiles is recorded. This may be done for individual words or all the words in the corpus. In looking at the changed profiles, the collocates are categorised into four sets:

'up' collocates
: those which have increased in significance in the target corpus;

'down' collocates
: those which have decreased in significance in the target corpus;

'new' collocates
: those which have appeared for the first time in the target corpus;

'gone' collocates
: those which appeared in the baseline corpus but which are no longer present in the target corpus.

The normal process of nym identification, as explained earlier, finds candidates which have as many collocates as possible in common with the target word. In monitoring the change in nymic relationships, however, only those collocates which are considered to have changed are involved in this process. The semantic proximity of the target word to the candidate nyms is therefore measured in terms of the number of changed collocates the two words have in common. If the task is to identify new nyms, then only 'up' and 'new' collocates are used; conversely, 'down' and 'gone' collocates are employed in finding nyms which have decreased in significance. This is exemplified in Table 8,

which presents the 1997 'up' and 'new' collocates for the node word *crisis*.

three	coal	IMF
hit	NHS	Outlook
South	funding	hospitals
Opera	prison	injury
opera	overcrowding	Montserrat
East	Far	LUCY
M	spread	Beef
financial	environmental	Thai
higher	Still	Coal's
currency	university	escalating
Business	CHRIS	Thailand's
Kong	Japan's	striker
Hong	Thailand	mid-life
recruitment	deepened	peso
Asia	Korean	millennium
Asian	South-east	MATTHEW
Pacific	lurched	WARD
illustrated	South-East	BSE
education	Guinea	blah
discipline	Chris	VINES
Blair	winter	
analysts	Asia's	

Table 8: 1997 'up'/'new' collocates of *crisis*

It can be seen that the collocates in Table 8 refer to a number of crises topical in 1997. The next stage involves the identification of words which share these collocates with the original node word *crisis* in the 1997 collocate database. The output from this is given in Table 9.

financial	Korea	problems
crisis	economy	investment
Asian	Hong	South-east
currency	market	fears
funding	Far	officials
East	Japan	authorities
South	Kong	stock
Asia	Letter	health
economic	tiger	Bangkok
markets	turbulence	IMF
yesterday	turmoil	economies
Business	education	

Table 9: 1997 'up' nyms of *crisis*

The nyms in Table 9, presented in descending order of strength of association, focus more clearly on the financial crisis looming in South East Asia.

Since one of the chief goals of this methodology is to provide up-to-date information on thesaural equivalents, it can also be used to find nyms which have declined in significance. The output is harder to interpret, since the difference in size between the baseline and target databases results in many more down/gone collocates than up/new ones. For *crisis*, as an example, there were 3,187 down/gone collocates but only 68 new/up ones. Nevertheless, the nyms which are generated by using the down/gone collocates can be useful and interesting. Tables 10a and 10b show nyms for *war*, using 1993 as the target corpus and all previous data (1988-1992) as the baseline.

civil	siege	ethnic
genocide	warring	fighting
crimes	fighters	enclave
war	Karadzic	besieged
Muslim-led	commanders	offensive
stronghold	embargo	UN
Gorazde	Yugoslavia	wars
cleansing	Muslims	convoys
Bosnian	atrocities	Izetbegovic
Bosnia	factions	Belgrade
aggression	Bosnia's	
shelling	Vance	

Table 10a: 'up' nyms of *war* (1993)

In Table 10a, the 'up' nyms are presented and it can be seen that these all relate to the civil war in the former Yugoslavia. In Table 10b, in contrast, those nyms are listed which in 1993 became less closely associated with *war*.

Gulf	Shias	military
war	HAERI	Egyptians
Iran-Iraq	MANAGUA	ADEL
IRAQI	Iraq	Scuds
Jordanians	al-Arab	CAIRO
NICOSIA	SAFA	Dhahran
US-led	KABUL	al-Assad
Khafji	emirate	starve
TEHRAN	ISLAMABAD	oilfields
waterway	Barco	UAE
invading	DARWISH	Saddam

Table 10b: 'down' nyms of *war* (1993)

The main reference reflected in the 'down nyms' of Table 10b is to the Gulf War, which followed Iraq's invasion of Kuwait. The implication of this evidence is that by 1993, the Gulf War had ceased to figure so prominently in

our corpus data and so had become less strongly associated with the concept of *war*.

Concluding Remarks

This paper has described the basic ACRONYM system, a set of tools which has relevance both to text retrieval applications and to linguistic description. The focus here has been on outlining the recent modifications which have been carried out to refine the nymic output to facilitate the linguistic task of describing the textual thesaurus. Several of them, in particular semantic clustering, are also intended to improve performance in document retrieval. The nymic output from ACRONYM intuitively appears to have the potential to increase both recall and precision, and initial tests of its effectiveness in this regard have been carried out, by using nymic output to extract article headlines. The next stage of the research will focus more closely on the evaluation and optimisation of the system as a text retrieval facility.

Bibliography

Blackwell, S. (1993) 'From dirty data to clean language' in *English Language Corpora: Design, Analysis and Exploitation, Papers from the 13th International Conference on English Language Research on Computerized Corpora*, Nijmegen 1992, Aarts, J., P. de Haan and N. Oostdijk (eds) Rodopi, Amsterdam, pp. 97-106.

Brown, P. F., P. V. de Souza, R. L. Mercer, V. J Della Pietra and J. C. Lai (1992), 'Class-Based *n*-gram Models of Natural Language' in *Computational Linguistics*, volume 18, number 4, ACL, MIT Press, pp. 467-479.

Collier, A. (1993) 'Issues of large-scale collocational analysis' in *English Language Corpora: Design, Analysis and Exploitation, Papers from the 13th International Conference on English Language Research on Computerized Corpora*, Nijmegen 1992, Aarts, J., P. de Haan and N. Oostdijk (eds) Rodopi, Amsterdam, pp. 289-298.

Collier, A. and M. Pacey (1997), 'A Large Scale Corpus System For Identifying Thesaural Relations', in *Corpus-based Studies in English - Papers from the seventeenth International Conference on English Language Research on Computerized Corpora (ICAME 17)*, Ljung, M. (ed) pp. 87-100.

Deese, J., (1964), 'The associative structure of some common English adjectives', in *Journal of Verbal Learning and Verbal Behaviour 3*, pp. 347-357.

Firth, JR (1957), *Papers in Linguistics, 1934-1951*. London: Oxford University Press.

Futrelle, R. P and S. Gauch (1993), 'Experiments in Syntactic and Semantic Classification and Disambiguation Using Bootstrapping', in *Acquisition of Lexical Knowledge from Text*. 1993, pp. 117-127. Columbus, OH. Assoc. Computational Linguistics.

Grefenstette, G, (1992), 'Finding Semantic Similarity in Raw Text: the Deese Antonyms', in *AAAI Fall Symposium Series: Probabilistic Approaches to Natural Language (Working Notes)*, pp. 54-60. Cambridge MA.

Kaufman, L. and Peter J. Rousseeuw (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley and Sons, NY.

Renouf, A. (1993) 'A word in time: First findings from the investigation of dynamic text' in *English Language Corpora: Design, Analysis and Exploitation, Papers from the 13th International Conference on English Language Research on Computerized Corpora*, Nijmegen 1992, Aarts, J., P. de Haan and N.Oostdijk (eds) Rodopi, Amsterdam, pp. 279-288.

Renouf, A. (1996), 'The ACRONYM Project: Discovering the textual thesaurus', in *Synchronic corpus linguistics - Papers from the 16th International Conference on English Language Research on Computerised Corpora (ICAME 16)*, Percy, C. E, C. F Meyer and Ian Lancashire (eds). pp. 171-188.

Schütze, H. and J. Pedersen (1993) 'A vector model for syntagmatic and paradigmatic relatedness' in *Proceedings of the 9th Annual Conference of the UW Centre for the New OED and Text Research*, pp. 104-113, Oxford, England.

Schütze, H. and J. O. Pedersen (1995), 'Information Retrieval Based on Word Senses', in *Fourth Annual Symposium on Document Analysis and Information Retrieval*, pp. 161-175, Las Vegas NV.