TALN 2010, Montréal, 19–23 juillet 2010

Approche quantitative en syntaxe : l'exemple de l'alternance de position de l'adjectif épithète en français

J. Thuilier⋆, G. Fox◇, B. Crabbé⋆
⋆ Université Paris VII Denis Diderot ◇ Université Paris III Sorbonne-Nouvelle
UFRL et INRIA (Alpage) ILPGA et EA 1483

Résumé. Cet article présente une analyse statistique sur des données de syntaxe qui a pour but d'aider à mieux cerner le phénomène d'alternance de position de l'adjectif épithète par rapport au nom en français. Nous montrons comment nous avons utilisé les corpus dont nous disposons (French Treebank et le corpus de l'Est-Républicain) ainsi que les ressources issues du traitement automatique des langues, pour mener à bien notre étude. La modélisation à partir de 13 variables relevant principalement des propriétés du syntagme adjectival, de celles de l'item adjectival, ainsi que de contraintes basées sur la fréquence, permet de prédire à plus de 93% la position de l'adjectif. Nous insistons sur l'importance de contraintes relevant de l'usage pour le choix de la position de l'adjectif, notamment à travers la fréquence d'occurrence de l'adjectif, et la fréquence de contextes dans lesquels il apparaît.

Abstract. This article presents a statistical analysis of syntactic data that aims to better understand the phenomenon of position alternation displayed by attributive adjectives with respect to nouns in French. We show how we used the corpora available for French (the French Treebank and the Est-Républicain corpus) as well as ressources provided by Natural Language Processing for our study. The proposed model contains 13 variables based on properties of the adjectival phrase, the adjectival item and on frequency constraints. This model is capable to predict the position of adjectives at more than a 93% rate. We especially focus on the importance of constraints based on usage for the choice of position for the adjective, in particular the frequency of contexts in which it appears.

Mots-clés : Syntaxe probabiliste, linguistique de corpus, adjectif épithète, régression logistique.

Keywords: Probabilistic syntax, corpus linguistics, attributive adjective, logistic regression.

1 Introduction

La position de l'adjectif épithète par rapport au nom en français est un problème bien connu en linguistique française. Ce phénomène met en jeu plusieurs dimensions linguistiques : phonologie, morphologie, syntaxe, sémantique et discours (voir notamment les travaux de (Forsgren, 1978; Wilmet, 1981; Nølke, 1996; Noailly, 1999; Abeillé & Godard, 1999)). Cependant, à notre connaissance, il n'existe pas de travaux qui rendent compte de l'interaction des contraintes qui sont associées à chacune des dimensions et qui permettent de déterminer la position de l'adjectif dans un contexte précis. Une seule contrainte impose catégoriquement une position spécifique à l'adjectif : la présence d'un dépendant post-adjectival. Les autres contraintes participant à l'alternance sont préférentielles, dans la mesure où elles n'expriment que des tendances.

Le but de notre travail est de quantifier ces contraintes préférentielles et de proposer une modélisation qui rende compte au mieux de la position des adjectifs à partir de contraintes interprétables. En étudiant les contraintes préférentielles, nous faisons l'hypothèse, à la suite de (Bresnan *et al.*, 2007), que les données qui relèvent de l'usage font partie de la connaissance de la langue et que les phénomènes d'ordre des mots ne sont pas uniquement guidés par des contraintes qui décident de la grammaticalité d'une phrase.

Pour rendre compte des contraintes préférentielles intervenant dans le choix de la position de l'adjectif épithète, nous présentons une étude quantitative s'appuyant sur le corpus French Tree Bank (désormais FTB) et le corpus de l'Est-Républicain (désormais ER). Cette étude quantitative s'appuie sur la construction d'un modèle de prédiction capable de décider de la position de l'adjectif étant donné un certain nombre de variables. Nous utilisons le caractère opératoire des modèles de prédiction pour évaluer l'importance de groupes de contraintes. L'un des problèmes posés par ce type d'approche en syntaxe est la nécessité d'un volume important de données richement annotées, telles que celles utilisées par (Wasow, 2002; Bresnan *et al.*, 2007) pour l'anglais. Cet article est l'occasion de montrer comment combiner différentes ressources et outils existants pour obtenir des informations quantitatives sur un problème de syntaxe du français.

Dans la section 2, nous décrirons les corpus et la méthode utilisés. Dans la section 3, les variables retenues seront décrites et nous présenterons les ressources et les outils utilisés pour obtenir ces variables. Dans la section 4, nous détaillerons le modèle de prédiction de la position de l'adjectif et nous discuterons l'apport des variables relatives à la fréquence afin de montrer l'importance de l'usage dans le phénomène étudié.

2 Méthodologie

French Treebank (FTB) Ce travail s'appuie sur diverses sources de données. La principale est le corpus arboré de Paris 7 Abeillé *et al.* (2003) dont nous utilisons la sous-partie annotée en constituants et en fonctions syntaxiques (Abeillé & Barrier, 2004). Disponible depuis 2003, il est le résultat d'un projet d'annotation supervisée d'articles du journal *Le Monde*, mené à l'université de Paris 7. Les annotations qui nous intéressent sont de type morphologique et syntaxique. L'annotation morphologique associe chaque forme à une des 13 principales catégories grammaticales (trait `cat`) comme par exemple *N* (nom) ou *A* (adjectif). L'annotation syntaxique propose une analyse en constituants de chaque phrase. Notons à ce sujet que l'annotation de mots composés a été neutralisée dans le cadre de notre étude : les mots composés sont traités comme une forme unique non structurée[1]. Le corpus utilisé comporte 12351 phrases pour 385 458 occurrences de formes et pour 24 098 types de mots distincts. Il contient 18 946 occurrences d'adjectifs annotés comme épithète, parmis lesquelles 17762 accompagnent une tête nominale. Une fois écartés les adjectifs numéraux, les adjectifs contenus dans des dates, les abréviations et certaines occurrences problématiques au niveau de l'annotation, il reste 15242 occurrences d'adjectifs épithètes qui constituent nos données de travail. Ces occurrences représentent 1992 types de mots (ou lemmes).

Est républicain (ER) Nous utilisons une version enrichie du corpus ER pour obtenir des comptes de fréquence plus fiable sur de gros volumes. Le corpus utilisé est celui distribué sur le site du CNRTL[2]. Il a été superficiellement nettoyé puis segmenté en utilisant la méthode décrite dans (Candito & Crabbé, 2009) ainsi qu'augmenté d'un jeu d'annotation en lemmes et en parties de discours décrites dans (Crabbé & Candito, 2008). Une annotation en lemmes et en parties de discours est réalisée par une adaptation au

[1] La raison vient du fait que la structure en constituants et l'annotation morphologique associée aux composants de composés dans le corpus est plus pauvre que l'annotation des formes standard. Il n'aurait pas été évident d'extraire des données à partir des composants sans risquer d'induire un nombre important d'erreurs.

[2] `http ://www.cnrtl.fr/corpus/estrepublicain/`

français de la méthode d'annotation conjointe décrite par (Chrupała *et al.*, 2008)[3] reposant sur une classification par perceptrons. Le corpus ainsi enrichi comporte environ 148 000 000 occurrences de mots pour environ 811 000 formes de mots qui se répartissent en 662 000 lemmes distincts. Parmi ceux-ci, on recense environ 10 000 000 d'occurrences et 57500 lemmes d'adjectifs. Sur base de ce corpus, nous avons considéré comme adjectifs en position épithète les mots étiquetés adjectifs adjacents aux mots étiquetés noms communs.

L'analyse que nous présentons s'appuie sur une table de données statistiques qui contient les 15242 occurrences d'adjectifs épithètes extraites du FTB et auxquelles est associée une variable prédite représentant la position dans laquelle chaque occurrence apparaît (antéposition ou postposition). Chacune de ces occurrences est caractérisée par 13 variables prédictrices qui ont été obtenues à l'aide d'outils issus du traitement automatique des langues et dont la sémantique est détaillée en section 3.

Inférence statistique L'analyse qui suit s'appuie sur l'étude des propriétés de modèles de régression logistique multivariée[4]. Formellement, un tel modèle permet de prédire π la probabilité de succès d'une variable binaire, appelée *variable prédite*, en fonction de n variables continues ou binaires $X_0 \dots X_n$, appelées *variables prédictrices*. Dans ce qui suit, nous posons que la variable binaire prédite comporte deux valeurs : *antéposition* et *postposition* de l'adjectif. Le *succès*, noté *ante*, correspond à l'antéposition et l'*échec* à la postposition. Dans notre cas, un modèle de régression logistique multivarié est un modèle de la forme :

$$\pi_{\text{ante}} = \frac{e^{\alpha+\beta_0 X_0+\ldots+\beta_n X_n}}{1 + e^{\alpha+\beta_0 X_0+\ldots+\beta_n X_n}} \quad (1)$$

où π_{ante} représente la probabilité d'antéposition étant données les variables prédictrices. Entraîner un modèle revient à spécifier les variables prédictrices et à calculer les valeurs des coefficients α et β par maximum de vraisemblance sur des données dans un espace logit (Agresti, 2007). En pratique, le calcul est réalisé par l'algorithme d'optimisation Newton-Raphson.

On interprète un coefficient positif (resp. négatif) β_i associé à une variable X_i comme un coefficient qui vote pour le succès (resp. échec). Les valeurs numériques des coefficients ne peuvent pas être comparées telles quelles, par contre on peut tester (avec un test de Wald) si un coefficient β_i est significativement différent du coefficient nul[5]. Les coefficients significatifs sont ceux que l'on cherche à interpréter. On peut interpréter les valeurs e^{β_i} comme des rapports de chance. Ainsi le nombre e^{β_i} s'interprète comme la multiplication du rapport de chance pour un incrément unitaire de la variable X_i.

Tel quel, un modèle de régression logistique donne une probabilité de succès pour une variable binaire. On augmente communément le modèle d'une procédure de décision, ce qui permet ainsi de le rendre

[3]Ce travail a été réalisé par D. Seddah et G . Chrupala qui ont adapté leur étiqueteur au français. L'étiqueteur est état de l'art et a une exactitude en parties de discours proche de 98% (D. Seddah pc.). Pour la lemmatisation, l'algorithme de (Chrupała *et al.*, 2008) repose sur des scripts d'édition ce qui entraîne quelques erreurs marginales. Par comparaison avec Flemm (Namer, 2000), l'intérêt de ce lemmatiseur est son caractère déterministe.

[4]Pour une introduction détaillée, voir (Agresti, 2007). On préfère utiliser des modèles de régression logistique aux analyses discriminantes et à l'analyse en composantes principales, dans la mesure où ces derniers supposent des hypothèses de normalité sur les données qui ne sont généralement pas satisfaites par les données portant sur le langage naturel. Nous avons par ailleurs testé l'utilisation de méthodes reposant sur l'usage d'arbres de décision tirant parti de l'algorithme CART (Breiman *et al.*, 1984) en nous inspirant de la méthodologie suggérée par (Baayen, 2008). Il est à noter que cet algorithme ne s'est pas révélé d'une granularité suffisante pour permettre une analyse fine des données : la nécessité de partitionner le domaine de variables continues rend les résultats difficilement interprétables en pratique.

[5]Ce test permet de détecter les coefficients qui prédisent une valeur constante. Autrement dit, il permet de détecter les variables qui, prises individuellement, n'influencent pas significativement le résultat de succès ou d'échec.

opérationnel. Ainsi, on pose que le résultat est un succès si $\pi > 0.5$, échec sinon. Cela permet de prédire effectivement la position de l'adjectif par rapport au nom sur des données pour lesquelles la valeur de la variable prédite est inconnue. A fin de comparaison, nous utilisons le caractère opérationnel du modèle de régression en nous appuyant systématiquement sur une évaluation 10 passes dont nous reportons la moyenne (μ) et l'écart-type (σ). Cette évaluation se fait en comparant la position prédite par le modèle et la position effectivement rencontrée dans le corpus FTB.

3 Variables prédictrices

Dans cette section, nous présentons un ensemble des contraintes, synthétisé dans le tableau 1, qui a été proposé dans la littérature comme intervenant dans le phénomène d'alternance de position de l'adjectif épithète[6] et que nous avons pu capturer grâce à des ressources existantes. Il est important de noter que nous restreignons l'objet de notre travail en ne prenant pas en compte les changements de sens liés à la position (comme dans : *une ancienne usine* vs. *une usine ancienne*) pour des raisons pratiques, la non disponibilité de ressources ne permettant pas de capturer ces changements de sens.

Variables	Types	Description
COORD	*booléen*	adjectif en coordination ou non
POST-ADJ	*booléen*	adjectif avec un dépendant post-adjectival ou non
ADJ-LONG	*réel*	longueur de l'adjectif en syllabes
SADJ-LONG	*réel*	longueur du SAdj en syllabes
FREQ	*réel*	fréquence de l'adjectif dans le corpus ER
DERIVÉ	*booléen*	adjectif dérivé ou non
NATIO	*booléen*	adjectif de nationalité ou non
COULEUR	*booléen*	adjectif de couleur ou non
INDEF	*booléen*	adjectif indéfini ou non
COLLOCANT	*réel*	score χ^2 pour le bigramme 'adjectif-nom'
COLLOCPOST	*réel*	score χ^2 pour le bigramme 'nom-adjectif'
ADJ-PREFANT	*booléen*	préférence statistique de l'adjectif pour l'antéposition ou non
ADJ-PREFPOST	*booléen*	préférence statistique de l'adjectif pour la postposition ou non

TAB. 1 – Les variables prédictives, leur type et leur description

Dépendant post-adjectival La présence d'un dépendant post-adjectival est l'unique contrainte catégorique intervenant dans le choix de la position de l'adjectif. Lorsque l'adjectif est suivi d'un complément ou d'un modifieur, il est obligatoirement postposé comme le montre l'exemple 1.

(1) a. un entretien long de deux heures

b. *un long de deux heures entretien

Cette contrainte est capturée par l'intermédiaire de la variable POST-ADJ. On extrait les valeurs de la variable POST-ADJ par reconnaissance de motifs sur les arbres du treebank. Pour cela, nous enrichissons dynamiquement le treebank d'annotations de têtes et de dépendants syntaxiques en réutilisant les tables de propagation de (Arun, 2004). L'adjectif tête d'un AP dépendant du nom ayant un dépendant droit est

[6]Dans un premier travail (Thuilier *et al.*, soumis), nous avons testé un nombre plus important de variables. Les variables présentées ici ont été selectionnées en fonction de leur importance dans le phénomène.

considéré comme un adjectif à complémentation. Dans nos données, les 438 adjectifs qui ont un dépendant post-adjectival sont en postposition, conformément à ce que nous attendions.

Coordination Les approches en termes de compétence du problème de la position de l'adjectif épithète, telle que celle de (Abeillé & Godard, 1999), montrent que la position des adjectifs apparaissant dans une coordination simple n'est pas restreinte, comme le montre l'exemple 2.

(2) une belle et longue table / une table belle et longue

Cependant, dans nos données, 94,8% des occurrences adjectivales en coordination (c'est-à-dire 788 occurrences) sont en postposition. Cela indique que des données basées sur l'usage semblent mettre en avant que la coordination est un facteur favorisant fortement la postposition. Les valeurs de la variable COORD sont extraites par motifs en réutilisant le même principe que pour la variable POST-ADJ.

Classes d'adjectif La position de l'adjectif peut également être influencée par la sémantique lexicale de l'adjectif (par exemple (Grevisse & Goosse, 2007)). Les adjectifs appartenant aux classes sémantiques du type couleur, forme, propriété physique, nationalité, termes techniques..., sont décrits comme postposés. Nous avons pu prendre en compte deux classes sémantiques représentatives. Ce sont deux classes pour lesquelles on dispose de dictionnaires largement exhaustifs : la classe *nationalité* dont on extrait un dictionnaire à partir de ProlexBase (Tran & Maurel, 2006) et la classe *couleur* dont on extrait les valeurs à partir du dictionnaire Chroma[7]. Un adjectif appartenant à l'un de ces dictionnaires sera ainsi marqué par la variable NATIO ou COULEUR. Nous prenons également en considération la classe des adjectifs indéfinis. La définition de cette classe ne repose pas sur la sémantique lexicale, mais sur le comportement syntaxique. En effet, les adjectifs indéfinis ont des caractéristiques d'adjectifs (co-occurrence avec un déterminant, possibilité d'être antéposé ou postposé au nom), mais ils peuvent se comporter comme des déterminants dans certaines circonstances. Ils sont marqués par la variable INDEF[8]. Pour d'autres classes comme la forme ou les propriétés physiques, l'absence de dictionnaire nous contraint à les ignorer.

Morphologie Certains adjectifs 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. Cette contrainte est capturée par la variable DERIVÉ, qui détermine si un adjectif est dérivé de nom ou de verbe grâce à une étape d'analyse morphologique dérivationnelle à l'aide du logiciel DERIF (Namer, 2002). En pratique, nous considérons comme mots dérivés les mots pour lesquels DERIF trouve un lemme de base appartenant au dictionnaire de catégorie Verbe ou Nom.

Longueur La longueur est un facteur souvent utilisé dans les études sur l'ordre des mots ou sur l'ordre des constituants. En ce qui concerne les adjectifs, (Forsgren, 1978; Wilmet, 1981) ont chacun mis à jour, dans une étude sur corpus, la relation entre longueur de l'adjectif et position : les adjectifs les plus courts ont tendance à être antéposés[9]. De plus, étant donné que l'adjectif peut être modifié qu'il soit en antéposition ou en postposition, nous supposons que la longueur du syntagme adjectival peut aussi être pertinente. Nous estimons la longueur de l'adjectif (ADJ-LONG) et du SAdj (SADJ-LONG) en termes de syllabes. Pour cela, nous avons procédé à une syllabation du corpus à l'aide du logiciel industriel de synthèse vocale ELITE qui nous a permis de calculer le nombre de syllabes pour chaque forme de nom et d'adjectif en tenant compte d'effets de liaison. A titre d'indication, nous avons repéré une vingtaine de résultats aberrants produits par

[7] Dictionnaire extrait du web : *http ://pourpre.com/chroma/*

[8] Les adjectifs considérés comme indéfinis sont : tel, autre, certain, quelques, divers, différent, maint, nul, quelconque, même.

[9] Dans (Thuilier *et al.*, soumis), nous avons montré que la longueur absolu de l'adjectif semble être un facteur plus pertinent que la longueur relative entre l'adjectif et le nom.

le syllabateur sur les 15242 occurrences de couples adjectifs/noms sur lesquels nous travaillons.

Fréquence La fréquence de l'item adjectival a aussi été pointée comme un facteur intervenant dans le choix de la position de l'adjectif. (Wilmet, 1981) observe une corrélation entre fréquence élevée du lemme adjectival et antéposition. Dans ce travail, nous estimons la fréquence des lemmes (variable FREQ) dans l'Est-Républicain : on préfère utiliser une grande masse de données même imparfaite pour estimer la fréquence, en nous fondant sur l'hypothèse qu'une grande taille d'échantillon nous permet d'estimer des fréquences qui reflètent mieux la probabilité réelle des mots[10].

Préférences statistiques de l'item adjectival La question de l'alternance a jusqu'ici été envisagée d'un point de vue général pour la catégorie adjectivale : au sein du SN, deux positions sont disponibles pour placer l'épithète. Si l'observation générale du phénomène montre effectivement cette possibilité, celle de l'usage des lemmes spécifiques dans notre table de données nous montre un tout autre état de fait :

	antéposés	*postposés*	*2 positions*	*Totaux*
nombre de lemmes	122	1685	185	1992
	6.1%	*84.6%*	*9.3%*	*100%*

TAB. 2 – Répartition des lemmes du corpus FTB

Comme le montre le tableau 2, sur les 1992 lemmes rencontrés dans le FTB, seuls 185 (9,3%) alternent en position. Les 90,7% restant ont une position fixe. Ceci semble indiquer que, même si les locuteurs ont connaissance des possibilités d'alternance, ils attribuent une position canonique à chaque lemme spécifique et ne se servent des possibilités d'alternance que dans un nombre de cas limité. La contrainte de la fréquence, proposée dans le paragraphe précédent, peut ainsi être restreinte à la fréquence par position. En effet, comme le notent entre autres (Bybee & Mcclelland, 2005; Goldberg, 2006), la fréquence d'occurrence d'un item joue un rôle dans la représentation mentale qu'en a le locuteur. Cependant, les items n'apparaissent pas en isolation, ils sont insérés dans une structure linguistique. Nous ne mémorisons donc pas simplement les occurrences, et ne constituons pas nos représentations mentales à l'aide de l'item seul, nous prenons aussi en compte son contexte d'apparition. En ce qui concerne les épithètes, cela suggère que nous enregistrons la position de l'adjectif par rapport au nom en même temps que son occurrence. Autrement dit, nous aurions deux représentations possibles de chaque adjectif : une en antéposition et une en postposition. Nous effecturions les comptes de fréquence pour chaque représentation séparément. À l'instar de (Gries & Stefanowitsch, 2004) qui étudient des phénomènes d'alternance liés au verbe, nous pensons qu'il y a un biais lexical dans l'alternance des adjectifs : les locuteurs sont sensibles au fait que l'une des deux positions possibles est généralement bien plus fréquente selon l'item. Ainsi, on peut penser que le choix même de l'adjectif est, dans la majorité des cas, un élément suffisant pour prédire sa position.

Afin de tester la validité d'une telle contrainte, nous élaborons deux dictionnaires, l'un regroupant les adjectifs montrant une préférence pour l'antéposition (DA) et l'autre pour ceux qui préférent la postposition (DP). Pour qu'un lemme adjectival soit membre d'un de ces dictionnaires, le nombre d'occurrences dans une position (antéposition ou postposition) doit être significativement différent (seuil $\alpha = 0.05$) du nombre d'occurrences que l'on peut attendre théoriquement sur la base d'une loi binomiale[11]. Nous avons défini 2

[10]Depuis les travaux de (Zipf, 1932), on admet généralement que la longeur et la fréquence des mots entretiennent une relation inverse. Pour nos données, le coefficient de corrélation rho de Spearman est $r_s = -0.44$, ce qui indique qu'il existe bien une corrélation. Cependant, comme nous le verrons dans la partie 4, les deux variables participent significativement au modèle malgré les procédures de maximisation, ce qui signifie que ce sont deux facteurs pertinents.

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

variables : ADJ-PREFANT représente l'appartenance de l'adjectif au DA et ADJ-PREFPOST au DP.

Préférences statistiques de la combinaison nom/adjectif Prolongeant l'idée que l'on mémorise une occurrence dans son contexte d'apparition, on peut penser que, dans certains cas, l'item nominal spécifique influence la position de l'adjectif. Par exemple, l'expression collocative *vibrant hommage* contient un adjectif dont les propriétés, telles que sa nature morphologique, favorisent la postposition (comme dans *voix vibrante*, *ton vibrant*). Or, la présence du nom *hommage* permet de prédire son antéposition en raison de la fréquence de leur apparition en co-occurrence dans cet ordre-là.

Pour obtenir une estimation statistique de la relation d'association entre l'item adjectival et l'item nominal, nous avons utilisé une métrique couramment utilisée pour identifier les collocations : le χ^2 (Manning & Schütze, 1999)[12]. Le calcul des collocations nécessite un gros volume de données, c'est pourquoi nous utilisons le corpus ER. Nous avons procédé à l'extraction des bigrammes de lemmes adjectif-nom et nom-adjectif et créé 2 listes, l'une servant à identifier les collocations avec adjectif antéposé, l'autre pour les collocations avec adjectif postposé. A partir de ces listes, nous avons calculé la valeur de χ^2 de chaque bigramme. Nous avons attribué cette valeur aux variables COLLOCANT (pour les bigrammes à adjectif antéposé) et COLLOCPOST (pour les bigrammes à adjectif postposé). Le rôle de ces variables n'est pas de déterminer si le bigramme constitue une collocation, mais plutôt de donner une estimation de la force de l'association de l'adjectif par rapport au nom (et réciproquement). La table 3 cherche à montrer que le calcul du χ^2 sur l'ER donne un résultat satisfaisant d'un point de vue qualitatif [13].

Adj-Nom	haut niveau ; immédiat après-guerre ; vibrant hommage ; luxuriante végétation ; chaude eau ; vain mot ; lourd tribut ; majeure partie ; courte durée ; profonde tristesse
Nom-Adj	poids lourd ; téléphone portable ; homicides involontaires ; République tchèque ; concurrence déloyale ; personne agée ; an passé ; accident mortel ; sens inverse ; acier inoxidable

TAB. 3 – Bigrammes adjectif-nom et nom-adjectif, présentant le score de χ^2 le plus élevé dans la table de données

4 Modèles de prédiction

Le modèle de prédiction est construit à partir de l'ensemble des variables et maximisé à l'aide d'une procédure par élimination arrière dirigée par une heuristique AIC (Akaike Information Criterion) (Akaike, 1974). Par cette procédure, les variables ADJ-LONG et COORD ont été éliminées car elles ne participent pas de façon significative au modèle. Dans le modèle présenté en figure 1, les variables ayant un coefficient positif votent pour l'antéposition, et celles ayant un coefficient négatif votent pour la postposition. Ces votes correspondent à ce que nous attendions théoriquement. L'exactitude du modèle est μ = 93.1% (σ = 0.007). Le détail des capacités de prédiction est présenté dans le tableau 4. A titre de comparaison, un modèle ne contenant aucune variable prédictrice et prédisant systématiquement la postposition (modèle Nul), a une exactitude μ = 71.9% (σ = 0.018).

$P(K \geq k) = \sum_{i=k}^{n} \binom{n}{k} p^k (1-p)^{n-k}$ avec une probabilité théorique $p = 0.5$.

[12]Nous avons également testé la métrique dérivée du test exact de Fisher jugée plus satisfaisante pour le langage naturel car il ne repose pas sur une hypothèse de normalité(Pedersen, 1996). Cependant, en pratique, sur de très gros corpus, celui-ci impose le calcul de factorielles sur de très grands nombres, ce qui entraîne des problèmes importants d'arrondi.

[13]A l'exception du bigramme *chaude eau* qui obtient une valeur χ^2 très élevée car "Chaude Eau" est un nom propre fréquemment cité dans l'ER.

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

$\beta X =$	$+0.29$		*
	-14.9	POST-ADJ = 1	
	-0.56	SADJ-LONG	***
	$+0.000008$	FREQ	***
	-0.33	DERIVÉ = 1	***
	$+0.55$	INDEF = 1	***
	-3.73	NATIO = 1	***
	-16.24	COULEUR = 1	
	$+0.0001$	COLLOCANT	***
	-0.00005	COLLOCPOST	***
	$+2.24$	ADJ-PREFANT	***
	-1.58	ADJ-PREFPOST	***

FIG. 1 – Formule du modèle de prédiction (effets significatifs codés *** p<0.001, ** p<0.01, * p<0.1)

		Position prédite		% Correct
		P	A	
Position	P	10493	522	95.3 %
observée	A	535	3692	87.3%

Précision totale : $\mu = 93.1\%$ ($\sigma = 0.007$)

TAB. 4 – Matrice de confusion du modèle de prédiction

Afin d'illustrer l'importance de la fréquence d'usage des mots ainsi que de structures plus larges dans le phénomène étudié, observons plus précisément les variables relatives à la fréquence. En s'appuyant sur les discussions présentées dans les parties relatives aux préférences statistiques de la section 3, nous faisons l'hypothèse que la fréquence d'occurrence d'un adjectif (estimée par la variable FREQ), ainsi que sa fréquence d'apparition dans des contextes particuliers (estimée par les variables COLLOCANT, COLLOCPOST, ADJ-PREFANT et ADJ-PREFPOST) influent sur la représentation mentale qu'en a le locuteur. Les capacités de prédiction d'un modèle logistique contenant ces 5 variables relatives à la fréquence, permettraient donc d'estimer la pertinence de l'usage. L'exactitude d'un tel modèle est de $\mu = 92.5\%$ ($\sigma = 0.008$)[14]. La fréquence d'usage semble donc avoir un rôle central dans le choix de la position de l'adjectif épithète. Ce constat semble confirmer l'idée selon laquelle les connaissances des locuteurs dans un phénomène d'alternance sont affectées par l'usage et plus particulièrement la fréquence. Notons que ces 5 variables sont issues du corpus ER. Ainsi, malgré la présence de bruits dans ce corpus et les approximations faites pour le calcul des fréquences, les tendances obtenues grâce à un gros volume de données sont des estimateurs satisfaisants pour un travail sur un problème spécifique de syntaxe.

[14]Modèle maximisé à l'aide d'une procédure par élimination arrière dirigée par une heuristique AIC. A titre de comparaison, le modèle contenant les 8 autres variables (POST-ADJ, COORD, NATIO, COULEUR, INDEF, DERIVÉ, ADJ-LONG, SADJ-LONG), variables qui ne relèvent pas de l'usage, a une exactitude de $\mu = 80.7\%$ ($\sigma = 0.019$).

5 Conclusion

Notre travail s'inscrit dans la lignée des travaux sur des phénomènes d'alternance en anglais (Bresnan, 2007; Wasow, 2002), qui introduisent l'idée de mener des études quantitatives fondées sur des données empiriques richement annotées en syntaxe. Adopter cette vision a deux implications majeures sur les plans théoriques et pratiques. Du point de vue de la théorie, ce type d'approche suppose que l'on considère que l'usage participe à la construction des connaissances langagières des locuteurs d'une langue. La fréquence d'occurrence de formes spécifiques et de structures est jugée comme l'une des propriétés centrales de l'usage affectant la connaissance des locuteurs. L'importance de la contribution dans le modèle de prédiction des contraintes liées à cette notion (FREQ, ADJ-PREFANT, ADJ-PREFPOST, COLLOCANTet COLLOCPOST) montre le rôle certain de l'usage dans le phénomène d'alternance de l'adjectif et tend à confirmer cette vision.

D'un point de vue technique, (Wasow, 2002) note la difficulté de mener à bien ce type de travail pour des langues ou des registres de langues dotés de peu de ressources. C'est le cas du français : nous ne disposons pour l'instant que d'un corpus arboré, qui est de taille trop réduite pour l'observation fiable d'effets de fréquence. Cependant, le fait de le combiner avec d'autres ressources, comme le corpus ER et les dictionnaires, nous a permis de proposer un modèle avec des résultats améliorés et de mieux cerner le problème de l'alternance de l'adjectif épithète. Notre travail montre donc qu'il est envisageable, grâce aux outils et ressources existants, d'adopter une telle approche pour l'étude du français.

Remerciements Les auteurs remercient Richard Beaufort et Sophie Roeckhoudt pour leur aide à la syllabation du corpus, Fiametta Namer pour nous avoir donné accès à son logiciel DERIF, Djamé Seddah, Grezgorz Chrupala et Marie Candito pour leur aide lors du traitement du corpus l'Est Républicain.

Références

ABEILLÉ A. & BARRIER N. (2004). Enriching a french treebank. In *Proc. of LREC'04*, Lisbon, Portugal.

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for french. In *Treebanks*. Dordrecht : Kluwer.

ABEILLÉ A. & GODARD D. (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.

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

ARUN A. (2004). Statistical parsing of the french treebank. Master's thesis, School of Informatics, University of Edinburgh.

BAAYEN R. (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge University Press.

BREIMAN L., FRIEDMAN J. H., OLSHEN R. A. & STONE C. J. (1984). *Classification and Regression Trees*. Belmont : Wadsworth.

BRESNAN J. (2007). Is syntactic knowledge probabilistic ? experiments with the english dative alternation. In S. FEATHERSTON & W. STERNEFELD, Eds., *Roots : Linguistics in Search of Its Evidential Base*, p. 77–96. Berlin : Mouton de Gruyter.

BRESNAN J., CUENI A., NIKITINA T. & BAAYEN. H. (2007). Predicting the dative alternation. In G. BOUME, I. KRAEMER & J. ZWARTS, Eds., *Cognitive Foundations of Interpretation*. Amsterdam : Royal Netherlands Academy of Science.

BYBEE J. & MCCLELLAND J. L. (2005). Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *The Linguistic Review*, **22**, 381–410.

CANDITO M. & CRABBÉ B. (2009). Improving generative statistical parsing with semi-supervised word clustering. In *Proceedings of International Workshop on Parsing Technologies*, Paris.

CHRUPAŁA G., DINU G. & VAN GENABITH J. (2008). Learning morphology with morfette. In *In Proceedings of LREC 2008*, Marrakech, Morocco : ELDA/ELRA.

CRABBÉ B. & CANDITO M. (2008). Expériences d'analyse syntaxique statistique du français. In *Actes de la 15ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN'08)*, p. 45–54, Avignon, France.

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.

GREVISSE M. & GOOSSE A. (2007). *Le bon usage*. 14ème édition : De Boeck Université.

GRIES S. T. & STEFANOWITSCH A. (2004). Extending collostructional analysis : A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics*, **9 :1**, 97–129.

MANNING C. D. & SCHÜTZE H. (1999). *Foundations of statistical natural language processing*. Cambridge : MIT Press.

NAMER F. (2000). Flemm : Un analyseur flexionnel de français à base de règles. In C. JACQUEMIN, Ed., *Traitement automatique des Langues pour la recherche d'information*, p. 523–547. Paris : Hermes.

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.

PEDERSEN T. (1996). Fishing for exactness. In *Proceedings of the South - Central SAS Users Group Conference*, p. 188–200, Austin (TX).

THUILIER J., FOX G. & CRABBÉ B. (soumis). Prédire la position de l'adjectif épithète en français : approche quantitative. *Lingvisticae Investigationes*.

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

WASOW T. (2002). *Postverbal behavior*. 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.

ZIPF G. K. (1932). *Selected studies of the principle of relative frquency in language*. Cambridge : Mass.