La complémentation des adjectifs à partir d'un grand corpus annoté automatiquement

Cécile Fabre
cecile.fabre@univ-tlse2.fr
UE TAL – 28/11/2008

Plan

- Point de départ : travaux antérieurs sur la sous-catégorisation des verbes
- Le traitement des adjectifs par Syntex
- Les données extraites
- Les mesures de filtrage
  - Indices sur la nature de la relation : sujet impersonnel, obligatorité, productivité
- Premiers résultats :
  - Mieux caractériser les patrons existants
  - Mettre au jour de nouveaux patrons

Plan

- Point de départ : travaux antérieurs sur la sous-catégorisation des verbes
- Le traitement des adjectifs par Syntex
- Les données extraites
- Les mesures de filtrage
  - Indices sur la nature de la relation : sujet impersonnel, obligatorité, productivité
- Premiers résultats :
  - Mieux caractériser les patrons existants
  - Mettre au jour de nouveaux patrons

Du *Treebank* au grand corpus annoté automatiquement

- Disponibilité de grands volumes de données issues de corpus annotés
- Travaux précédents sur les compléments verbaux
- Démarche :
  - Etudier à grande échelle les propriétés combinatoires des SP
  - Capter la distinction complément-ajout
  - Traduire les tests usuels en mesures applicables sur corpus

Travaux sur les verbes

- Extraction des cadres de sous-catégorisation :
  - Brent 1993, Federici et al. 1998… :

    Fréquence de la cooccurrence, critère de placement (pp argument jouxte le verbe)
- Focalisation sur la distinction complément-ajout :
  - Fabre et Frérot 2002
    - Dépendance au verbe vs autonomie
  - Merlo et Leybold 2001, Merlo et Ferrer 2006
    - Head dependence
    - Optionality
    - Iterativity
    - Verb classes

Distinction complément/ajout

- Mesure graduelle de l'autonomie des SP : (Fabre et Bourigault 2008), (Fabre, Rebeyrolle et Ho-Dac 2008)
- Tests usuels :
  - Caractère obligatoire ou facultatif
  - Déplacement
  - Détermination exercée par le verbe sur la préposition
- Traduits par des mesures sur corpus :
  - Placement du SP dans le corpus (étude de la position préverbale)
  - Calcul du degré d'autonomie du SP par rapport au verbe

Degré d'autonomie

- SP d'autant plus autonome qu'il apparaît plutôt avec des verbes qui sélectionnent faiblement la préposition
- SP d'autant moins autonome qu'il apparaît plutôt avec des verbes qui sélectionnent fortement la préposition

(à, question$_D$)	**(à, dehors$_D$)**
faible autonomie	forte autonomie
Verbes associés :	Verbes associés :
soustraire *s'intéresser* *consacrer* *soumettre* *renoncer* *ressembler* *s'attendre* *s'accrocher*	*transpirer* *tendre* *manger* *travailler* *retenir* *passer* *se précipiter* *pousser*

Sous-catégorisation des adjectifs

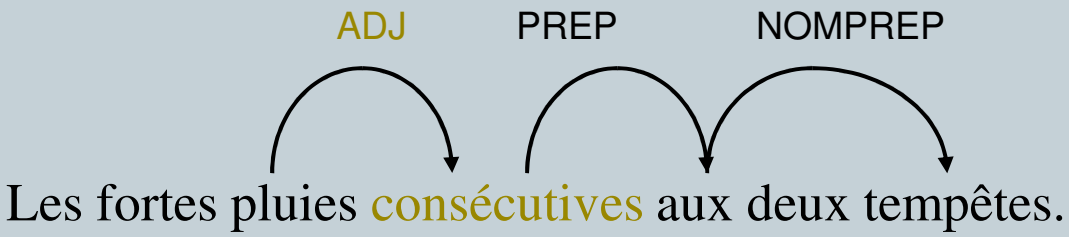
- Objectifs :
  - Apporter des indices pour caractériser les patrons existants
  - Découvrir de nouvelles instances des patrons ou de nouveaux patrons
- Démarche : mettre au point des indices de complémentation pour les adjectifs
  - Critères de valence moins bien définis :
    - Compléments très rarement obligatoires
    - Critères de permutabilité / pronominalisation (encore) moins bien établis
    - Plus grande variabilité des prépositions

Plan

- Point de départ : travaux antérieurs sur la sous-catégorisation des verbes
- Le traitement des adjectifs par Syntex
- Les données extraites
- Les mesures de filtrage
  - Indices sur la nature de la relation : sujet impersonnel, obligatorité, productivité
- Premiers résultats :
  - Mieux caractériser les patrons existants
    - Le cas des constructions impersonnelles
    - Ordonner les patrons
  - Mettre au jour de nouveaux patrons

I – Le traitement des adjectifs par Syntex

- Source : Bourigault 2007
- Relation épithète : ADJ



AdjFP|consécutif|consécutives|4|ADJ;3|PREP;5

- cat
- lemme
- forme
- num
- gouverneur
  REL;numGouv
- dépendant
  REL_1;num$Dép_1$, …

ADJ - Procédure d'attachement

Recherche des gouverneurs candidats

- À droite
- À gauche

Indices endogènes et stratégie de désambiguïsation

- L'indice pour le candidat c_i est égal à la fréquence du triplet (c_i,*ADJ*,*a*)

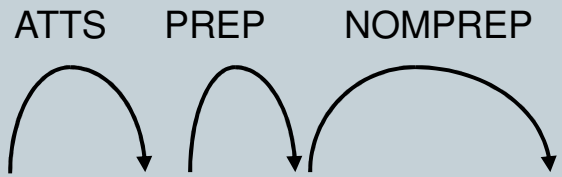
ex : *plainte pour discrimination déposée :* on compare la fréquence de (*plainte, ADJ, déposée*) et (*discrimination, ADJ, déposée*)

- Si le dépendant est un participe passé, on utilise la fréquence du triplet (*a*,*OBJ*,c_i)

ex : (*déposer, OBJ, plainte*)

Autres relations entre un adj et son gouverneur

Relation ATTS



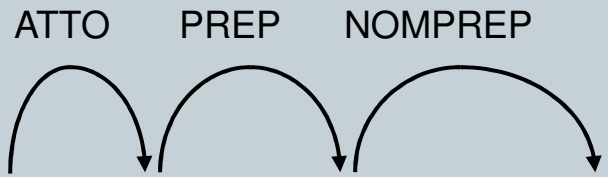
Nous serions enclins à nous consacrer davantage à notre sphère privée.

= relation non ambiguë

L'analyseur exploite une liste de verbes susceptibles de se construire avec un attribut du sujet (*être, devenir, paraître…*)

Autres relations entre un adj et son gouverneur

Relation ATTO



Ces gènes doivent subir une mutation qui les rend insensibles à toute régulatio

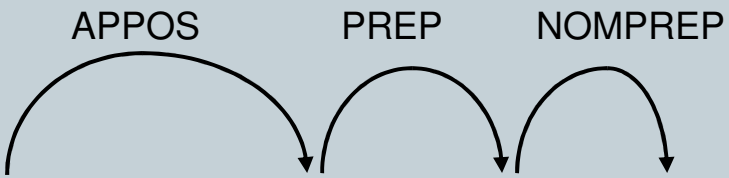
= relation non ambiguë (un seul gouverneur candidat)

L'analyseur exploite une liste de verbes susceptibles de se construire avec un attribut de l'objet (*considérer, croire, déclarer, juger…*)

Seuls les cas non ambigus sont traités :
Ce droit est estimé légitime
Ce droit que j'estime légitime
Je l'estime légitime
J'estime légitime ce droit mais pas : *J'estime ce droit légitime*

Autres relations entre un adj et son gouverneur

Relation APPOS



L'Europe du sud, soucieuse de défendre avant tout une alimentation raisonnée...

Gouverneur non identifié

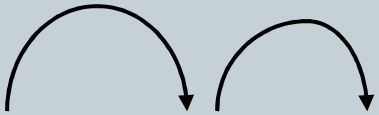
(NOGOV)

- Incise en début de phrase :

Mais, soucieuse d'affirmer son indépendance, la banque centrale refusa de céder…

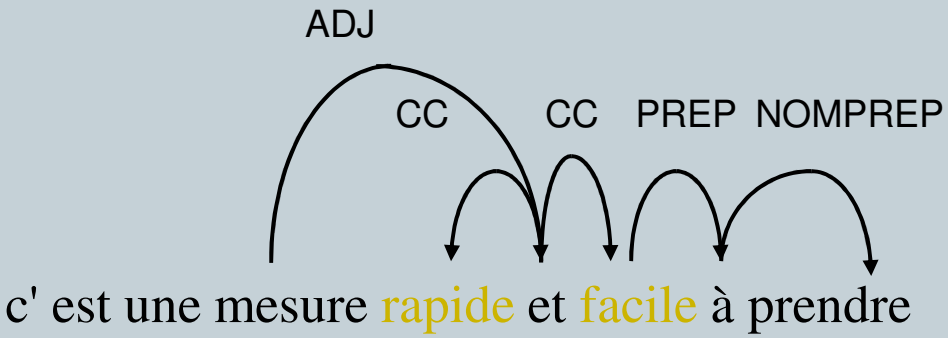
- Constructions ambiguës non résolues :

L' Inde rend Islamabad responsable du détournement du vol 814 en Afghanistan



des triangulaires suicides pour la droite et extrêmement bénéfiques pour la gauche

Cas de coordination entre adjectifs



C’est la conjonction de coordination qui porte la relation => on récupère la relation à partir d’elle

Plan

- Point de départ : travaux antérieurs sur la sous-catégorisation des verbes
- Le traitement des adjectifs par Syntex
- Les données extraites
- Les mesures de filtrage
  - Indices sur la nature de la relation : sujet impersonnel, obligatorité, productivité
- Premiers résultats :
  - Mieux caractériser les patrons existants
    - Le cas des constructions impersonnelles
    - Ordonner les patrons
  - Mettre au jour de nouveaux patrons

Les données extraites

- 10 ans du *Monde*, 200 millions de mots
- Extraction :
  - Tous les adj munis d'une relation PREP

    patrons ADJ [$_{PP}$ prep] et ADJ [$_{VPinf}$ prep]
  - Principales informations extraites :

adj	prep	dep	cat-dep	rel	suj	seq
capable	de	rouler	VINF	NO GOV	Undef	[…] avec son rêve de voitures capables de rouler […]
idéal	pour	cyclo-cross	N	ATTS	ce	C'est idéal pour le cyclo-cross […]
interne	à	Ps	N	ADJ	document	Dans un document interne au PS […]

Répartition des relations

REL	%
ATTS	30
ADJ	29
NO GOV	28
APPOS	11

Filtrage des constructions spéciales

- Constructions éliminées :
  - Constructions superlatives :

  un des plus ADJ [$_{PP}$ de] *un des plus talentueux de tout Paris*

  le plus ADJ [$_{PP}$ de] *le plus anglais des Irlandais*
  - Constructions marquant l'intensité :

  (trop, suffisamment, assez) ADJ [$_{VPinf}$ pour]

  assez souple pour supporter
- Il reste des constructions impossibles à filtrer :
  - Ce tableau est abominable de niaiserie

Filtrage par la productivité

- Productivité du couple adj, prep : nombre de dépendants différents
- On conserve les patrons de prod >= 3
- Résultats

	Treebank	Le Monde
nb d'adjectifs différents	285	2684
patrons *prep* (VINF | N)	26	136
patrons *adj prep* (VINF |N)	376	6778 542438 occurrences

Des données très bruitées

- Étiquetage:

abbatial	de	N	*Abbatiale de Souillac*
sauf	par	N	*sauf par grand froid*

- Segmentation :

polémique	à	N	*terriblement polémique au contraire*
influent	de	N	*l'un des plus influents du côté protestant*

- Analyse syntaxique :

ponctuel	à	N	*offrir une aide ponctuelle à des élèves*
vif	de	N	*le ton très vif de cette réponse*

Plan

- Point de départ : travaux antérieurs sur la sous-catégorisation des verbes
- Le traitement des adjectifs par Syntex
- Les données extraites
- Les mesures de filtrage
  - Indices sur la nature de la relation : sujet impersonnel, obligatorité, productivité
- Premiers résultats :
  - Mieux caractériser les patrons existants
    - Le cas des constructions impersonnelles
    - Ordonner les patrons
  - Mettre au jour de nouveaux patrons

1er indice : repérer les constructions impersonnelles

- Séparer les constructions impersonnelles des véritables *frames* :
- ATTS
  - Il est fier de dire que … => VPinf[de]
  - Il est absurde de dire que … => Ø
- Indices
  - Proportion de sujets « il », « cela », « ce »

mesure **propImpers** = nb de sujets impersonnels / nb total de sujets

calculée pour la relation ATTS

fier de VINF : < 0,1

Patron SUJ:VPinf[de] - Treebank

- 40 patrons

propImpers >= 0,9	**propImpers <0,15**
27 patrons	12 patrons
absurde, acceptable, anormal, dangereux, déconcertant, difficile, exact, fâcheux, facile, impératif, important, impossible, indispensable, inexact, intéressant, inutile, judicieux, nécessaire, pertinent, possible, préférable, prématuré, prudent, rare, ruineux, superflu, utile	capable, conscient, content, coupable, désireux, fier, heureux, incapable, libre, soucieux, sûr, susceptible

- Seule valeur intermédiaire : nouveau [$_{\text{VINF}}$ de] (prod = 5)

  L'idée n'est pas nouvelle de mettre

  L'utopie n'est pas nouvelle de pimenter …

Patron SUJ:VPinf[de] – *Le Monde*

- 431 patrons (prod >= 3)

propImpers >= 0,8	**propImpers <= 0,2**
362 patrons	32 patrons
aberrant, abominable, affreux, agaçant, aléatoire, américain, artificiel, chic, chouette, cocasse, comique, commun, cool, consternant, débile, fantastique, fastidieux…	aimable, aise, anxieux, avide, capable, certain (0,18), conscient, content, coupable, désireux, digne, fier, fort, furieux, heureux, impatient, incapable, inquiet, las, libre, mécontent, mince, nul, partisan, prêt…

- 37 patrons de valeurs intermédiaire : malheureux [$_{VINF}$ de], gentil[$_{VINF}$ de], lourd[$_{VINF}$ de], curieux[$_{VINF}$ de], fou[$_{VINF}$ de]…

  c'est lourd de lui faire porter… / l'erreur serait lourde de le cantonner

  Julie est trop malheureuse d'avoir été / c'est malheureux de parler de ça

2ème indice : le critère d'optionalité

- La réalisation syntaxique des adjectifs est généralement optionnelle
- Valeur *propAvecPrep* :

  part des occurrences de l'adjectif trouvées avec un rattachement prépositionnel
- 100 adjectifs pour lesquels Syntex a trouvé un rattachement prépositionnel dans + de la moitié des cas :

  passible, désireux, susceptible, inhérent, incapable, enclin, issu, originaire, capable, annonciateur, tributaire, apte, servi, exempt, appartenant, attentatoire, soucieux, synonyme, prompt, perclus, infoutu, générateur, féru …

3ème indice : la productivité

- « apprécier le rendement des constructions syntaxiques » (Legallois)
- Productivité (adj, prep) : nbre de dépendants différents
- **Productivité pondérée** : part de la préposition dans la productivité totale de l'adjectif

		prod	prod pondérée
étonnant	de N	35	0,3
	de VINF	17	0,14
	dans N	28	0,24
	pour N	9	0,07
	par N	7	0,06
			…
			prod totale : 115

Des valeurs de valence ?

- Existence d'une complémentation => *propAvecPrep*
- Régularité du patron => productivité
- Degré d'association Adj Prép => productivité pondérée

Valeurs moyennes obtenues sur le corpus *Le Monde*

	Patrons *Treebank*	**Patrons *Le Monde***
propAvecPrep	25%	4%
Productivité	293	20
Productivité pondérée	0,5	0,3

Plan

- Point de départ : travaux antérieurs sur la sous-catégorisation des verbes
- Le traitement des adjectifs par Syntex
- Les données extraites
- Les mesures de filtrage
  - Indices sur la nature de la relation : sujet impersonnel, obligatorité, productivité
- Premiers résultats :
  - Mieux caractériser les patrons existants
  - Mettre au jour de nouveaux patrons

Caractérisation des patrons Treebank
adj [PP prép]

Valeurs hautes

- le patron est productif
- l'adjectif est trouvé la plupart du temps avec une expansion prépositionnelle
- l'expansion prépositionnelle de cet adjectif est majoritairement introduite par cette prép

31 patrons relèvent nettement de ce cas de figure (prod>50, propAvecPrep>0,5, prodPond>0,5):

avare, dépendant, conscient, constitutif, originaire, préjudiciable [$_{PP}$ de]

allergique, propice, favorable, conforme, préjudiciable [$_{PP}$ à]

compatible/incompatible [$_{PP}$ avec]

Caractérisation des patrons Treebank
adj [PP prép]

Valeurs basses

- le patron est peu productif
- l'adjectif est rarement trouvé avec une expansion prépositionnelle
- cette expansion est très rarement introduite par cette préposition

24 patrons relèvent nettement de ce cas de figure (prod<10, propAvecPrep<0,1, prodPond<0,1):

courant, gros, grand, grave, habituel, particulier, vrai [$_{PP}$ *en*]

dynamique [$_{PP}$ *depuis*] *épars* [$_{PP}$ *de*]

obligatoire, rare, traditionnel [$_{PP}$ *à*]

rare, satisfaisant [$_{PP}$ *sur*]

lourd, trompeur [$_{PP}$ *pour*]

ferme [$_{PP}$ *vis-à-vis de*]

Premier résultat : ordonnancement des patrons existants

patron	prod	prod Pond	propAvecPrep	propImpers
âgé de N	265	0,91	0,69	
aisé à VINF	110	0,32	0,16	0,04
applicable à N	450	0,74	0,48	
applicable sur N	1	0,01	0,48	
conforme à N	544	0,97	0,68	
dangereux de VINF	199	0,33	0,1	0,93
grave en N	3	0,02	0,002	
propre à N	144	0,58	0,005	
Difficile à N	29	0,01	0,37	

Découvrir de nouvelles instances des patrons

- Extraire les nouvelles instances qui satisfont les critères de valence optimaux
- Valeurs de valence hautes (propAvecPrep>=50%, prodPond>=0,5)
  - sur adj productifs (prod >= 50)
    33 patrons [$_{PP}$ à], [$_{PP}$ de], [$_{VPinf}$ à], [$_{VPinf}$ de]
    attribuable, inapte, réductible, dévolu … [$_{PP}$ à]
    hérissé, perclus, avide, composé, typique … [$_{PP}$ de]
    soluble [$_{PP}$ dans]
  - sur adj moins productifs (prod < 50)
    23 patrons [$_{PP}$ à], [$_{PP}$ de], [$_{VPinf}$ à], [$_{VPinf}$ de], [$_{PP}$ par],
    déductible, indétachable, recru, ruisselant … [$_{PP}$ de]
    attentatoire [$_{PP}$ à]
    poursuivi [$_{PP}$ pour]
    infoutu [$_{VPinf}$ de]
    servi [$_{PP}$ par]

Valeurs intermédiaires

- propAvecPrep entre 20 et 50%, prodPond entre 0,2 et 0,5

  58 patrons

  marri [$_{VPinf}$ de]
  croulant [$_{PP}$ sous] dubitatif [$_{PP}$ sur], chatouilleux [$_{PP}$ sur]
  loyal [$_{PP}$ à]
  fondé, payable, repérable, utilisable [$_{PP}$ par]
  inégalé, invaincu, vacant [$_{PP}$ depuis]…

Découvrir de nouveaux patrons ?

- prép n'apparaissant pas dans les patrons Treebank

a) Prép réputées « argumentales »

(valeur argumentale dans PrepLex, pour les V)

	prodPond	prod	propAvecPrep
livrable à partir de N	0,21	17	0,49
furieux contre N	0,20	39	0,19
valable jusqu'à N	0,22	91	0,26
novice en matière de N	0,24	8	0,16
indisponible pendant N	0,22	7	0,26

b) Prép réputées « non argumentales »

méfiant à l'égard de N	0,6	103	0,2
+ indulgent, injurieux, circonspect			
sceptique quant à N	0,24	25	0,26
+ optimiste, pessimiste, circonspect, dubitatif			
Minoritaire au sein de N	0,14	46	0,11
+ majoritaire			

⇒ De l'acquisition de patrons de valence à l'observation de schémas de « colligation »

« comportement de cooptation, de préférence mesurée statistiquement à partir de corpus, sans [...] qu'il soit redevable à quelque principe structural » (Legallois)

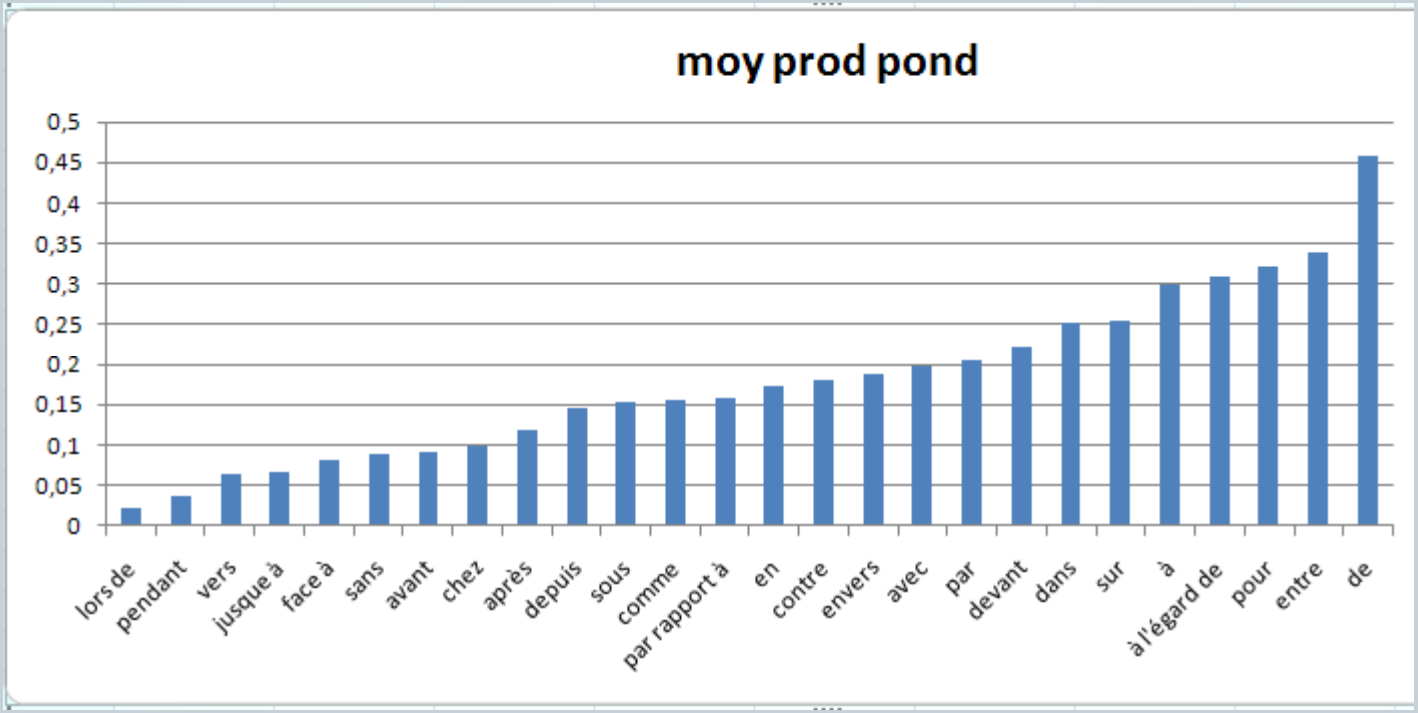
Conclusion

- Complémentarité Treebank / Le Monde annoté :
  - Le Monde : volume +, fiabilité –
  - Treebank : volume -, fiabilité +
- Grand volume de données annotées automatiquement permettent :
  - De fournir de nouveaux éléments pour filtrer les patrons déjà acquis
  - D'étendre les patrons
- Valence / « colligation »
  - Du jugement binaire à l'appréciation graduelle

Perspectives

- Examiner de plus près les données…
  - Propriétés d'adjectifs particuliers
    - Comparatif entre (essai, étude, réflexion, résultat, test…)
  - Alternance de prép
    - Incrédule devant/face à
    - Inquiet de/face à/quant à/devant/à l'idée de

    => étude des noms dépendants
- Comparaison complémentation des verbes / des adjectifs :
  - Des propriétés distinctes
  - Des méthodes d'acquisition à mieux distinguer
- Evaluation ?



Valeurs calculées sur patrons propres à LM

Bibliographie

- Bourigault, D. (2007). *Un analyseur syntaxique opérationnel : SYNTEX*. Thèse d'habilitation à diriger des recherches, Université Toulouse 2-Le Mirail.Brent M-R. (1993), From Grammar to Lexicon : Unsupervised Learning of Lexical Syntax, *Computational Linguistics*, Vol.19 : 2, pp.243-262.
- Fabre, C., Rebeyrolle, J. et Ho-Dac Mai (2008) "Examen du statut des syntagmes prépositionnels à la lumière de données issues de corpus annotés", CMLF (Congrès Mondial de Linguistique Française).
- Fabre, C. et Bourigault, D. (2008), "Exploiter des corpus annotés syntaxiquement pour observer le continuum entre arguments et circonstants", *Journal of French Language Studies*, 18(1) , 87-102.
- Fabre, C. et Frérot, C. (2002). Groupes prépositionnels arguments ou circonstants : vers un repérage automatique en corpus, *Actes du colloque TALN*, Nancy, 215-224.
- Federici S., Montemagni S., Pirrelli V., Calzolari N. (1998), Analogy-based Extraction of Lexical Knowledge from Corpora: the SPARKLE Expérience, Actes de *LREC*, Grenade, 75-82.
- Legallois, D. (2005). Du bon usage des expressions idiomatiques dans l'argumentation de deux modèles anglo-saxons: la grammaire de construction et la grammaire contextualiste, *Les Cahiers de l'Institut de Linguistique de Louvain (CILL)*, 31, 2-4, 109-127.
- Merlo P. and M. Leybold (2001) " Automatic Distinction of Arguments and Modifiers: the Case of Prepositional Phrases", Workshop on Computational Language Learning (Connl 2001), Toulouse, France.
- Merlo P., E. Esteve Ferrer (2006) "The Notion of Argument in PP Attachment", Computational Linguistics 32(2).