Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Syntaxe
DOI 10.1051/cmlf/2010161

Fréquence, longueur et préférences lexicales dans le choix de la position de l'adjectif épithète en français

Juliette Thuilier

Université Paris VII Denis Diderot & Alpage (Paris 7 – INRIA)
jthuilier@linguist.jussieu.fr

Gwendoline Fox

Université Paris III Sorbonne-Nouvelle , ILPGA & EA 1483
gfox@univ-paris3.fr

Benoît Crabbé

Université Paris VII Denis Diderot & Alpage (Paris 7 – INRIA)
bcrabbe@linguist.jussieu.fr

1 Introduction

Dans ce travail, nous étudions la position de l'adjectif par rapport au nom en nous appuyant sur des données de corpus. Le but de cet article est de fournir une modélisation du phénomène de placement de l'adjectif et d'interpréter les facteurs intervenant dans le phénomène ainsi que d'évaluer leur importance. Nous nous intéresserons particulièrement aux facteurs concernant l'item adjectival et nous discutons plus spécifiquement l'impact de contraintes comme la longueur et la fréquence sur sa position par rapport au nom.

En français, les adjectifs épithètes ont la possibilité d'être antéposés ou postposés au nom. Le placement de l'adjectif par rapport au nom est un phénomène complexe qui engage des facteurs d'ordre divers : phonologie, syntaxe, sémantique, discours (Forsgren, 1978 ; Wilmet, 1981 ; Nølke, 1996 ; Noailly, 1999 ; Abeillé & Godard, 1999). Notre travail repose sur l'idée centrale que le choix de la position de l'adjectif ne repose pas simplement sur des contraintes catégoriques, mais qu'il est également très largement influencé par des contraintes préférentielles. L'étude de ce phénomène doit donc passer par la mise à jour de ces contraintes. Un travail sur les contraintes de préférence implique la prise en considération de l'usage des locuteurs. Elle suppose donc que le savoir linguistique des locuteurs ne se limite pas à la compétence langagière, mais contient également des préférences d'usage (notamment Bresnan *et al.*, 2007). Ce type d'approche est rendu possible par la disponibilité d'un corpus annoté (le French Tree Bank, dorénavant FTB) et par une approche quantitative basée sur les méthodes statistiques inférentielles modernes.

Par ailleurs, nous restreignons l'objet du présent travail : nous étudions le phénomène du point de vue de la forme, laissant de côté pour le moment les différences de sens liées à la position de l'adjectif. D'abord, il n'existe pas de relation régulière entre le sens de l'adjectif et sa position (Abeillé & Godard, 1999). De plus, les adjectifs qui peuvent apparaître dans les deux positions sont, pour la plupart d'entre eux, interprétables de la même façon, en termes véricconditionnels, en antéposition et en postposition (Noailly, 1999; Abeillé & Godard, 1999). Enfin, les cas où il existe des effets sémantiques créés par la position de l'adjectif, comme dans les exemples (2) et (3), constituent un aspect du phénomène qui ne fait pas partie de l'objet de notre travail.

(1) un gros fumeur / un fumeur gros

(2) un ancien couvent / un couvent ancien

L'article est structuré de la façon suivante : dans la partie 2, nous exposerons les méthodes et les outils utilisés pour le travail de prédiction de la position de l'adjectif. La partie 3 sera consacrée à la présentation des contraintes que nous avons prises en compte et à leur description dans nos données. Dans

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Syntaxe
DOI 10.1051/cmlf/2010161

la partie 4, nous exposerons le modèle de prédiction de la position de l'adjectif et nous le décomposerons afin de mieux comprendre le phénomène du placement de l'adjectif. Nous nous attacherons notamment à montrer que la longueur et la fréquence sont des facteurs non négligeables dans le choix de la position de l'adjectif et que les informations relatives à l'item adjectival permettent de capturer, à elles seules, une partie importante du phénomène de placement de l'adjectif.

2 La méthode

2.1 Extraction de données sur corpus et enrichissement

Ce travail repose sur l'exploitation du corpus arboré du français : le French Tree Bank (Abeillé *et al*., 2003). Nous avons extrait l'ensemble des adjectifs épithètes apparaissant avec une tête nominale, puis éliminé les adjectifs numéraux, les adjectifs contenus dans des dates, les abréviations et certaines occurrences problématiques au niveau de l'annotation. À partir des adjectifs extraits, nous avons constitué une table de données contenant pour chaque occurrence, sa position par rapport au nom, ainsi que 10 variables. Celles-ci ont été élaborées sur la base des contraintes trouvées dans la littérature sur l'adjectif dans la limite des ressources dont nous disposons (voir partie 3).

Les variables concernant les classes lexicales ont été obtenues à l'aide de dictionnaires : dictionnaire d'adjectifs indéfinis, CHROMA[1] pour les adjectifs de couleur, PROLEXBASE (Tran & Maurel, 2006) pour les adjectifs de nationalité. Les adjectifs dérivés d'une autre partie du discours ont pu être repérés à l'aide de DERIF (Namer, 2002). De plus, grâce au logiciel de synthèse vocale ELITE, nous avons ajouté les informations relatives à la longueur en syllabes des adjectifs et des SAdj.

La table contient 15324 occurrences d'adjectifs, dont 4309 (28.1%) en antéposition et 11015 (71.9%) en postposition. Les occurrences se répartissent en 1993 lemmes. Le tableau 1 montre que les adjectifs qui présentent une alternance dans notre corpus sont peu nombreux du point de vue des lemmes (9,3% des lemmes). Cependant, en termes d'occurrences, ces lemmes représentent 5718 adjectifs, soit 37,3 % des occurrences. Cela signifie que les lemmes qui se rencontrent dans les deux positions ont tendance à être des adjectifs fréquents.

	Antéposé	Postposé	2 positions	Totaux
Nombre de lemmes	123 *6,2%*	1684 *84,5%*	186 *9,3%*	1993 *100%*
Occurrences	484 *3,2%*	9122 *59,5%*	5718 *37,3%*	15324 *100%*

Tableau 1 : Répartition des lemmes et des occurrences selon la position.

Notons que, parmi les occurrences des lemmes présentant la possibilité d'alternance de position, la proportion antéposition/postposition s'inverse par rapport aux données globales : 3825 antéposés (66,9 %), 1893 postposés (33,1 %). Les adjectifs qui alternent ont donc tendance à être antéposés. Ainsi, la tendance générale est que les occurrences postposées représentent des lemmes peu fréquents qui n'apparaissent qu'en postposition, alors que les occurrences antéposées renvoient à des lemmes plus fréquents qui montrent plus d'alternance[2].

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Syntaxe
DOI 10.1051/cmlf/2010161

2.2 Inférence statistique

À partir de la table de données, nous avons construit des modèles de régression logistique[3] (Agresti, 2007) qui devaient modéliser au mieux le choix de la position de l'adjectif dans le FTB à partir des variables prédictrices. L'intérêt principal du modèle de régression logistique est qu'il permet de prédire une variable binaire à partir d'un ensemble de variables prédictrices. Ainsi, considérant que la position de l'adjectif est une variable binaire (antéposition = 1 et postposition = 0), nous construisons un modèle de régression logistique qui prend un ensemble de variables, binaires ou continues, et qui donne comme résultat la probabilité pour un adjectif donné d'être antéposé. Si cette probabilité est inférieure à 0,5, la position prédite par le modèle est la postposition, et si cette probabilité est supérieure à 0,5 le modèle prédit l'antéposition. Par exemple, la figure 1 présente un modèle de régression logistique construit à partir de 6 variables : 2 variables continues (la longueur syllabique de l'adjectif – ADJ-SYLL ; la fréquence de l'adjectif – FREQ) et 4 variables binaires (l'adjectif dérive ou non d'une autre partie du discours – DERIVE ; l'adjectif appartient ou non à la classe des adjectifs indéfinis – ADJ-INDEF ; l'adjectif est un adjectif de nationalité ou non – NATIO ; l'adjectif est un adjectif de couleur ou non – COULEUR).

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

+1,21	
+0,007	FREQ
−1,15	ADJ - SYLL
−0.41	DERIVE = 1
+1.78	ADJ-INDEF = 1
−6.67	NATIO = 1
−15.29	COULEUR = 1

Figure 1 : Exemple de modèle de régression logistique

Chaque variable est affectée d'un coefficient. Ici, si le coefficient est positif, la variable vote en faveur de l'antéposition, inversement s'il est négatif, elle vote pour la postposition. Regardons, par exemple, ce que ce modèle prédit pour un adjectif dérivé d'une autre partie du discours de 2 syllabes et de fréquence 50, autrement dit un adjectif pour lequel ADJ-SYLL = 2, FREQ = 50, DERIVE = 1, ADJ-INDEF = 0, NATIO = 0 et COULEUR = 0

$$\pi_{ante} = \frac{e^{1,21+0,007*50-1,15*2-0,41*1+1,78*0-6,67*0-15,29*0}}{1+e^{1,21+0,007*50-1,15*2-0,41*1+1,78*0-6,67*0-15,29*0}} = 0,24$$

D'après le modèle en figure 1, la probabilité pour cet adjectif d'être antéposé est de 0,24, autrement dit le modèle prédit la postposition.

Afin de déterminer la pertinence et l'importance des contraintes, nous procédons à des comparaisons de modèles de prédiction. Notre démarche consiste à confronter les capacités de prédictions de modèles reposant sur différents groupes de contraintes. De cette façon, nous pouvons évaluer le pouvoir prédicteur et donc la pertinence des variables. La capacité de prédiction du modèle est calculée en 10 passes. Les données sont divisées en 10 parties égales. Le modèle est entraîné sur 9/10 des données et testé sur le 1/10 restant, et ce 10 fois en modifiant à chaque fois les données d'entraînement et les données test. La capacité de prédiction correspond à la moyenne des prédictions faites sur les 10 ensembles de données test. Elle est notée μ, et l'écart-type de cette moyenne est notée σ.

Le modèle qui sert de référence pour les comparaisons est le modèle Nul. Ce modèle ne contient aucune variable prédictrice et prédit toujours la postposition. Son exactitude est $\mu = 71,9$ ($\sigma = 0,018$). Ses capacités de prédiction sont exposées dans le tableau 2 : il prédit à 100% la postposition, mais est incapable de prédire l'antéposition.

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Syntaxe
DOI 10.1051/cmlf/2010161

		Position prédite		% de prédiction
		POST	ANT	
Position observée	POST	11015	0	100,00%
	ANT	4309	0	0,00%

Tableau 2 : Matrice de confusion du modèle Nul

La matrice de confusion présentée dans le tableau 2 permet d'estimer plus précisément la qualité de la prédiction : les lignes représentent les données réelles que l'on observe dans le corpus, et les colonnes les données prédites par le modèle. Le croisement des lignes et des colonnes permet de voir quel nombre d'occurrences le modèle prédit correctement dans chaque position et de le comparer au nombre de données mal prédites. Dans le cas du modèle Nul, la colonne « Position prédite – Post(position) » contient l'ensemble des occurrences car ce modèle ne prédit que la postposition. Le croisement avec les lignes représentant les données réelles nous permet de voir que les 4309 occurrences antéposées sont mal prédites par ce modèle.

3 Les variables

Nous présentons l'ensemble des variables à partir desquelles nous étudions la position de l'adjectif dans cet article. L'objectif est de capturer les contraintes issues de la littérature sur la problématique du placement de l'adjectif épithète, à travers des variables qui apparaissent dans notre table de données, soit sous forme de variable binaire (valeur 0 ou 1), soit sous forme de variable continue. Il est important de comprendre ce que représente chaque variable, dans la mesure où les modèles de prédiction, qui seront présentés dans la partie 4, sont construits à partir de ces dernières. Dans cette partie, nous décrirons succinctement chaque variable, puis nous détaillerons leur répartition en antéposition et en postposition, dans nos données.

Dans un travail précédent (Thuilier et al., soumis), nous avons montré que les variables relatives à la combinatoire de l'adjectif[4] n'apportent pas d'information suffisante pour prédire correctement la position des adjectifs d'un point de vue quantitatif. Ainsi, dans le cadre de cet article, nous nous concentrons sur les contraintes concernant l'item lexical (propriétés morphologiques, classes lexicales, longueur et fréquence) ainsi que sur les données relatives à la longueur du syntagme adjectival (Sadj) et du nom combiné avec l'adjectif.

Les variables sont présentées en trois groupes : premièrement, les variables qui ont trait aux propriétés lexicales de l'adjectif, c'est-à-dire aux informations de nature linguistique qui apparaissent dans l'entrée lexicale ; deuxièmement, les variables renvoyant aux caractéristiques ayant une influence sur le traitement cognitif des éléments utilisés (longueur et fréquence) ; troisièmement, les variables portant sur les préférences de chaque item pour une position.

3.1 Variables concernant les propriétés lexicales de l'adjectif

Variable concernant la morphologie de l'adjectif (DERIVE). Certains adjectifs sont issus d'une autre partie du discours. Ils peuvent dériver de verbes (participes passés, participes présents, suffixes -ible (prédictible)/-able (faisable)/-uble (soluble)/-if (attractif)) ou de noms (métallique, présidentiel, scolaire). Ces adjectifs sont généralement décrits comme préférant la postposition.

La variable dans la table : DERIVE est égale à 1 quand l'adjectif est dérivé d'un nom ou d'un verbe, 0 autrement.

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Syntaxe
DOI 10.1051/cmlf/2010161

Variables concernant des classes lexicales (NATIO ; COULEUR ; ADJ-INDEF). La plupart des grammaires de référence (Riegel *et al.*, 1994 ; Grevisse & Goose, 2007) donnent comme règle générale que les adjectifs à caractère objectif (i.e dont le sémantisme est perceptible ou inférable à partir de l'observation directe) sont postposés au nom. Ces adjectifs sont divisés en classes plus précises : forme, couleur, description de propriété physique, nationalité, technique... Nous ne prenons en compte ici que deux classes, la nationalité et la couleur, à titre d'exemple du caractère prédicteur qu'elles peuvent avoir. Par ailleurs, les adjectifs appartenant à la classe des indéfinis ont un comportement particulier, car ce sont des éléments qui partagent des caractéristiques à la fois avec les adjectifs et avec les déterminants. Nous avons choisi de le prendre en considération dans la mesure où ils peuvent être en co-occurrence avec un déterminant et apparaître en antéposition ou en postposition. Les adjectifs identifiés comme appartenant à cette classe dans notre corpus sont : *tel*, *autre*, *certain*, *quelques*, *divers*, *différent*, *maint*, *nul*, *quelconque*, *même*.

Les trois variables dans la table de données : NATIO est égale à 1 quand l'adjectif désigne une nationalité, 0 autrement ; COUL est égale à 1 lorsque l'adjectif dénote une couleur, 0 autrement ; ADJ-INDEF est égale à 1 quand l'adjectif appartient à la classe des indéfinis, 0 autrement.

3.1.1 Les variables concernant les propriétés lexicales dans la table de données

Les adjectifs de couleur (COULEUR = 1) n'apparaissent qu'en postposition, et seule une occurrence d'adjectif de nationalité (NATIO = 1) apparaît en antéposition. Cela signifie que ces deux classes sémantiques ont une très forte préférence pour la postposition. Les adjectifs indéfinis (ADJ-INDEF = 1), quant à eux, se présentent à 87,2% en antéposition. Enfin les adjectifs issus de dérivation morphologique ont une forte préférence pour la postposition, puisqu'ils sont à plus de 91% postposés. Le tableau 3 montre cependant que les variables renvoyant aux informations lexicales ne concernent qu'un nombre restreint de données, au maximum un cinquième du corpus pour la variable DERIVE.

	NATIO = 1		COULEUR = 1		ADJ-INDEF = 1		DERIVE = 1	
Antéposé	1	*0,06%*	0	*0%*	628	*87,2%*	266	*8,5%*
Postposé	1754	*99,94%*	63	*100%*	98	*12,8%*	2874	*91,5%*
Totaux	1755	*100%*	63	*100%*	720	*100%*	3140	*100%*
Proportions par rapport aux données globales	11,5%		0,4%		4,7%		20,5%	

Tableau 3 : Répartition antéposé/postposé pour les variables relatives à l'information lexicale.

3.2 Variables relatives à la longueur syllabique et à la fréquence

Dans de nombreux travaux sur l'ordre des mots, la longueur est envisagée comme un facteur à prendre en compte : pour les adjectifs du français (Forsgren, 1978 ; Wilmet, 1981), pour les alternances de mots (Cooper & Ross, 1975 ; Benor & Levy, 2006) et pour les alternances de constituants dans d'autres langues (Hawkins, 2000 ; Wasow, 2002 ; Rosenbach, 2005 ; Bresnan *et al*., 2007). L'idée centrale peut être résumée par le principe *court avant long* : les éléments les plus courts ont tendance à être placés avant les plus longs. Nous envisageons la longueur en termes de syllabes et nous la déclinons en trois variables distinctes : la longueur absolue de l'adjectif, la longueur absolue du Sadj et la longueur relative de l'adjectif par rapport au nom.

Longueur en syllabes de l'adjectif seul (ADJ-SYLL). Wilmet (1981) constate, dans son étude sur corpus, que les cent adjectifs les plus fréquents sont majoritairement monosyllabiques et, parallèlement, que les adjectifs les plus fréquents sont généralement antéposés. Il en conclut qu'il existe probablement

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Syntaxe
DOI 10.1051/cmlf/2010161

une relation entre longueur et position. Il semblerait donc que plus les adjectifs sont courts, plus ils ont tendance à être antéposés.

La variable dans la table de données : ADJ-SYLL est égale au nombre de syllabes que contient l'adjectif[5].

Longueur en syllabes du Sadj (SADJ-SYLL). Etant donné que le Sadj peut contenir d'autres éléments que l'adjectif tels qu'un modifieur, un complément ou une coordination d'adjectifs, on peut penser que c'est la longueur du syntagme, plutôt que de l'adjectif seul, qui est pertinente pour décider de la position du Sadj et donc de l'adjectif.

La variable dans la table de données : SADJ-SYLL est égale au nombre de syllabes que contient le Sadj.

Longueur relative de l'adjectif par rapport au nom (A-N-SYLL). La tendance généralement observée est que les éléments les plus courts précèdent les plus longs. Forsgren (1978) fait cette observation pour les couples adjectif/nom. On s'attend donc à la postposition pour un adjectif plus long que le nom, et à l'antéposition pour un adjectif plus court que le nom.

La variable dans la table de données : A-N-SYLL est égale à la différence du nombre de syllabes entre l'adjectif et le nom. Une valeur positive signifie que l'adjectif est plus long que le nom et, inversement, une valeur négative indique que l'adjectif est plus court que le nom.

Fréquence de l'adjectif (FREQ). Comme mentionné précédemment, Wilmet (1981) observe que la fréquence élevée est corrélée à l'antéposition. Afin de vérifier ce constat dans notre corpus, nous avons compté le nombre d'occurrences par lemme.

La variable dans la table de données : FREQ est égale au nombre d'occurrences du lemme dans nos données. Par exemple, l'adjectif *allemand* apparaît 174 fois dans nos données. La valeur de FREQ pour l'ensemble des occurrences de ce lemme est donc 174.

3.2.1 Fréquence et longueur dans la table de données

Les variables de longueur montrent toutes les trois le même type de tendance et correspondent à ce que nous attendions, c'est-à-dire qu'elles suivent le principe *court avant long*.

Par exemple, pour la longueur de l'adjectif (ADJ-SYLL), plus l'adjectif est court plus il a tendance à être antéposé. La valeur pivot autour de laquelle la répartition antéposé/postposé semble s'organiser est 2 syllabes. En effet, les adjectifs de moins de 2 syllabes sont, en large majorité, antéposés, et représentent 87,1% des adjectifs dans cette position, alors que ceux de plus de 2 syllabes sont, en quasi-totalité, postposés (cf. tableau 4).

	Adjsyll<2		Adjsyll = 2		Adjsyll>2	
ANT	1847	*72,0%*	1906	*36,7%*	556	*7,4%*
POST	720	*28,0%*	3288	*63,3%*	7007	*92,6%*
Totaux	2567	*100%*	4194	*100%*	7563	*100%*

Tableau 4 : Répartition antéposé/postposé selon la longueur de l'adjectif en syllabes.

En ce qui concerne la fréquence, la tendance qui se dessine est moins prononcée que pour la longueur syllabique. Cependant, comme le montre la figure 2, on observe que la probabilité d'antéposition de l'adjectif augmente avec la fréquence. Lorsque la fréquence est autour de 1, la probabilité d'antéposition est inférieure à 0,2, tandis que quand la fréquence est supérieure à 400, la probabilité est supérieure à 0,6.

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Syntaxe
DOI 10.1051/cmlf/2010161

Cela montre donc que la fréquence et la position de l'adjectif entretiennent une relation telle que : plus la fréquence est élevée dans la table de données, plus l'adjectif a tendance à être antéposé.

3.3 Variables de préférences pour une position

Afin de compléter les informations dont nous disposons sur l'item adjectival (NATIO, COUL, ADJ-INDEF, DERIVE), nous avons approximé les caractéristiques lexicales de chaque lemme. Ces approximations ont pour but de détecter si l'adjectif a une préférence statistique pour l'une ou l'autre position. Elles ont été calculées sur la table de données.

Nous avons élaboré deux dictionnaires à partir de la table de données : un dictionnaire des adjectifs anormalement antéposés et un dictionnaire des adjectifs anormalement postposés. Un membre du dictionnaire antéposé (respectivement postposé) est un lemme pour lequel le nombre de positions observées (antéposition ou postposition) est significativement différent (au seuil $\alpha = 0.05$) du nombre de positions que l'on peut attendre théoriquement sur la base d'une loi binomiale[6].

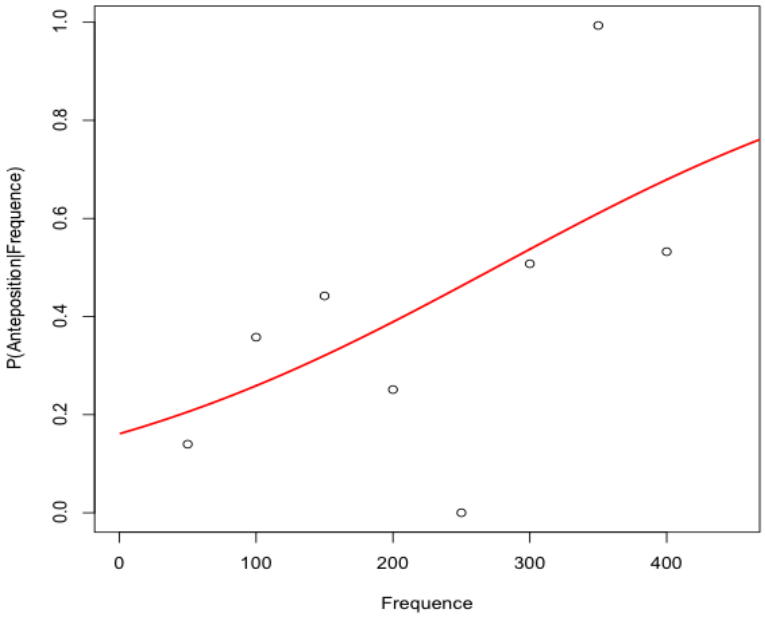

Figure 2 : Courbe représentant la probabilité que l'adjectif soit antéposé étant donné sa fréquence dans la table de données.

De ces deux dictionnaires, on dérive deux traits dont les valeurs sont définies comme suit :

– PREF-ANT : l'adjectif appartient au dictionnaire des adjectifs anormalement antéposés 1 et 0 sinon.

– PREF-POST : l'adjectif appartient au dictionnaire des adjectifs anormalement postposés 1 et 0 sinon.

Dans la table de données, 7502 occurrences adjectivales (49%) ont une valeur positive pour A-PREF-POST et 3851 (25,1%) pour A-PREF-ANT. Ces occurrences représentent respectivement 304 et 49 lemmes adjectivaux.

4 Modèles de prédiction

Dans cette partie, nous présentons le modèle de prédiction construit à partir de l'ensemble des variables, puis nous le décomposons en plusieurs sous-modèles afin de mettre en lumière la contribution des différents groupes de variables et tenter de mieux comprendre le phénomène de placement de l'adjectif.

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Syntaxe
DOI 10.1051/cmlf/2010161

Nous observerons ainsi l'apport des contraintes de longueur et de fréquence (modèle Longueur-Fréquence), et nous montrerons l'importance de l'information relative à chaque item adjectival pour prédire correctement la position de l'adjectif.

4.1 Modèle global

Nous décrivons ici le modèle de prédiction global, qui a pour but de prédire la position de l'adjectif à partir de toutes les variables présentées en partie 3. Ce modèle est construit sur les 10 variables et maximisé à l'aide de la procédure AIC[7]. Le modèle obtenu, présenté en figure 3, contient 9 variables, la variable FREQ ayant été éliminée grâce à la procédure AIC. Dans ce modèle, les variables ADJ-INDEF, A-PREF-ANT votent pour l'antéposition étant donné leur coefficient positif. À l'inverse, les coefficients négatifs des variables DERIVE, NATIO, COULEUR et A-PREF-POST indiquent que ces dernières votent pour la postposition.

La variable SADJ-SYLL vote pour la postposition de façon plus ou moins forte selon que la valeur de la variable est plus ou moins importante. De plus, contrairement à ce que nous attendions, la variable ADJ-SYLL vote pour l'antéposition de telle façon que plus l'adjectif est long, plus l'antéposition est favorisée. Cette contradiction apparente s'explique par le fait que la variable ADJ-SYLL est un contrepoids à la variable SADJ-SYLL. En effet, la valeur de la variable ADJ-SYLL associée à son coefficient s'ajoute à la valeur de SADJ-SYLL multipliée par son coefficient. Le tout reste négatif car le Sadj est au minimum aussi long que l'adjectif (SADJ-SYLL ≥ ADJ-SYLL) et le coefficient de SADJ-SYLL est supérieur à celui de ADJ-SYLL. Ainsi, le groupement SADJ-SYLL et ADJ-SYLL vote pour la postposition.

Enfin, la variable DIFF-SYLL-A-N peut prendre des valeurs négatives ou positives. Cela implique que la préférence de cette variable dépend du signe de la valeur de la contrainte : lorsque la longueur du nom est supérieure à celle de l'adjectif (valeur négative pour DIFF-SYLL-A-N), la contrainte vote pour l'antéposition ; lorsque l'adjectif est plus long que le nom (valeur positive pour DIFF-SYLL-A-N), la contrainte vote pour la postposition. Cela est conforme à ce que nous attendions.

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

+0,014	
+0,14	ADJ-SYLL
−0,57	SADJ - SYLL
−0.29	DIFF-SYLL-A-N
−0.35	DERIVE = 1
+0,91	ADJ-INDEF = 1
−3,47	NATIO = 1
−14,24	COULEUR = 1
+2,81	A-PREF-ANT = 1
−2,60	A-PREF-POST = 1

Figure 3 : Formule du modèle global

Ce modèle a une exactitude $\mu = 91{,}8\%$ ($\sigma = 0{,}008$), ce qui est très largement supérieur à l'exactitude du modèle Nul (71,9%). Le tableau 5 montre plus en détail ses capacités de prédiction. On observe qu'il est capable de prédire correctement 81,3% des antéposés et 95,8% des postposés.

Afin de mieux comprendre quel est l'apport des différentes contraintes, nous avons construit des modèles autour de groupes de contraintes plus restreints.

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Syntaxe
DOI 10.1051/cmlf/2010161

		Position prédite		% de prédiction
		P	A	
Position observée	P	10554	456	95,8%
	A	806	3503	81,3%

Tableau 5 : Matrice de confusion du modèle de prédiction global

4.2 Modèle basé sur les propriétés lexicales de l'adjectif

Le modèle basé sur les propriétés lexicales (désormais modèle Lexical) permet de prédire la position de l'adjectif en fonction des variables concernant la morphologie (DERIVE) et les classes lexicales (NATIO, COULEUR et ADJ-INDEF).

Le modèle contient les 4 variables et a une exactitude de $\mu = 75{,}4\%$ ($\sigma = 0.019$). Ainsi, les variables permettent de construire un modèle dont les capacités de prédiction sont sensiblement meilleures que celles du modèle Nul. Il semble donc que les classes d'adjectifs retenues sont pertinentes pour le choix de la position de l'adjectif

		Position prédite		% de prédiction
		P	A	
Position observée	P	10920	95	99,1%
	A	3681	628	14,6%

Tableau 6: Matrice de confusion du modèle Lexical

4.3 Modèle Longueur-Fréquence de l'adjectif

Dans cette partie, nous construisons un modèle qui prédit la position de l'adjectif à partir de la longueur et de la fréquence de l'adjectif, afin de montrer que ces deux variables ont un impact sur le choix de la position. Le modèle Longueur-Fréquence contient les variables FREQ et ADJ-SYLL. Comme le montre la formule de ce modèle (figure 4), la variable FREQ vote pour l'antéposition, et la variable ADJ-SYLL vote pour la postposition, et ce de façon plus ou moins forte selon la valeur de chacune de ces deux variables.

L'exactitude de ce modèle est $\mu = 80{,}7\%$ ($\sigma = 0{,}010$). Le tableau 7 indique que ces deux variables permettent de prédire correctement plus de la moitié des antéposés et près de 90% des postposés. L'apport de ces variables dans le choix de la position est donc plus important, en termes quantitatifs, que celui des 4 variables du modèle Lexical.

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

+1.56
–1.33 ADJ-SYLL
+0.005 FREQ

Figure 4 : Formule du modèle Longueur-Fréquence

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Syntaxe
DOI 10.1051/cmlf/2010161

		Position prédite		% de prédiction
		P	A	
Position observée	P	9846	1169	89,4%
	A	1785	2524	58,6%

Tableau 7 : Matrice de confusion du modèle Longueur-Fréquence

Ce modèle montre l'importance de ces deux contraintes dans le phénomène de placement de l'adjectif épithète. Les deux facteurs renvoient à un problème de traitement cognitif. D'abord, la longueur est en général un signe de complexité : plus le mot est long, plus son articulation et sa reconnaissance sont complexes (Cooper & Ross, 1975 ; Pinker & Birdsong, 1979). La tendance est alors de placer les éléments les plus simples en premier pour faciliter le traitement de l'ensemble de mots envisagés. D'un point de vue quantitatif, le placement des adjectifs semble respecter cette tendance générale. En ce qui concerne la fréquence, un nombre important de travaux a montré qu'elle joue un rôle dans la représentation cognitive des mots (voir notamment Bybee, 2006). Ainsi plus la fréquence d'un mot est élevée, plus son accès dans le lexique mental des locuteurs est rapide. Notre travail fait apparaître la tendance selon laquelle plus un adjectif est fréquent, et donc accessible, plus il a tendance à être antéposé.

Par ailleurs, des travaux (notamment Zipf, 1932 ; Fenk-Oczlon, 1989) ont montré que la longueur et la fréquence des mots entretiennent une relation inverse : plus la fréquence d'un mot est haute, plus ce mot à tendance à être court. Cela signifierait que les deux contraintes, fréquence et longueur, sont redondantes et qu'une seule suffirait dans la prédiction de l'ordre des mots. Cependant, dans nos données, la corrélation entre la fréquence et la longueur des adjectifs est très faible (coefficient de corrélation rho de Spearman[8] $r_s = -0{,}15$ (p.value < 0,001)). Les deux variables sont donc pertinentes et, comme le montre le modèle Longueur-Fréquence, elles ont une influence non négligeable sur le choix de la position de l'adjectif.

Les variables de longueur et de fréquence concernent l'item lexical et l'usage que les locuteurs en font (notamment la fréquence d'apparition). Les bonnes capacités de prédiction du modèle Longueur-Fréquence ainsi que celles du modèle Lexical vont donc dans le même sens : les informations relatives à chaque item lexical sont centrales, quelle qu'en soit la nature (intrinsèque ou d'usage). Notons que le modèle combinant la longueur, la fréquence et les 4 variables relatives aux propriétés lexicales a une exactitude de $\mu = 84{,}8\%$ ($\sigma = 0{,}011$). La combinaison de ces variables améliore encore la prédiction, ce qui confirme l'importance des caractéristiques lexicales dans le choix de la position de l'adjectif.

Les modèles Longueur-Fréquence et Lexical suggèrent que le phénomène de placement de l'adjectif s'explique par des facteurs de nature différente : d'un côté, des contraintes proprement linguistiques (les classes sémantiques, la dérivation morphologique) et de l'autre, des facteurs relatifs à l'usage et au traitement cognitif. En d'autres termes, les tendances générales observables pour la fréquence et la longueur ne sont pas suffisantes pour choisir correctement la place de l'ensemble des adjectifs, il y a d'autres contraintes qui agissent, parfois en contradiction avec la fréquence et la longueur.

4.4 Modèle basé sur les préférences pour une position

Dans cette section, nous présentons le modèle basé sur les 2 variables de préférence lexicale (désormais modèle Préférence), afin de montrer que l'approximation statistique des préférences est une estimation pertinente qui indique que le composant lexical relatif à chaque item adjectival doit être analysé plus en détail. Le modèle Préférence contient les variables A-PREF-ANT et A-PREF-POST et a une exactitude de $\mu = 91{,}2\%$ ($\sigma = 0{,}007$). Comme le montre le tableau 8, ce modèle prédit correctement l'antéposition de 79% des adjectifs effectivement antéposés.

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Syntaxe
DOI 10.1051/cmlf/2010161

		Position prédite		% de prédiction
		P	A	
Position observée	P	10567	448	95,90%
	A	904	3405	79,00%

Tableau 8 : Matrice de confusion du modèle Préférence

Les capacités de prédiction du modèle Préférence sont très proches de celles du modèle global. Ce sont donc les deux variables approximant les préférences qui permettent en grande partie la prédiction dans le modèle global. Ces approximations statistiques mettent en lumière que chaque item lexical semble avoir une préférence pour une position et que nous prenons cette préférence en compte dans nos choix de placement de l'adjectif. Cela suggère que la dimension lexicale que nous avons prise en compte (DERIVE, NATIO, COULEUR, ADJ-INDEF, ADJ-SYLL et FREQ) n'est pas suffisante pour bien décrire le choix de la position de l'adjectif. L'enjeu est alors de comprendre ce que capture les approximations A-PREF-ANT et A-PREF-POST. On peut penser que les classes lexicales jouent un rôle important dans l'élaboration des préférences. Il faudrait donc étendre les variables du type NATIO et COULEUR à d'autres classes lexicales à l'aide de dictionnaires appropriés.

La comparaison des modèles montre que l'essentiel de l'information nécessaire pour une prédiction correcte à près de 92% est une information relative à chaque item adjectival. L'étude des erreurs du modèle global, c'est-à-dire des 8% de données prédites de façon incorrecte par le modèle global, laisse apparaître qu'il faut aussi prendre en compte des informations relatives à l'item nominal et à d'autres éléments présents dans la structure. En effet, nous observons que le modèle n'est pas en mesure de capturer les effets de figements, tels que *(à) juste* titre ou *libre échange*. De plus, il ne rend pas compte de constructions comme *l'été dernier/ le mois dernier/ la semaine dernière*, où l'adjectif postposé *dernier* se combine avec un nom exprimant un notion de temps et un article défini. Ces exemples suggèrent que, pour mieux capturer et modéliser le problème de la position de l'adjectif, il faut tenir compte de l'item nominal et de la structure spécifique dans laquelle l'adjectif est instancié. Pour cela, nous envisageons d'utiliser des outils déjà existants, comme le calcul de collocations (Manning & Schütze, 1999) et de collostructions (Gries, 2003) sur de grands corpus.

5 Conclusion

À l'aide d'une approche quantitative sur corpus, nous avons montré que le problème de placement de l'adjectif épithète en français est un problème qui relève en grande partie des caractéristiques de chaque item adjectival. Nous avons souligné l'importance des facteurs d'usage, longueur et fréquence, qui permettent de prédire la position de plus de 80% des adjectifs de nos données. Ce travail sur le choix de la position de l'adjectif doit être développé en affinant la dimension lexicale qui a été approximée à l'aide des variables A-PREF-ANT et A-PREF-POST. Il doit également l'être par la prise en compte de l'item nominal et de la construction dans laquelle l'adjectif apparaît, comme le fait apparaître l'étude des erreurs du modèle global. Notre travail montre également qu'il est possible de modéliser le choix effectif de la place de l'adjectif, grâce à une approche probabiliste et de données annotées en syntaxe. Cela appuie l'hypothèse que nous avons formulée en introduction, à savoir que le choix de la position de l'adjectif repose en grande partie sur des contraintes préférentielles. De plus, notons que la modélisation de la position de l'adjectif est relativement satisfaisante malgré la restriction au niveau de la sémantique liée à la position. Les tendances générales régissant le phénomène étudié peuvent donc être capturées au niveau de la forme.

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Syntaxe
DOI 10.1051/cmlf/2010161

Plus généralement, notre travail laisse ouverte la question de la représentativité et de la généralisation. Le corpus sur lequel nous avons mené cette étude est un corpus journalistique, ce qui ne satisfait pas de critères de représentativité nous permettant de généraliser nos conclusions au français. Cependant, nous considérons que les tendances générales dégagées ne seront pas invalidées par la variation entre les différents genres de corpus. À titre d'exemple, Bresnan *et al* (2007) modélisent l'alternance dative en anglais à partir d'un corpus oral composé de conversations téléphoniques et d'un corpus journalistique, et les modèles de régression ne varient que légèrement selon le genre du corpus. Il serait intéressant d'étudier le phénomène sur d'autres corpus afin de permettre une véritable généralisation et d'affiner la description du phénomène.

Bibliographie

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

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

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

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19(6)*, 716–723.

Benor, S. B. & Levy, R. (2006). The chicken or the egg ? a probabilistic analysis of english binomials. *Language*, *82(2)*, 28–55.

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

Bybee, J. (2006). From usage to grammar : the mind's response to repetition. *Language*, *82(4)*, 711-733.

Cooper, W. E. & Ross, J. R. (1975). World order. In Grossman, R. San, L. & Vance, T. (éd.), *Papers from the Parasession on Functionalism*, Chicago : Chicago Linguistic Society, 63–111.

Fenk-Oczlon, G. (1989). Word frequency and word order in freezes. *Linguistics*, *27*, 517–556.

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

Grevisse M. & Goosse A. 2007. *Le bon usage.* 14ème édition : De Boeck Université.

Gries, S. T. (2003). Collostructions : Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, *8 :2*, 209–243.

Hawkins, J. (2000). The relative order of prepositional phrases in english : Going beyond manner-place-time. *Language Variation and Change*, *11*, 231–266.

Manning, C. D. & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge : MIT Press.

Namer, F. (2002). Acquisition automatique de sens à partir d'opérations morphologiques en français : étude de cas. *TALN-2002*, 235-244.

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

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

Pinker, S. & Birdsong, D. (1979). Speakers' sensitivity to rules of frozen word order. *Journal of Verbal Learning and Verbal Behavior*, *18(4)*, 497–508.

Riegel M., Pellat J.-C. & Rioul R. 1994. *Grammaire méthodique du français*. PUF.

Rosenbach, A. (2005). Animacy versus weight as determinants of grammatical variation in English. *Language*, *81(3)*, 613–644.

DOI 10.1051/cmlf/2010161

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

Thuilier, J. Fox, G. & Crabbé, B. (soumis). Prédire la position de l'adjectif épithète en français : approche quantitative. Lingvisticae Investigationes.

Wasow, T. (2002). *Postverbal behavior*. Stanford : CSLI publications.

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

---

[1] *http ://pourpre.com/chroma/*

[2] Nous rappelons que, dans notre table de données, nous n'avons pas d'information concernant la sémantique des adjectifs. Cela signifie que les différences de sens liées à la position ne sont pas identifiables automatiquement. Nous ne pouvons donc pas estimer le pourcentage de données qu'elles représentent.

[3] Formellement, une fonction logistique est une fonction à valeurs dans l'intervalle [0,1] et dont la forme analytique est la suivante :

$$\pi_{ante} = \frac{e^{\beta X}}{1 + e^{\beta X}}$$

où π_{ante} s'interprète comme la probabilité d'antéposition de l'adjectif et β abrège la séquence de coefficients de régression α, $\beta_0 ... \beta_n$ associés respectivement à chacune des variables prédictrices $X_0 ... X_n$ notées X. Pour un nuage de points donné, le calcul de la régression consiste à estimer par maximum de vraisemblance les paramètres α (coefficient d'interception) et β_i de chaque variable du modèle dans un espace *logit*.

[4] Variables combinatoires prises en compte dans Thuilier *et al.* (soumis) :

- configuration du Sadj : coordination de l'adjectif, présence d'un adverbe modifiant l'adjectif, présence d'un dépendant post-adjectival ;
- configuration du SN : présence d'autres adjectifs, présence d'un SPrep, présence d'une subordonnée relative ;
- nature du déterminant du SN.

Notons que la variable relative à la présence d'un dépendant post-adjectival renvoie à la seule contrainte catégorique intervenant dans le phénomène du placement de l'adjectif épithète : un adjectif ayant un dépendant post-adjectival est obligatoirement postposé, et cela est vérifié dans notre corpus. Cependant, la proportion des occurrences présentant une valeur 1 pour cette variable étant très faible (2,9% des données), cette dernière ne permet pas de faire des prédictions correctes sur l'ensemble du corpus.

[5] Le nombre de syllabes de certains adjectifs est un nombre décimal car il correspond à la moyenne pondérée des valeurs données par le syllabateur ELITE. En effet, pour certains mots et en fonction du contexte, le nombre de syllabes peut varier pour un même lemme.

[6] Plus formellement, soit *n* le nombre d'occurrences de l'adjectif en corpus et *k* le nombre de fois où il est antéposé (resp. postposé), l'adjectif appartient au dictionnaire des adjectifs anormalement antéposés (resp. postposés) si $P(K \geq k) < 0.05$ où

$$P(K \geq k) = \sum_{i=k}^{n} \binom{n}{k} p^k (1-p)^{n-k}$$

avec une probabilité théorique $p = 0{,}5$. Concrètement, cela signifie qu'un adjectif présentant 6 occurrences de postposition et aucune occurrence d'antéposition est marqué comme ayant une préférence pour la postposition. Les adjectifs ayant moins de 6 occurrences ne peuvent pas avoir une préférence marquée pour une position, car la fréquence n'est pas assez élevée pour évaluer si la distribution de position est différente de la distribution théorique.

Neveu F., Muni Toke V., Durand J., Klingler T., Mondada L., Prévost S. (éds.)
Congrès Mondial de Linguistique Française - CMLF 2010
978-2-7598-0534-1, Paris, 2010, Institut de Linguistique Française

Syntaxe
DOI 10.1051/cmlf/2010161

[7] La procédure AIC permet d'identifier le modèle le plus compact, c'est-à-dire qui maximise la vraisemblance des données en minimisant le nombre de variables. Plus précisément, la procédure de sélection utilisée est une procédure par élimination arrière dirigée par une heuristique AIC (Akaike Information Criterion) (Akaike, 1974).

[8] Le coefficient r_s est compris entre -1.00 et 1.00. Une valeur de 1.00 ou -1.00 indique une corrélation parfaite. Une valeur de 0 indique qu'il n'y a aucune corrélation.