Prédire la position de l'adjectif épithète en français

Juliette Thuilier[1], Gwendoline Fox [2] & Benoît Crabbé [1]

[1]Alpage (Paris 7 - INRIA)

[2]Paris 3 & EA 1483

19 novembre 2009

LingLunch

1. Introduction

Continuum de possibilités pour l'antéposition et la postposition des adjectifs

(1) a. un entretien long de deux heures
b. *un long de deux heures entretien

(2) a. Même dans les pays où la progression salariale n'a pas été freinée, les *importants* gains de productivité qui auront marqué 1992 ... (FTB)
b. Même dans les pays où la progression salariale n'a pas été freinée, les gains de productivité *importants* qui auront marqué 1992 ...

(3) a. Une banque d'affaire *britannique*
b. ? ? ? Une *britannique* banque d'affaire
c. ... la très *britannique* banque d'affaire et de marché vient d'acheter ... (FTB)

(4) a. La location, échappant à toute réglementation, offre une souplesse de gestion *réelle* au propriétaire
b. La location, échappant à toute réglementation, offre une *réelle* souplesse de gestion au propriétaire (FTB)

Facteurs d'ordre divers : phonologie, syntaxe, sémantique, discours
Forsgren (1978); Wilmet (1981); Nølke (1996); Noailly (1999); Abeillé and Godard (1999)

Hypothèse

La position des adjectifs est guidée en grande partie par des contraintes **préférentielles**

- Etude sur le placement de l'adjectif $\Rightarrow$ mise à jour de ces contraintes préférentielles
- Outils disponibles :
  - corpus annoté : French Tree Bank (Abeillé et al., 2003; Abeillé and Barrier, 2004)
  - méthodes statistiques inférentielles : la régression logistique (Agresti, 2007)
- on essaie de tirer des généralités sur des questions de préférences à partir de l'étude de corpus (Bresnan et al., 2007)

Restriction

Les résultats de prédictions peuvent être satisfaisants en se basant seulement sur la forme

- résultats satisfaisants sans prendre en compte la sémantique liée à la position
- les différences sémantiques liées position de l'adjectif ne font pas partie de l'objet de notre travail

(5) l'*ancien* couvent / le couvent *ancien*

(6) un *gros* fumeur / un fumeur *gros*

Objectifs

- prédire la position de l'adjectif
- modéliser le phénomène de placement de l'adjectif
- interpréter les facteurs intervenant, évaluer leur importance

2. Méthodes

Corpus et extraction de données

- French Tree Bank
  - adjectifs épithètes apparaissant avec une tête nominale
  - élimination des cardinaux, des adjectifs contenus dans des dates, des abréviations, des occurrences problématiques au niveau de l'annotation
  - données : **15324 occurrences**

- Table de données : pour chaque occurrence
  - la position par rapport au nom
  - 20 variables prédictrices (ex : adjectif de nationalité, présence ou non d'un adverbe, longueur de l'adjectif...)

- Variables prédictrices
  - informations extraites directement du FTB
  - enrichissement grâce à des dictionnaires (ProlexBase Tran and Maurel (2006), Chroma (extrait du web))
  - syllabation du corpus à l'aide du logiciel de synthèse vocale eLite
  - analyse morphologique dérivationnelle (logiciel Derif (Namer, 2002))

	adj	nom	position	art_def	det_poss	det_dem	coord	post_adj
21	principal	représentant	1	0	0	0	0	0
22	strict	alignement	1	0	0	0	0	0
23	actuelle	heure	0	1	0	0	0	0
24	seule	arrivée	1	1	0	0	0	0
25	financière	commission	0	1	0	0	0	0
26	dernier	congrès	1	1	0	0	0	0
27	excessif	nombre	0	1	0	0	0	0
28	principal	enjeu	0	1	0	0	0	0
29	internes	débats	0	1	0	0	0	1
30	profond	renouvellement	1	0	0	0	0	0

	co_ocAnt	co_ocPost	sprep	prel	adjindef	adjsyll	Sadj_syll	diffsyll_a_n	natio
21	0	0	1	0	0	3.000000	3.0000000	-1.0000000000	0
22	0	0	1	0	0	1.000000	1.0000000	-3.0000000000	0
23	0	0	0	0	0	3.000000	3.0000000	2.0000000000	0
24	0	0	1	0	0	1.000000	1.0000000	-2.0000000000	0
25	0	0	0	0	0	3.000000	3.0000000	0.0000000000	0
26	0	0	0	0	0	2.000000	2.0000000	0.0000000000	0
27	0	0	1	0	0	3.000000	3.0000000	1.2250000000	0
28	0	0	1	0	0	3.000000	3.0000000	1.0000000000	0
29	0	0	0	0	0	2.500000	8.5591220	0.5000000000	0
30	0	0	1	0	0	2.000000	2.0000000	-2.0000000000	0

FIG.: Extrait de la table de données

Description des données

- 15324 occurrences
  - 4309 en antéposition (28.1%)
  - 11015 en postposition (71.9%)
- 1993 lemmes

Répartition antéposés/postposés des lemmes et des occurrences

	antéposés	*postposés*	*2 positions*	*Totaux*
nombre de lemmes	123	1684	186	1993
	6.2%	*84.5%*	*9.3%*	*100%*
occurences	484	9122	5718	15324
	3.2%	*59.5%*	*37.3%*	*100%*

Inférence statistique : Régression logistique

- permet de modéliser le comportement d'une variable binaire en fonction de variables prédictrices
- position de l'adjectif = variable binaire

  postposition = 0
  antéposition = 1

Fonction logistique

Fonction logistique = fonction à valeurs dans l'intervalle $[0, 1]$

$$\pi_{\text{ante}} = \frac{e^{\beta \mathbf{X}}}{1 + e^{\beta \mathbf{X}}} \qquad (1)$$

où

- π_{ante} = probabilité d'antéposition de l'adjectif
- β = coefficients de régression $\alpha, \beta_0 \ldots \beta_n$
- $\mathbf{X}$ = variables prédictrices $X_0 \ldots X_n$

$$\pi_{\text{ante}} = \frac{e^{\mathbf{x}\beta}}{1+e^{\mathbf{x}\beta}}, \text{ où}$$

$$X\beta =$$

-0.90	
$+0.87$	DET-DEM $= 1$
$+0.20$	DET-POSS $= 1$
-0.34	ART-DEF $= 1$
-0.29	ADV $= 1$
-15.58	POST-ADJ$= 1$
$+0.71$	SPREP $= 1$
$+0.32$	CO-OCPOST $=1$
-0.34	CO-OCANT $= 1$
-1.98	COORD $= 1$

FIG.: Exemple de formule d'un modèle de régression logistique

Outillage pour l'interprétation

1. Coefficients affectés à chaque variable sont interprétables
2. Comparaison des capacités de prédiction pour déterminer contraintes pertinentes

Calcul de l'exactitude (évaluation 10 passes)

- entraînement sur 9/10 des données
- test sur 1/10 des données
- opération répétée 10 fois sur 10 ensembles de données différents
- exactitude = moyenne μ des 10 scores obtenus (et σ l'écart-type)

Modèle Nul

- Modèle servant de point de référence
  - ne contient **aucune variable prédictrice**
  - prédit systématiquement la **postposition**
- Exactitude $\mu = 71.9\%$ ($\sigma = 0.018$)

Matrice de confusion du modèle Nul

		Position prédite		% de prédiction
		P	A	
Position observée	P	11015	0	100%
	A	4309	0	0%

3. Modèle combinatoire

Contraintes

Configuration du SAdj

1. POST-ADJ : présence ou non d'un dépendant post-adjectival
2. ADV : présence ou non d'un modifieur adverbial
3. COORD : adjectif en coordination avec un autre adjectif ou non

Configuration du SN

4. CO-OCANT : co-occurrence ou non avec un adjectif antéposé

   (7) a. Les *autres grands* constructeurs ont fait la même proposition (FTB)
   b. la situation de son *nouvel* empire *hollywoodien* (FTB)

5. CO-OCPOST : co-occurrence ou non avec un adjectif postposé

   (8) Interrogé sur la compensation *salariale partielle* (FTB)

Contraintes (suite)

Configuration du SN (suite)

6. SPREP : présence ou non d'un SPrep dans le SN
7. PREL : présence ou non d'une proposition relative dans le SN

Déterminant du SN

8. ART-DEF : le déterminant est un article défini ou non
9. DET-DEM : le déterminant est démonstratif ou non
10. DET-POSS : le déterminant est possessif ou non

Modèle Combinatoire

- contient 9 variables
- variable PREL écartée car elle ne participe pas significativement au modèle
- Exactitude totale : $\mu = 71.8\%$ ($\sigma = 0.016$)

Matrice de confusion du modèle Combinatoire

		Position prédite		% de prédiction
		P	A	
Position observée	P	10887	128	98.8%
	A	4203	106	1.0%

- Exactitude comparable à celle du modèle Nul
- la prédiction correcte pour l'antéposition ne s'améliore pas sensiblement

$\Rightarrow$ les contraintes combinatoires ne sont pas de bons prédicteurs

4. Modèle Lexical

Contraintes

Morphologie

1. DERIVE : l'adjectif est issu ou non d'une autre partie du discours (verbe ou nom)

Classes d'adjectifs

2. NATIO : l'adjectif désigne une nationalité ou non
3. COULEUR : l'adjectif dénote une couleur ou non
4. ADJINDEF : l'adjectif appartient à la classe des indéfinis ou non

indéfinis : *tel, autre, certain, quelques, divers, différent, maint, nul, quelconque, même*

Longueur et fréquence

5. ADJSYLL : longueur de l'adjectif en syllabes
6. ADJFREQ : fréquence du lemme dans la table de données

Modèle Lexical

- contient les 6 variables qui concernent l'item adjectival
- Exactitude totale : $\mu = 84.8\%$ ($\sigma = 0.011$)

Matrice de confusion du modèle Lexical

		Position prédite		% de prédiction
		P	A	
Position observée	P	10337	678	93.8%
	A	1661	2648	61.6%

$\Rightarrow$ impact important des caractéristiques lexicales

$\Rightarrow$ La prédiction correcte de l'antéposition est à plus de 60%

5. Modèle de Préférences

Approximation des caractéristiques lexicales

1. ADJ-PREFANT : l'adjectif a une préférence pour l'antéposition ou non
2. ADJ-PREFPOST : l'adjectif a une préférence pour la postposition ou non

Calcul des préférences

- approximations pour détecter si, statistiquement, l'adjectif a une préférence pour une position
- deux dictionnaires : celui des adjectifs anormalement antéposés, celui des adjectifs anormalement postposés
- membre du dictionnaire = adjectifs dont le nombre de positions observées est significativement différent du nombre de positions attendu sur la base d'une loi binomiale (seuil $\alpha = 0.05$)

Exemples pour les préférences

Lemme *vieux*

- 14 occurrences en antéposition
- 2 occurrences en postposition

- ADJ-PREFANT = 1
- ADJ-PREFPOST = 0

Lemme *variable*

- 0 occurrences en antéposition
- 5 occurrences en postposition

- ADJ-PREFANT = 0
- ADJ-PREFPOST = 0

Dans la table de données :

- 7502 occcurrences (49.0%) ont une valeur positive pour ADJ-PREFPOST,
- 3851 occurrences (25.1%) ont une valeur positive pour ADJ-PREFANT.

Modèle de Préférences

- Modèle contenant les 2 variables de préférence
- Exactitude totale : $\mu = 91.2\%$ ($\sigma = 0.007$)

Matrice de confusion du modèle de Préférences

		Position prédite		% de prédiction
		P	A	
Position observée	P	10567	448	95.9%
	A	904	3405	79.0%

$\Rightarrow$ préférences lexicales améliorent nettement la prédiction de l'antéposition (79%)

6. Modèle Global

- construit à partir de l'ensemble des contraintes vues précédemment
- contraintes supplémentaires

SAdj-syll : longueur du SAdj

diffsyll-a-n : différence entre la longueur de l'adjectif et celle du nom
Une valeur positive indique que l'adjectif est plus long que le nom
Une valeur négative indique que l'adjectif est plus court que le nom

- contient 15 variables
  - Co-ocAnt, art-def, Coord, adjsyll et adjfreq éliminées
- Exactitude totale : $\mu = 91.9\%$ ($\sigma = 0.008$)

Matrice de confusion du modèle global

		Position prédite		% de prédiction
		P	A	
Position observée	P	10500	515	95.3%
	A	729	3580	82.3%

⇒ Meilleur modèle de prédiction obtenu

⇒ Faible différence d'exactitude entre modèle global (91,9%) et modèle de préférence (91.2%)

⇒ Les préférences lexicales sont centrales

$\pi_{\text{ante}} = \frac{e^{\mathbf{x}\beta}}{1+e^{\mathbf{x}\beta}}$, où $X\beta =$

-0.33	
$+0.81$	DET-DEM $= 1$
$+0.48$	DET-POSS $= 1$
$+1.00$	SPREP $= 1$
$+0.34$	PREL $= 1$
$+0.60$	CO-OCPOST $=1$
$+1.05$	ADJINDEF $= 1$
$+2.80$	ADJ-PREFANT $= 1$
-0.49	ADV $= 1$
-15.99	POST-ADJ$= 1$
-0.45	SADJ-SYLL
-0.30	DIFFSYLL-A-N
-0.37	DERIVE $= 1$
-3.32	NATIO $= 1$
-17.20	COULEUR $= 1$
-2.70	ADJ-PREFPOST $= 1$

Interprétation des coefficients

Toutes choses égales par ailleurs :

- si DET-DEM $= 1$, l'adjectif a 2,3 chances en plus d'être antéposé
- si ADJ-PREFANT $= 1$, l'adjectif a 16,4 chances en plus d'être antéposé
- ADV $= 1$, l'adjectif a 1,6 chances en plus d'être postposé
- ADJ-PREFPOST $= 1$, l'adjectif a 14,9 chances en plus d'être postposé

7. Résultats

- Utilisation et exploitation de méthodes statistiques pour prédire et analyser la place des adjectifs
- hypothèse selon laquelle il s'agit de contraintes de préférence semble vérifiée
- modélisation satisfaisante malgré la restriction au niveau de la sémantique liée à la position
- contraintes combinatoires $\Rightarrow$ effet nul en terme quantitatif
- contraintes lexicales $\Rightarrow$ contraintes les plus importantes

Etude des erreurs

Mise à jour de patrons qui ressortent fréquemment dans les erreurs

(9) (à) juste titre

(10) l'été dernier / la semaine dernière / le mois dernier

(11) différent :

a. 26 antéposés / 26 postposés

b. déterminant défini $\Rightarrow$ 25 antépositions
déterminant non-défini $\Rightarrow$ 24 postpositions

- effets de figement (nom + adjectif), constructions
- montre que pour mieux décrire le phénomène, il faut prendre en compte d'autres informations lexicales (notamment le nom) et du contexte plus large
- intégration au modèle de prédiction ? par ex. collocations

Perspectives

1. Décomposition du composant lexical :
   - > à quoi renvoient les caractéristiques lexicales ?
     - – on a observé la fréquence et la longueur séparément
     - – modèle fréquence/longueur : exactitude $\mu = 80,7\%$ ($\sigma = 0,010$)
   - > les classes sémantiques ? annotation automatique envisageable ? annotation manuelle vaut la peine ?
2. Analyse en termes de constructions
   - > Grammaire de construction (Kay and Fillmore, 1999; Croft, 2001; Goldberg, 2006)
3. Plausibilité des modèles probabilistiques d'un point de vue expérimental
   - > modèle probabiliste proposé correspond-il à une forme de savoir linguistique ?

Abeillé, A. and N. Barrier (2004). Enriching a french treebank. In *Proceedings of Language Ressources and Evaluation Conference (*LREC*)*, Lisbon.

Abeillé, A., L. Clément, and F. Toussenel (2003). Building a treebank for french. In *Treebanks*. Dordrecht : Kluwer.

Abeillé, A. and D. Godard (1999). La position de l'adjectif èpithète en français : le poids des mots. *Recherches linguistiques de Vincennes 28*, 9–32.

Agresti, A. (2007). *An introduction to categorical data analysis*. Wiley interscience.

Bresnan, J., A. Cueni, T. Nikitina, and H. Baayen. (2007). Predicting the dative alternation. In G. Boume, I. Kraemer, and J. Zwarts (Eds.), *Cognitive Foundations of Interpretation*. Amsterdam : Royal Netherlands Academy of Science.

Croft, W. (2001). *Radical Construction Grammar*. Oxford University Press.

Forsgren, M. (1978). *La place de l'adjectif épithète en français contemporain, étude quantitative et sémantique*. Stockholm : Almqvist & Wilksell.

Goldberg, A. (2006). *Constructions at Work : the nature of generalization in language*. Oxford University Press.

Kay, M. and C. J. Fillmore (1999). Grammatical constructions and linguistic generalizations :the what's x doing y ? construction. *Language 75*, 1–34.

Namer, F. (2002). Acquisition automatique de sens à partir d'opérations morphologiques en français : étude de cas. In *Traitement Automatique de la Langue Naturelle (TALN)*.

Noailly, M. (1999). *L'adjectif en français*. Ophrys.

Nølke, H. (1996). Où placer l'adjectif épithète ? focalisation et modularité. *Langue française 111*, 38–57.

Tran, M. and D. Maurel (2006). Prolexbase : Un dictionnaire relationnel multilingue de noms propres. *Traitement automatique des langues 47*(3), 115–139.

Wilmet, M. (1981). La place de l'épithète qualificative en français contemporain : étude grammaticale et stylistique. *Revue de linguistique romane 45*, 17–73.