Extraction de collocations fondée sur des méthodes statistiques

De la concordance aux collocations

Benoit.Robichaud@UMontreal.ca

Plan de la présentation

- Les collocations : *lesquelles, pour qui, où, comment ?*
- Les concordances : *interrogations, contextes et kwic*
- Calculs statistiques des collocations :
  - Fréquence de bigrammes
  - Distance et variance moyennes (les bigrammes distants)
  - Tests d'hypothèses et d'indépendance :
    - Test *t*
    - Test du *khi-carré*
    - Proportion de vraisemblance
    - Calcul de l'information mutuelle
- Conclusions

Les collocations

- C'est quoi ? :
  - Une suite de mots (plus ou moins) consécutifs présentant un caractère *idiosyncrasique* (c.-à-d. dont le sens global ne peut simplement être induit par le sens des mots qui la constituent).

    'café noir'

    'prendre une décision'

    'to tell off'

    (exceptionnellement les termes complexes : 'hydrolic oil filter')

- Tests traditionnels pour le repérage et l'évaluation de collocations :
  - Leur premier sens n'est pas compositionnel
  - Leurs composantes ne sont pas substituables
  - Elles sont difficilement modifiables

Les collocations

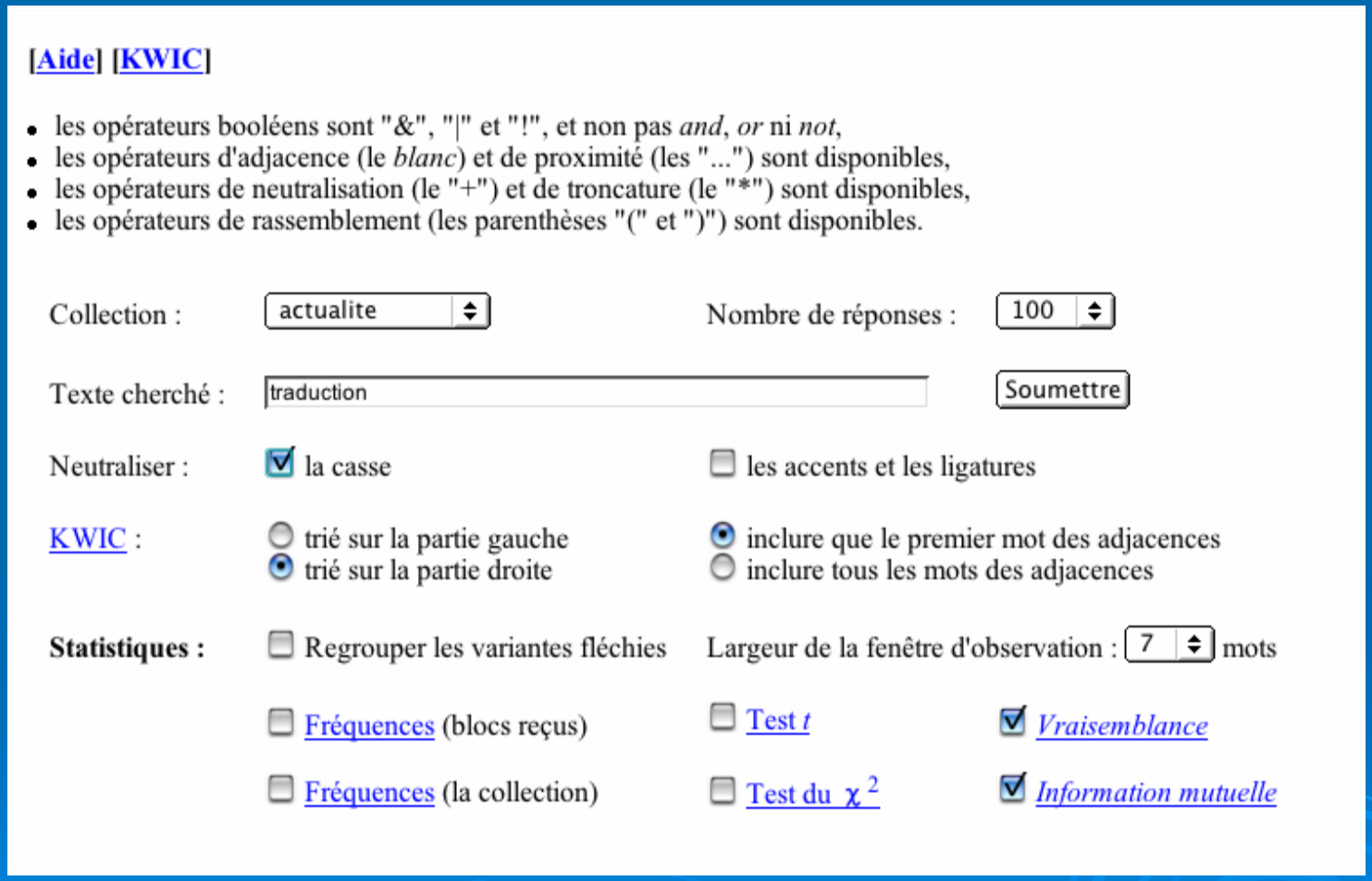
- Qui s'y intéresse ?
  - Lexicologie et lexicographie
    - Trouver les collocations les plus importantes à mettre dans les entrées de dictionnaire
  - Terminologie et traduction
    - Repérage des groupes terminologiques dans les textes de sous-domaines de spécialité
    - Vérification de la constance d'emploi des termes
  - Linguistique de corpus
    - études sociolinguistiques : *Firth, Halliday, Sinclair* …
  - Linguistique informatique
    - applications logicielles

Benoit.Robichaud@UMontreal.ca

Repérage des collocations

- Contextes de repérage des collocations :
  - Extraire d'un texte particulier les suites de mots les plus susceptibles d'être des *collocations*
    - conjointement avec d'autres méthodes de repérage
  - Interroger une banque de textes (une *collection*) à l'aide d'un *concordancier* pour extraire des contextes d'utilisation
    - repérer les mots cooccurrents des mots cherchés pour et induire les *collocations* les plus probables

Le concordancier : *interrogations*



Le concordancier : *segments trouvés*

Blocs reçus

Bloc # 15350:

Mais chaque matin, devant son ordinateur Toshiba portatif, Michel Tremblay préparait sa rentrée. il ruminait son prochain livre, terminait une **traduction**, révisait les dialogues d'un film pour Diane Dufresne ou s'inquiétait de l'affiche de sa dernière pièce, Marcel poursuivi par les chiens.

Bloc # 20246:

En voici une **traduction** non officielle, que nous a fait parvenir le professeur Jacques Dunnigan, de l'Université de Sherbrooke.

Bloc # 21901:

En fait, il est plus présent dans son île natale, Funem, et sa ville Odense, qui a consacré un musée à l'auteur le plus connu hors du Danemark avec le père de l'existentialisme Kierkegaard et l'écrivain Karen Blixen (La Femme africaine, **traduction** française de Out of Africa, Sept Contes gothiques, etc.).

Bloc # 22450:

Ergo, delenda sunt cégeps, comme on le dit, en **traduction** française bien entendu, dans certains milieux.

Bloc # 33860:

Le pauvre homme manifestement a envie de suivre le spectacle au lieu d'être forcé de faire de la **traduction** simultanée à un visiteur ignorant.

Le concordancier : ‘kwic’

196388	emplacer le télejournal de Bernard Derome par une	**traduction**	de celui de Peter Mansbridge.
222988	e (17 h), Quatre Saisons diffuse Flash Modes, une	**traduction**	de FashionTV, produite par la station torontoise
61995	qui tenait le haut du palmarès en 1964, était une	**traduction**	de Hold Me Tight des Beatles.
331892	(La	**traduction**	de l'anglais au français est celle de La Presse.)
265418	On l'occupe à la	**traduction**	de la correspondance et à de menus travaux de com
155440		**Traduction**	de Marguerite Yourcenar, mise en scène de Martine
58077	ar Payot et réédité par Folio, dans une brillante	**traduction**	de Marie-Odile Fortier-Masek.
418423	Traces d'étoiles de Cindy Lou Johnson,	**traduction**	de Maryse Warda, mise en scène de Pierre Bernard,
300592	a première fois en français à Montréal, selon une	**traduction**	de Michel Tremblay, La Descente d'Orphée est la p
265990	Voici la	**traduction**	de quelques-uns d'entre eux: Vas-y, Jésus, bottes
360566	inconnue du public francophone jusqu'à la récente	**traduction**	de son roman Niagara.
109059	e dernier livre de Bob Woodward, Chefs de guerre,	**traduction**	de The Commanders, édité en langue française par
207558	Ce document est une	**traduction**	de votre permis de conduire en plusieurs langues:
58119	des romans canadiensanglais qui nous arrivent en	**traduction**	depuis quelque temps sont signés par des gens don
58055	la publication des Dainty Monsters, il lisait en	**traduction**	des auteurs de langue française : Jacques Godbout
273279	promis de l'OPEP, cet accord ne vaudra que par la	**traduction**	des intentions dans les faits», estime Peter Bogi
407608	econdaire 4 et 5, d'assumer au coût de 125000$ la	**traduction**	des manuels de mathématiques.
356031	Il s'agit d'une	**traduction**	du guide anglais «Disney World and Beyond» publié
263354	pédagogique original, qui soit autre chose qu'une	**traduction**	du matériel francophone.
116765	ser le changement politique, comprend-on selon la	**traduction**	du texte écrit en arabe.
103028	t, comportant une nouvelle instrumentation et une	**traduction**	en allemand du texte anglais.
423064	La	**traduction**	en français des paroles de la chanson ne se rappo
169317	cet enroulement sur le message du gène empêche sa	**traduction**	en produit (protéine) et fait que la quantité du
265170	a Christie, que Le Masque édite dans une nouvelle	**traduction**	enfin complète et dotée d'un appareil critique ju
226477	si ce dernier titre à cause de sa parenté avec la	**traduction**	espagnole d'un western d'Anthony Mann, Tambores L
110968	La	**traduction**	est de La Presse.
34743	La	**traduction**	est, me semble-t-il, convenable;
213403	vernement n'est pas capable de t'en donner (de la	**traduction**	et des documents en français), fais-y des représe
159378	ollars aux contribuables en commissions, frais de	**traduction**	et procédures judiciaires, n'en déplaise aux fran
22450	Ergo, delenda sunt cégeps, comme on le dit, en	**traduction**	française bien entendu, dans certains milieux.
38502	comprendre qu'il s'agissait tout simplement de la	**traduction**	française de l'expression « french studies » !
21901	d et l'écrivain Karen Blixen (La Femme africaine,	**traduction**	française de Out of Africa, Sept Contes gothiques
226476		**Traduction**	française de Tacones Lejanos, Talons aiguilles (o
34770	Sinon, attendez un peu : la	**traduction**	française de Voice-Over est en cours.
121305	dio-Canada compte offrir aux abonnés du câble une	**traduction**	française des émissions d'information télévisées
58020	La	**traduction**	française du roman, L'Homme flambé (L'Olivier), e
196387	gmenter leurs tarifs d'abonnement, et le canal en	**traduction**	française leur donnerait un argument à plaider de
129714	ondres, rédigea en anglais ses mémoires, dont une	**traduction**	française ne fut établie qu'en 1971.
64791	lui consacrait une importante biographie dont la	**traduction**	française s'est fait attendre jusqu'à cette année
172147	Avec la	**traduction**	française, côte à côte, s'il y a lieu.
296775	nfin, dans le meilleur de ce qui la constitue, en	**traduction**	française.

Benoit.Robichaud@UMontreal.ca

Calculs des collocations : *fréquences*

- Comment :
  - repérer tous les mots adjacents au mot cherché,
  - compiler pour chacun leur fréquence,
  - les présenter en les triant sur les fréquences.
- Si on étiquette les échantillons de textes
  - On peut filtrer les mots grammaticaux
  - On peut neutraliser les variantes flexionnelles
- Problèmes :
  - On ne prend pas en compte les mots qui ne sont pas directement cooccurrents des mots cherchés
  - On ne peut trier entre eux les mots qui ont la même fréquence
  - On ne normalise pas la fréquence des couples formés de mots dont la fréquence est elle-même basse

fréquences	*mots*
176	française
144	anglaise
085	littéraire
027	libre
013	technique
004	automatique
	...

Calculs des collocations : moyenne et variance

- Comment :
  - On définit une fenêtre d'observation fixe d'une dizaine de mots

-4	-3	-2	-1	*traduction*	+1	+2	+3	+4

  - On repère les cooccurrents dans cette fenêtre
  - En plus de la fréquence, on prend en compte cette fois la distance moyenne et la variance moyenne de cette distance (ou l'*écart type moyen*) pour chacun des cooccurrents par rapport au mot cherché
  - On fait le tri en fonction de la variance
    - Lorsque la distance est fixe, la variance est nulle ('forte collocation')
    - Lorsque la variance est faible, on risque aussi d'être en présence d'une collocation

Calculs des collocations : moyenne et variance

- *Distance moyenne* : $\overline{d} = \frac{\sum_{i=1}^{n}}{n}$
- *Variance* : $s^2 = \frac{\sum_{i=1}^{n}(di - \overline{d})^2}{n-1}$
- *Écart type* : $s = \sqrt{s^2}$

Benoit.Robichaud@UMontreal.ca

Calculs des collocations : moyenne et variance

- Un exemple (tiré de Manning & Schutze (1999), chap 5) :

variances	*distances*	*fréquences*	*mot 1*	*mot 2*
1.13	2.57	7	powerful	organization
1.07	1.45	80	strong	support
1.01	2.00	112	Richard	Nixon
...	...	...	...	...
0.49	3.87	131	hundreds	dollars
...	...	...	...	...

Tests d'hypothèses

- Problèmes du calcul des fréquences, des distances et des variances :
  - Une haute fréquence et/ou une basse variance peuvent être accidentelles (par exemple, lorsque les membres d'une collocation sont indépendamment fréquents)
  - On ne peut trier les basses fréquences semblables entre elles
- Hypothèse nulle (*H0*) :
  - Il n'y a pas d'association collocative entre deux mots si la probabilité de voir les cooccurrents ensembles est égale (ou inférieure) au produit de leur probabilités respectives (avec une marge de confiance < 0.005)
  - indépendance *:* $P(m1m2) = P(m1) \times P(m2)$

Test *t*

- Permet d'estimer si une collocation donnée est probable ou non en se basant sur la moyenne et la variance d'un échantillon de données par rapport une *hypothèse nulle* basée sur une distribution normale :

$$t = \frac{P(m1m2) - P(m1) \times P(m2)}{\sqrt{P(m1m2)/N}}$$

t(m1m2)	*f (m1 m2)*	*f (m1)*	*f (m2)*	*m1*	*m2*
4.4854	20	42	20	Ayatollah	Ruhollah
4.4721	20	41	27	Bette	Midler
4.4332	20	30	117	Agatha	Christie
2.3714	20	14907	9017	first	made
1.2176	20	14093	14776	like	people

Test *t*

- Un exemple pour ‘nouvelle’ et ‘traduction’ :

  P(nouvelle) = 15828 / 14307668

  P(traduction) = 4675 / 14307668

  Hypothèse nulle = (15828 / 14307668) × (4675 / 14307668)
  = 3.615×10^{-7}

  Calcul du test *t* = 0.999932

- Le résultat est plus petit que la marge de confiance pour le test *t* (2.576)
- On ne peut rejeter l’hypothèse nulle : le couple ‘nouvelle traduction’ est donc compositionnel et n’est pas une *collocation*
- Dans le contexte du tri des différents cooccurrents, la marge de confiance elle-même n’est pas très importante
- Problème : on fait l’hypothèse que les probabilités sont normalement distribuées (Church et Mercer, 1989)

Test du *khi-carré*

- On calcule une table de contingence :

	m1 = *nouvelle*	m1 ≠ *nouvelle*
m2 = *traduction*	8 (*nouvelle traduction*)	4667 (ex : *ancienne traduction*)
m2 ≠ *traduction*	15820 (ex: *nouvelle vie*)	14287181 (ex: *ancienne vie*)

- On compare les fréquences observées de la table de contingence avec les fréquences estimées pour notre hypothèse nulle :

$$X^2 = \sum_{i,j} \frac{(Oij - Eij)}{Eij}$$

$$X^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}$$

Benoit.Robichaud@UMontreal.ca

Test du *khi-carré*

- Notre exemple :

  Hypothèse nulle = N × (8 + 4667) / N × (8 + 15820) / N ≈ 5.2

  Calcul du *khi-carré* ≈ 1.55

- Le résultat est plus petit que la marge de confiance pour le test du *khi-carré* (3.841)

- On ne peut rejeter l'hypothèse nulle : le couple 'nouvelle traduction' est donc compositionnel et n'est pas un bon candidat comme *collocation*

- Problème : on obtient une mauvaise évaluation lorsque :
  - l'échantillon des données est trop petit (< 20)
  - les fréquences observées sont trop basses (< 5) avec un échantillon trop restreint (< 40)

Calcul du coefficient de vraisemblance (en anglais *log-likelihood ratio*)

- Plus approprié que le test du *khi-carré* pcq le résultat est moins biaisé et plus facilement 'interprétable' avec les basses fréquences
- Fonctionne bien aussi pour les données peu fréquentes
- Deux hypothèses de base :

  H1 (indépendance) : $P(m2 \mid m1) = p = P(m2 \mid \neg\, m1)$

  H2 (dépendance) : $P(m2 \mid m1) = p1 \neq p2 = P(m2 \mid \neg\, m1)$

  Lorsque l'occurrence d'un mot est (ou non) indépendante de l'occurrence d'un autre mot

  On s'attend à p1 > p2 lorsque l'hypothèse *H2* est vraie

Benoit.Robichaud@UMontreal.ca

Calcul du coefficient de vraisemblance (en anglais *log-likelihood ratio*)

- On calcule les valeurs pour *p*, *p1* et *p2* :

$$p = \frac{f(m1)}{N} \qquad p1 = \frac{f(m1m2)}{f(m1)} \qquad p2 = \frac{f(m2) - f(m1m2)}{N - f(m1)}$$

- En assumant une distribution binômiale : $b(k;n,p) = \binom{n}{k} p^k (1-p)^{(n-k)}$
- On calcule le coefficient de vraisemblance :

$$\log \lambda = \log \frac{L(H_1)}{L(H_2)}$$

$$\log \lambda = \log \frac{b(f(m1m2); f(m1), p) \times b(f(m2) - f(m1m2); N - f(m1), p)}{b(f(m1m2); f(m1), p1) \times b(f(m2) - f(m1m2); N - f(m1), p2)}$$

Pour comparer avec une marge de confiance du *khi-carré* : $-2 \log \lambda$

Benoit.Robichaud@UMontreal.ca

Tri en fonction du coefficient de vraisemblance

AdjQ	6.455	simultanée	[33860, 62602, 181734, 213395]
AdjQ	5.663	française	[21901, 22450, 34770, 38502, 58020, 64791, 121305,
AdjQ	5.536	anglais	[356031]
AdjQ	5.061	anglaise	[105157, 269988, 320920]
AdjQ	4.640	littérale	[45206, 47648, 295611]
AdjQ	4.355	libre	[81413, 109500, 201763, 296284, 337639, 395942]
AdjQ	2.903	mauvaise	[233599]
AdjQ	2.748	francophone	[263354]
AdjQ	2.230	officielle	[20246, 215170]
AdjQ	2.132	célèbre	[109500]
AdjQ	1.947	people	[233599]
AdjQ	1.676	redoutables	[64623]
AdjQ	1.578	public	[125940]
AdjQ	1.100	fastidieux	[140864]
AdjQ	0.966	africaine	[21901]
AdjQ	0.820	multicolores	[395942]
AdjQ	0.698	européennes	[373065]
AdjQ	0.670	récente	[360566]
AdjQ	0.530	rares	[47648]
AdjQ	0.390	nouvelle	[265170]
AdjQ	0.384	postale	[109500]
AdjQ	0.031	brillante	[58077]
AdjQ	-0.176	chrétienne	[233884]
AdjQ	-0.301	grand	[54063]
AdjQ	-0.422	espagnole	[226477]
AdjQ	-0.648	coûteux	[215170]
AdjQ	-0.659	réaliste	[85640]
AdjQ	-0.793	judiciaires	[159378]
AdjQ	-1.768	de guerre	[109059]
AdjQ	-2.671	complète	[265170]
AdjQ	-3.158	bonne	[51742]
AdjQ	-4.328	capital	[373065]

Benoit.Robichaud@UMontreal.ca

Calcul du coefficient d'*information mutuelle*

- Évalue plus justement la force du rapport entre les fréquences relatives d'occurrence de deux mots dans une même fenêtre d'observation par rapport à leur fréquence d'occurrence respective

$$I(m1m2) = \log 2 \frac{P(m1m2)}{P(m1) \times P(m2)}$$

- *H1* (indépendance parfaite) :

$$I(m1m2) = \log \frac{P(m1m2)}{P(m1) \times P(m2)} = \log \frac{P(m1) \times P(m2)}{P(m1) \times P(m2)} = \log 1 = 0$$

- *H2* (dépendance parfaite) :

$$I(m1m2) = \log \frac{P(m1m2)}{P(m1) \times P(m2)} = \log \frac{P(m1)}{P(m1) \times P(m2)} = \log \frac{1}{P(m2)}$$

Tri selon le coefficient d'*information mutuelle*

AdjQ	7.620	littérale	[45206, 47648, 295611]
AdjQ	7.602	simultanée	[33860, 62602, 181734, 213395]
AdjQ	4.243	anglaise	[105157, 269988, 320920]
AdjQ	3.978	fastidieux	[140864]
AdjQ	3.750	française	[21901, 22450, 34770, 38502, 58020, 64791, 121305, 129714,
AdjQ	3.642	redoutables	[64623]
AdjQ	3.511	anglais	[356031]
AdjQ	3.294	people	[233599]
AdjQ	3.227	multicolores	[395942]
AdjQ	2.821	libre	[81413, 109500, 201763, 296284, 337639, 395942]
AdjQ	2.472	francophone	[263354]
AdjQ	2.423	mauvaise	[233599]
AdjQ	2.373	postale	[109500]
AdjQ	2.202	africaine	[21901]
AdjQ	2.027	officielle	[20246, 215170]
AdjQ	1.900	célèbre	[109500]
AdjQ	1.878	brillante	[58077]
AdjQ	1.845	européennes	[373065]
AdjQ	1.645	chrétienne	[233884]
AdjQ	1.411	espagnole	[226477]
AdjQ	1.404	récente	[360566]
AdjQ	1.289	rares	[47648]
AdjQ	1.230	coûteux	[215170]
AdjQ	1.222	réaliste	[85640]
AdjQ	1.128	judiciaires	[159378]
AdjQ	0.957	public	[125940]
AdjQ	0.395	complète	[265170]
AdjQ	0.333	grand	[54063]
AdjQ	0.304	de guerre	[109059]
AdjQ	0.166	capital	[373065]
AdjQ	0.116	bonne	[51742]
AdjQ	-0.408	nouvelle	[265170]

Calcul de l'*information mutuelle*

- Problème :
  - Les couples avec une plus petite fréquence ont un résultat proportionnellement plus fort que les couples ayant une haute fréquence

    Opposé de ce qu'on voudrait puisque les hautes fréquences d'occurrence sont un bon indice de base pour le repérage des collocations

- Solution 1 : Ne pas tenir compte des couples dont la fréquence est trop basse (inférieure à 3)

- Solution 2 : Compenser le biais en multipliant le résultat par la fréquence d'occurrence conjointe : $f(m1m2) \times I(m1m2)$

Conclusions

- Plusieurs raisons de tirer profit du repérage et du classement statistique des *collocations* :
  - Lexicographie et terminographie
  - Recherche d'information et classement de document
  - Génération de texte et traduction automatique
  - …

Références

- Bélisle & Desrosiers (1984) *Introduction à la statististique*. Québec : Morin
- Benson, M. (1989) « The structure of the collocation dictionary », In *International Journal of Lexicography*. 2:1-14
- Manning & Schutze (1999) *Fondations of Statistical Natural Language Processing*. Cambridge (Mass.) : MIT Press
- McKeown, K. & R.R. Dragomir (2000) « Collocation », In Dale, Moisi & Somers. (Eds.) *Handbook of Natural Language Processing*, New-York : Dekker.
- Pierrel, J.-M. Ed. (2000) *Ingénierie des langues*, Paris : Hermes
- Jones, D. & H. Somers (1997) *New methods in language processing*, London : UCL Press