LES NYMS: EN QUETE DU THESAURUS DES TEXTES

ANTOINETTE RENOUF *Research and Development*
Unit for English Studies, University of Liverpool *

1. Introduction

La recherche dont il sera question ici est en cours dans le cadre du projet ACRONYM, lancé en 1994, pour une durée de trois ans, par le gouvemement et l'industrie britanniques. I1 s'agit d'une collaboration dont les partenaires sont l' *Université de Liverpool, FT Info* (département 'archives' du *Financial Times)* et la *Press Association* (agence de presse qui alimente la presse regionale). Le projet a pour objectif de développer un systeme d'identification automatique des éléments constituant le thesaurus de textes informatisés. ACRONYM est un acronyme pour 'The Automatic Collocational Retrieval of Nyms'.

2. Definition du 'nym'

Les nyms composent des paires de mots (ou d'items lexicaux) qui apparaissent dans une base de donnees textuelles dans des contextes semblables (collocations) et dont la relation, semantique ou autre, est consideree comme non accidentelle. Nous les appelons nyms parcequ'ils ne correspondent pas a des categories semantiques conventionnelles. Les nyms associes au mot *luxury,* par exemple, rassemblent des termes approximativement synonymes *commefive-star, S-class,* des termes plut6t antonymes comme *no-frills, threewheeled,* des hyponymes contextuels *(jaguar, lexus)* et des termes plus inattendus. Le choix de I' appellation nym traduit le fait que, dans plusieurs applications, la nature precise de la relation n'est pas cruciale.

2.1. Applications du systéme

Une fois répertoriés, les nyms ont deux utilisations. La première concerne les programmes de recherche élargie dans de grosses bases de données textuelles. Les nyms pourront constituer des alternatives utiles pour le choix des motsclés en recherche documentaire. Il n' est pas fondamental ici de savoir ce que recouvre la notion de nym, pour autant que les nyms selectionnent, presque par définition, des séquences de texte similaires.

Dans les 1ogiciels de recherche textuelle, la fonction 'thésaurus' est sou vent une structure de travail vide, dans laquel les utilisateurs sont invités à introduire leurs propres synonymes ou des references de remplacement pour un mot donné. Par exemple, ils peuvent donner le sigle de leur société ou organisation pour que, chaque fois qu'un utilisateur presénte le nom complet comme mot-clé, la fonction 'thésaurus' soit activée et qu'une recherche 'secrete' soit déclenchée en meme temps sur la forme abrégée.

Les essais pour rattacher au texte des thesaurus existants se sont montres peu satisfaisants. Un thesaurus general est susceptible d' offrir des alternatives trop eloignees d'un texte particulier. L'entree du mot *luxury* dans le *Collins Thesaurus* (1995) est typique a cet egard. Les synonymes proposés sont les suivants:

Luxury :
1) *affluence, hedonism, opulence, richness, splendour, sumptuousness, voluptuousness;*
2) *bliss, comfort, delight, enjoyment, gratification, indulgence, pleasure, satisfaction, wellbeing;*
3) *extravagance,jrill, indulgence, nonessential, treat.*

Ces termes alternatifs n'appartiennent pas au vocabulaire journalistique, particulierement de la presse des affaires ou financière. Ils sont dans I'ensemble trop litteraires. *Frill* est le seul à avoir un caractère journalistique, et il apparaît en effet dans une base de données, bien que seulement sous la forme *no-frills.*

Un autre problème est que les thésaurus traditionnels sont grammaticalement marqués. C' est -á-dire qu' ils assignent une catégorie grammaticale a un item et ne proposent que des mots appartenant á cette même catégorie. En réalité, l' équivalence de sens passe sou vent par des sus les divages entre catégories grammaticales. Un problème lié, illustré par le mot *luxury*, est que ce sont souvent les noms, plutôt que les adjectifs, qui servent à modifier d'autres noms. Ainsi, *luxury* est, de preférénce à *luxurious*, le modifieur conventionnel dans les textes journalistiques. Le thésaurus des termes alternatifs de ce substantif doit donc comprendre les adjectifs.

Quelques thésaurus spécialisés ont été créés manuellement, mais l'opération, longue et coûteuse, donne des résultats inégaux et impressionnistes.

Les thésaurus existants sont, de plus, difficilement actualisables et ne peuvent donc prendre en compte les changements continuels dans le lexique et le thésaurus des textes. Des néologismes et de nouveaux usages apparaissent dans les corpus (Renouf 1993). La coréférence est le domaine où se produisent les changements les plus rapides et les plus marquants, dans la mesure oll les textes traitent des évènements du monde réel, en particulier en ce qui concerne le changement d'identite des gens qui assument des rùles publics.

Notre système pourra, nous l' espérons, améliorer cette situation à plusieurs égards. Notre thésaurus sera une représentation de la réalité du texte et non d'un lexique mental, parcequ'il sera extrait directement de la base de donnée particulière sur laquelle il opère. Ses informations, spécifiques, seront le reflet du domaine concerné et du style de la base de données. Dans la mesure où il est constitué automatiquement, ce thésaurus sera aussi aisément actualisable.

La seconde utilisation des nyms, une fois identifiés, se situe au niveau de la description linguistique. Les nyms constituent des éléments du thesaurus textuel que nous espérons décrire dans les prochaines années.

2.2. Hypothese sous-jacente

L'hypothèse de base d' ACRONYM est que deux mots aux profils collocationnels semblables doivent avoir en commun un sens, un référent ou un usage. Cette position a été développée à partir de l'affirmation de Firth (1957), qui pense que la collocation est une representation du sens. Cette idée est fondamentale dans l' orientation de notre recherche des quinze dernières années. Elle coïncide avec notre expérience que l'automatisation des procédures relatives à la manipulation des textes peut être réalisée directement sur le texte brut, et permet un accès à la sémantique au niveau de la surface.

2.3. Test préliminaire Cl l'hypothèse

En 1992, des tests d'extraction de profils collocationnels simples (sur des textes du journal *Times)* ont confirmé dans une certaine mesure l'hypothèse. Nos observations sont les suivantes (tableaux 1 a 4):

DROPPED + FELL	DROPPED only	FELL only
bombs 36.972	goal 122.474	turnover 24.363
sharply 10.704 points	goals 67.823	wickets 22.652
8.907 dramatically	penalty 60.579	foul 21.464
7.633 fifth 6.489	penalties 40.305	love 15.296
below 6.225	shots 29.033	victim 14.390
per 5.740	tries 20.930	rain 13.878
ball 5.720	conversion 18.045	nikkei 13.724 operating
cent 5.456	catches 14.560	12.644
catch 5.316	conversions 14.485	end-september 12.144
cents 5.233	pears 14.322	snow 11.974
consumption 4.959	shot 14.147	dm11.616
kerry 4.361	try 13.904	fence 11.570
quarter 3.859	mullen 13.538	dow 11.273
figure 3.841	thorburn 13.459	imports 10.916
half 3.607	leaflets 11.331	apart 10.874
third 3.582	davies 10.095	trap 9.761
index 3.540	stephens 9.358	lowest 9.725
profits 3.534	slip 9.130	sterling 9.514
prices 3.468	steele 8.880	dollar 9.335
immediately	evans 8.528	million 9.308
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
average 3.205 shares	planes 7.884	darkness 8.428
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	lloyds 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 second	temperatures 6.147	barrel 7.282
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	lb 5.534	roof 6.567
month 2.160	olazabal5.513	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

Tableau 1. *Collocational Profiles for DROPPED and FELL in* The Times 1991

Le **tableau 1** montre que les mots *dropped* et *fell*, que nous acceptons intuitivement comme étant sémantiquement liés comme synonymes, ont en effet plusieurs cooccurrents communs. Ils semblent se recouper en ce qui concerne le domaine financier. Cependant, comme tous les synonymes, ils ont des differences seémantiques (ou referéntielles), et par consequent des cooccurrents differents. Par exemple, *dropped* paraît plus étroitement associé aux bombes, aux témperatures, aux sports du rugby et du golf, alars que *fell* est lié aux prix du marché financier, aux valeurs de la monnaie et au sport du cricket.

Le **tableau** 2 représente les résultats d'une recherche sur tous les mots qui ont des cooccurrents communs avec le mot *head:* ces résultats montrent que *head,* un mot normalement fortement polysémique et donc pas un bon mot-clé, sélectionne en fait plusieurs nyms qui lui sont synonymes dans son sens de *being in charge* ('dirigeant'), comme *director, chairman, executive, chief.* Il semble qu'il n'y ait pas de nyms attaché au sens de *visage.* Le peu d'ambigu'ite dans l'utilisation du mot *head* est probablement due a la nature du domaine traite, qui joue ici en notre faveur.

Le **tableau** 3 donne une liste impressionante de nyms ayant des cooccurrents communs avec le mot *establish.* Le résultat est intéressant dans la mesure où le rôle de ce mot est de structurer le texte plutôt que de transmettre du sens et, comme tel, il pourrait ne pas avoir de synonymes conventionnels.

61 head	9 firm
33 former	9 japanese
32 director	9 joined
24 has	9 two
22 chairman	8 desk
20 securities	8 european
18 team	8 govett
17 corporate	8jon
16 executive	8 morgan
15 chief	8 sir
15 group	8 years
15 managing	7 arm
14 aged	7 csfb
13 management	7 de
13 sales	7 gills
13 uk	7 investment
12 analyst	7 james
12 its	7 leonard
12 left	7 part
12 research	7 partner
12 senior	6 ash worth
Iljohn	6 broker
11 new	6 capel
11 trading	6 carol
10 company	6 david
10 division	6 department
10 equities	6 financial
10 equity	6 fund
10 international	6 grenfell
10 says	6 house
9 after	6june
9 bank	61ondon
9 benson	6 manager
9 county	6 run
9 diary	6 trust
9 finance	

Tableau 2. *Nyms for HEAD from* The Times *City Page*

establish 497	providing 102	determine 93
create 171	management 102	towards 92
develop151	encourage 102	regiona192
established 150	community 102	powers 92
maintain 147	co-operation 102	legislation 92
ensure 146	protection 101	avoid 92
provide 145	proper 100	standards 91
build 140	keep 100	social 91
establishing 135	agreed 100	creating 91
secure 126	its 99	based 91
protect 126	government 99	agreement 91
introduce 125	funding 99	statutory 90
restore 123	extend 99	operate 90
proposed 122	environmental 99	help 90
improve 122	bring 99	european 90
achieve 117	allow 99	preserve 89
impose 114	require98	including 89
promote III	peace 98	decide 89
control III	international 98	authority 89
prevent 110	authorities 98	any 89
reform 109	apply 98	creation 88
reduce 109	planning 97	review 87
make 109	implement 97	necessary 87
existing 109	governments 97	meet 87
support 108	seek 96	information 87
development 107	companies 96	financial 87
ec 106	required 95	defend 87
retain 105	legal 95	raise 86
un 104	setting 94	pursue 86
economic 104	produce 94	iraq 86
law 103	monetary 94	developing 86
government's 103	separate 93	committee 86
give 103	policy 93	carry 86
securi ty 102		

Tableau 3. *Nymsfor Node ESTABLISH.*

Le **tableau 4** répertorie les sorties des nyms du mot *rowing.* On peut noter qu'il n'y a pas de vrais synonymes (en réalité, on pourrait seulement s'attendre à trouver les presque synonymes *sculling* ou *punting).* Néanmoins, nous avons un groupe sémantique différent ici: des 'co-taxonyms' tels que *athletics, football, golf, cricket, boxing* et *cycling.* On trouve aussi des mots ayant un lien lexical plus Hiche avec le mot *rowing,* comme *Oxford, Cambridge, boat, beat, college, crews.*

85 rowing
36 oxford
32 athletics
32 men's
31 cambridge
31 mike
29 david
~ju~m 21fumili
28 football
27 britain's
27 england's
27 golf
27 hall
27 John
26 cricket
26 former
26 jones
25 aged
25 boxing
25 championship
25 championships
25 chris
25 george
24 alan
24 barry
24 boat
24 cycling
24 evans
24 keiili
24 league
23 kings ton
22 andrew
22 birmingham
22 colin
22 correspondent
22 elliott
22 ian
221eicester
22 manchester
21 beat
21 college
21 hope
21 james
21 kent
21 leading
211ewis
21 michael
20 all an
20 amateur
20 british
20 chess
20 club
20 crews
20 cup
20 davis
20 division
20 dr
20 event
20 gary
20 international
19 brighton
19 bryan
19 cardiff
19 christopher
19 class
19 coach
19 coventry
19 craig
19 crew
19 edinbmgh
19 hill
19 indom
19 irish
19 liverpool
19 mark
19 martin
19 memorial
19 middles ex
19 min
18 australian
18 badminton
18 baili
18 brian
18 bristol
18 captain
18 champion
18 clark
18 commonwealili
18 don
18 england
18 final
18 fletcher
18 gloucester
18 grand
18 having
18 huge
18june
18 1eeds
18 lengilis
18 lightweight 18
london

Tableau 4. *Nymsfor Node ROWING.*

2.4. Les objectifs du projet

L'étude préliminaire nous a encouragés à mettre le projet à exécution. Les objectifs détaillés en sont:

1) identification automatique des nyms;
2) identification automatique des nyms composés;
3) maintenance des nyms dans le temps (dans un flot de textes qUI evolue);
4) application du système à des textes de domaines spécialisés;
5) étude préliminaire pour un application au français;
6) conception d'un système de caractérisation des paires de nyms, assiste par ordinateur.

2.5. Materiel informatique

Le logiciel est développé sous Unix avec 200 mega-octets de mémoire, 2 unites centralés et un disque de 20 gigas-octets. On utilise un modèle clientserveur dans lequel de grosses tâches résident dans le systeme, de sorte que les index, par exemple, sont constarnment en mémoire.

2.6. Les données

Les données textuelles sont foumies par nos partenaires industriels. Nous disposons d' environ 200 millions de mots des joumaux *Financial Times* et *The Independent,* de 1988 á 1994, cinq millions de mots du *McCarthy Business Reports* et d'un acces par satellite a diverses données de la *Press Association.* Toutes sortes d'autres données continuent d'etre accumulées.

Dans la suite, nous parlerons des résultats de la recherche sur la nature des nyms et du thésaurus, tels qu'ils apparaissent après les deux premières étapes de fonctionnement du nouveau système prototype.

3. Etape 1: identification automatique des nyms

La procédure d'identification des nyms est la suivante : pour chaque mot-type d'un corpus de textes donné, considéré comme focus ou nœud, on établit un

profit collocationnel complet de tous les mots qui figurent dans une fenêtre de quatre mots à sa droite et à sa gauche. En fait, certains types de mots sont ignorés ou traités comme des 'mots vides'; ce sont ceux qui sont très fréquents, principalement les mots grammaticaux, mais aussi les items condidérés comme peu utiles en tant que nyms dans la mesure où ils se retrouvent dans l'environnement de la plupart des autres mots. Les profits collocationnels sont enregistrés dans une **base de données collocationnelles**. Cette base contient des informations comme:

- longueur de la fenêtre;
- position dans la fenêtre;
- ordre dans la fenetre;
- lemmatisation;
- ponctuation;
- typographie;
- statistique sur les possibilités combinatoires.

Les profits sont comparés pour déterminer les nœuds qui ont des marques collocationnelles similaires à celles du terme cible, qui peut être fourni par l'utilisateur. Ces nœuds, ou 'candidats nyms' du terme cible, sont ensuite classes en fonction de divers critères, selon leur indice de similarité, basé sur une mesure de fréquence relative; c'est-à-dire le rapport entre occurrence observée et occurrence attendue, tenant compte de la fréquence du nœud dans la totalité du corpus et de son comportement comme collocation du mot cible.

3.1. Premiers resultats

3.1.1. Recherche par un seul mot-de

Le premier objectif du projet est l'etablissement de paires de nyms. En termes de technologie de l'information, cela veut dire que l'utilisateur d'une base de donnees soumettra un terme unique pour la recherche et que notre systeme fournira une serie de mots alternatifs.

Le **tableau 5** montre les sorties concernant les nyms du mot *director* dans un corpus de textes economiques tires du journal *The Independent.* La liste est ordonnee: le tri a ete effectue a partir des donnees chiffrees de la colonne de droite. La question est: que peut-on esperer trouver comme informations dans le thesaurus, comme mots-cles alternatifs, pour un mot donne ? En fait, on ne trouve pas id, dans les textes economiques, de vrais synonymes. Nous dispo

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

Tableau 5. *Nyms for Node DIRECTOR.*

sons des mots *chief* et *head,* qui pourraient être des synonymes proches, et aussi des mots *chairman* et *director-general,* qui sont en contraste avec *director.* On trouve aussi plusieurs noms propres, qui ne sont pas désambiguïsés parceque, à ce stade, les mots composés ne sont pas reconnus, sauf s' ils ont été prealablement marques. Si ya avait ete le cas, *chief executive* serait aussi un nym.

Chose interéssante, on peut creer une serie d'hyponymes en ajoutant au mot dominant *director* des modifieurs adjectivaux. Ici *managing, finance, non-executive, deputy* et *marketing.* Ce type de composition n'est pas surprenant, dans le cas de *director,* parceque les éléments sont, en un sens, métonymiques. Il s'agit de certains aspects du concept de *director.* Mais on a aussi observé cela avec d'autres mots. Par exemple le mot *bid* ales hyponymes *hostile bid, take-over bid, union-led bid.* Nous considérons ces combinaisons

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

Tableau 6. *Nyms for Node LUXURY.*

de termes hyponymiques ou métonymiques comme une des caractéristiques du thésaurus textuel. Elles prévalent peut-être dans les textes économiques, ou les noms dominants sont souvent des nominalisations de verbes. Nous étudierons ce phénomène dans une étape ultérieure.

Le **tableau** 6 présente la sortie des nyms du mot *luxury,* extraits du *Financial Times.* Cette forme de base semble être préféréé à la forme dérivée *luxurious* comme modifieur de noms. I1 s' agit probablement d 'une question de

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

Tableau 7. *Nyms of LUXURY and HOTEL*

choix stylistique: une nuance différente est suggérée par la forme de base, ce qui est probablement assez courant. Ceci n'apparaît cependant pas dans un thésaurus traditionnel.

I1 ressort de notre liste d'items que *luxury* est ambigu, ou plutôt multi-référentiel. On trouve une variété de nyms qui, dans des contextes différents, sont associés à la notion de luxe et/ou pourrait remplacer le mot *luxury* dans le texte. Par exemple *f-type, jaguar, lexus, infiniti, xj, s-class, bmw.* Dans le contexte des hôtels, nous avons *Steigenberger, Periquito, Gleneagles, fivestar, posh.* Une *datcha* est une maison de luxe en Russie, un *pullman* est un train de luxe, *Hennessy-Louis* un cognac de luxe. Enfin, et c' est rassurant, nous avons les formes flechies du mot : *luxuries, luxurious.*

3.1.2. Recherche par deux mots-cles

La recherche la plus typique dans une base de données se fait sur plus d'un seul terme. Nous avons donc étendu le système à l'identification des nyms de deux mots-c1és. Ceci a favorisé notre approche, dans la mesure où les deux termes se désambiguent mutuellement, en servant de contexte l'un à l'autre, ce qui permet à nos sorties d' être mieux focalisées sur les aspects pertinents des mots, précisés par le contexte commun.

Le **tableau** 7 donne une liste de nyms beaucoup plus limitée que la précédente, correspondant aux termes-c1es *luxury* et *hotel.* On y trouve des

quasi-synonymes du mot *luxury* dans le contexte *hotel: five-star, posh, luxurious, plush* et des quasi-antonymes: *three-star, no-frills.* Nous avons aussi une taxinomie des hôtels: *Intercontinental, Steigenberger, Savoy, Mediterranee, Sofitel, Marriott, Grosvenor, Hyatt, Ritz.*

Cette sortie est interessante du point de vue de la technologie de l'information, car elle montre le rôle crucial des noms propres dans le thésaurus textuel. Sur le plan linguistique, le nombre réduit des synonymes et des antonymes attire notre attention sur le fait que le thésaurus textuel, tel que réalisé dans n'importe quel domain specialise, est limite et n'utilise pas toutes les possibilités existantes du lexique mental. Les synonymes et antonymes reellement utilises sont bons, mais egalement inattendus.

Jusqu'ici, dans notre étude, les noms propres sont vus comme un trait distinctif central du thésaurus textuel. L'image qui s'en dégage est celle d'une hiérarchie sémantique assez plate: un terme dominant, *hotel,* et en dessous, très souvent, une simple liste de noms propres.

3.1.3. Les noms propres comme mats-des

Les utilisateurs de bases de données présentent souvent des noms propres comme terme de recherche.

Le **tableau** 8 donne les nyms pour le mot *conference* combiné avec le nom propre *Marriott.* L'hôtel Marriott est en effet le siège de nombreux colloques,

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

Tableau 8. *Nyms for MARRIOIT and CONFERENCE*

l'idée etant qu'un utilisateur pourrait vouloir trouver des textes supplémentaires ayant trait aux colloques et aux hôtels où ils ont lieu. Les nyms sont répartis en trois groupes référentiels distincts: les noms de colloques, comme *Siggraph, Interforest, Helfex, Supercomputing,* d'autres lieux de colloques: *QUll, Plaisterers, Wallop, Accra, G-Mex,* et des hôtels: *Novotel, Gleneagles, Intercontinental, Ramada.* Cette sortie est utile du point de vue de la technologie de l'information, dans la mesure ou plusieurs de ces nyms sont des mots-clés alternatifs raisonnables.

3.1.4. Identification automatique de la polysémie

La demière tache, dans cette première étape, a été de trouver un systeme automatique d'identification de la polysémie. Ce travail non trivial doit assurer une correspondance plus étroite entre le mot-clé et les listes de nyms. Si les différents sens (ou références, ou usages) du mot-clé peuvent être etablis, alors l'utilisateur peut decider de celui qu'il avait en tête en sélectionnant ce terme. Evidemment, bien que tous les mots ne soient pas polysémiques en euxmêmes, chaque mot peut en fait être multi-contextuel et multi-référentiel dans le texte. La metaphore est la cause majeure de cette situation : elle retire le terme de son contexte normal pour l'insérer de façon incongrue dans un autre contexte, afin de produire un effet stylistique.

Là encore, c'est la collocation que nous utilisons pour identifier la polysémie. A partir de la liste des cooccurrents pour un mot donné, nous avons examiné les profits collocationnels de chacun et nous les avons rnis en groupes sur la base de leur degré de ressemblance. Un logiciel du domaine publique
effectuant ce type de regroupements, appelé *Pam,* a été adapté pour cette tâche.

Le **tableau** 9 montre les groupes créés à partir du profit collocationnel du mot *air.* On peut les rattacher aux sens suivants:

air:
1) atmosphére;
2) domaine d'action militaire;
3) élément associé au transport aérien;
4) un des nombreux moyens de transport civil;
5) terme en relation avec le confIit dans l'ex- Y ougoslavie.

11 s'agit encore ici d'un résultat préliminaire, le système étant encore à l'étude, mais c'est une grosse satisfaction de constater que sur cette simple base, nous pouvons distinguer non pas deux sens d'un mot (ce qui serait déjà un résultat),

- Run 4 Cluster I -
air 0.19
water 0.06
hot 0.03
pollution 0.02
breath 0.02
warm 0.01
breathe 0.01
conditioners 0.01
filters 0.01
compressed 0.01

- Run 4 Cluster 3 -
aircraft 0.41
jets 0.20
fighter 0.06
charter 0.02
carriers 0.01

- Run 4 Cluster 5 -
Bosnian 0.73
Bosnia 0.58
Serb 0.56
Serbs 0.55
embargo 0.29
forces 0.25

- Run 4 Cluster 4 -
rail 0.52
freight 0.36
road 0.26
express 0.21
services 0.13
links 0.11
passengers 0.04
strike 0.02

- Run 4 Cluster 2 missile 0.44
- Iraq 0.11
surface 0.08
Iraqi 0.06
attack 0.06
launched 0.04

Tableau 9. *AIR:* 5 *Collocate Clusters*

Run 4 Cluster I
slap 0.03
cliff 0.03
sampling 0.03
flies 0.0 1
quota 0.01

_ Run4 Cluster 2 -
severe 0.56
disruption 0.12
losses 0.11
loss 0.04
consequences 0.04
difficulties 0.04
criticism 0.03
problems 0.02

- Run 4 Cluster 3 -
criminal 0.77
charges 0.68
trial 0.50
prosecution 0.48
disciplinary 0.40

Run 4 Cluster 4
daunting 0.66
task 0.29
challenge 0.20
prospect 0.20
must 0.10

- Run 4 Cluster 5 -
eyes 0.81
hair 0.37
tears 0.37
nose 0.36
staring 0.26

Tableau 10. *FACE:* 5 *Collocate Clusters*

mais plusieurs sens. Les groupes sont plus ou moins reconnaissables en termes sémantiques traditionnels, mais il faut comprendre que dans certains cas, comme pour la catégorie (5) ci-dessus, il s'agit d'un ensemble de références qui ne correspond pas à un sens discret (isolable).

Un autre obstacle à la démarche pour donner une représentation du sens' des mots réside dans le fait que les mots les plus communs de la langue jouen un rôle majeur dans la phraseologie, sans contribuer au contenu propositionne des séquences (Renouf 1992,1993). Des expressions comme *on the face ofi!.' at the end of the day,* sont des paraphrases d'items lexicaux comme1 *superficially, ultimately,* et les mots qui composent ces expressions ne fonc tionnent pas comme des signifiants indepéndants (pour reprendre la terminolo gie de Saussure 1931), c'est-à-dire qu'ils ne réfèrent pas a des objets ou de concepts.

Malgré cela, même pour un mot fréquent comme *face,* les résultats d notre analyse, donnes dans le **tableau 10**, sont prometteurs.

Les categories sont ici les suivantes :

face:

1) compos ant de plusieurs expressions comme: *slap in the face, the cliff face, fly in the face of,*
2) 'être confronté à', dans le contexte de choses négatives comme *loss, consequences, criticisms, problems;*
3) 'être confronte à', dans un contexte juridique;
4) 'avoir comme possibilité', dans le cadre d'un défi, mais pas nécessaire ment dans un contexte negatif;
5) 'partie de la tête'.

4. Etape 2 : identification automatique des noms composés

On sait que les unités lexicales sont sou vent constituées de plus d'un mot simple. Il a donc été nécessaire, pour améliorer notre système, de trouver un moyen d'identifier automatiquement les items composes dans un texte. Le but serait qu' on puisse entrer *John Major* comme terme de recherche unique et que

le système puisse le faire correspondre, par exemple, à *Prime Minister*. Pour cette étape du projet, nous ne sommes qu' au tout début du travail, qui porte d'abord sur les séquences de deux mots.

En mettant en correspondance les profits collocationnels des mots, nous avons pu isoler ceux qui partagent les mêmes cooccurrents et apparaissent ensemble de façon significative.

Dans le **tableau 11,** les paires de mots concernent respectivement les entrees *Slair, Heseltine* et *Euro.* Il s'agit dans tous les cas de véritables mots composés, sauf *Blair Labour* et *Heseltine Secretary.* Ces accidents sont dus à l'absence de ponctuation, dans le journal, entre un nom propre et une apposition: *Tony Blair Labour Party Spokesman of Employment* et *Michael Heseltine Secretary of State for Defence/the Environment.* Certaines combinaisons paraissent étranges parcequ'elles font partie en fait d'une unité multimots plus large. *Tony Euro,* par exemple, appartient a la sequence *Tony Euro Rebel/Sceptic*. Globalement, pourtant, ce premier résultat semble encourageant.

Le **tableau 12** donne la liste des mots qui font paire avec l~ mot *party,* et la aussi là résultat est prometteur: *advance party, garden party, party line, party dress, dinner party, guilty party, interested party, working party* sont tous de bons candidats. Ces sorties extensives po sent en fait le problème de

BLAIR	HESEL TINE	EURO
Neil Mr	Mr	Atlantic
Les	Secretary	Brokers
Hunter	Michael	sceptic
RN	plan	Mix
Labour	Politics	Disney
Barley		sceptics
Mrs		Africa
Tony		federalism
Colonel		Disneyland
Stewart		enthusiasts
bt		Tory
		bank
		elections
		MP
		MPs

Tableau 11. *Ranked Lists of Words Pairing with the Nodes BLAIR, HESELTINE and EURO.*

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

Tableau 12. *Ranked List of Words Pairing with the Node PARTY*

savoir quelles sont les éléments composés que I' on veut garder et ceux que l'on desire rejeter. Et au-dela, qu'est-ce qu'un vrai compose? Doit-il être métaphorique ? Doit-il avoir un sens qui est plus que la somme de ses parties? Etc. Cet aspect a été souvent étudié en détail et toute une série de critères a été proposee pour definir la classe des noms composés. Nous devrons préciser ces notions dans la suite du travail sur la description des thésaurus textuels. Dans l'optique de la technologie de l'information, cependant, il suffit probablement d'adopter une définition assez floue, dans la me sure où l'utilisateur va soumettre des composes de toutes sortes, avec des degrés de figement variés.

5. **Conclusion**

Nos recherches en cours sur les thésaurus de textes, en particulier économiques, continuent de nous faire découvrir des faits nouveaux. Les observations suivantes donnent une idée des constatations faites jusqu'ici.

L'éventail des realisations lexicales trouvées dans les textes pour chaque type de relation sémantique est plus large et plus varié que celui fourni par les thésaurus traditionnels, en partie à cause du fait qu'il dépasse la notion de frontière de mot. Dans le cas d'un texte ou d'un type de texte concernant un

domaine particulier , cependant, l'éventail des choix lexicaux sera spécifique et plus restreint.

En liaison avec cette question, nous avons observé que, dans chaque texte le système par lequelles objets ou les concepts sont mis en relation, à différents niveaux de généralite, est plutôt plat, comme pour *poodle, dog, animal, mammal.* Jusqu'ici, dans les données que nous avons traitées, il se reduit simplement a deux niveaux, comme dans *hotel: Sheraton.*

L'hyponymie est donc une relation nymique dominante dans nos données. Ceci se manifeste en partie par la prépondérance des noms composés, points ultimes de la hiérarchie dans un système de reference. Les noms propres apparaissent en fait comme étant au centre des relations nymiques dans les textes joumalistiques. Etant donné ce caractère crucial, les noms propres demandent une attention beaucoup plus grande que celle qui leur a été accordée jusqu'ici dans les descriptions linguistiques.

Les nyms sont souvent, nous l'avons dit, des mots composés. C'est-à-dire que les relations nymiques et collocationnelles ne peuvent être décrites indépendamment les unes des autres. De nombreux nyms multi-mots sont des hyponymes formés par adjonction d'un modifieur sur un hyperonyme, comme par exemple *managing director* ou *take-over bid.*

NOTE

* Cet article a fait l'objet d'une communication au 14eme Colloque sur la Grammaire et le Lexique Compares des Langues Romanes (Tel Aviv, sept. 1995). Je remercie David Tiomajou et Christian Leclere de I' aide qu' ils m' ont apportee pour la traduction fran~aise. Une version en anglais figure dans *Synchronic corpus linguistics. Papers from the sixteenth International Conference on English Language Research on Computerized Corpora (ICAME* 16), C.E. Percy, C.F. Mayer and I. Lancashire (eds), Amsterdam: Rodopi, 1996.

REFERENCES

Firth, J.R. 1957. *Papers in Linguistics.* 1934-1951. London: OUP.

HarperCollins (ed.). 1995. *Collins Thesaurus: The Ultimate Word Finder*. London: HarperCoJlins.

Renouf, AJ. 1992. What Do You Think of That: A Pilot Study of the Phraseology of the Core Words of English, in *New Directions in English Language Corpora*, G. Leitner (ed.), Berlin: Monton de Gruyters, pp.30l-317.

Renouf, AJ. 1993. What the Linguist has to say to the Information Scientist, *The Journal of Document and Text Retrieval* vol.l/2, pp. 173-190.

Saussure, Ferdinand de. 1931. *Cours de Linguistique Generale,* Paris: Payot.

SUMMARY