Predicting the Position of Attributive Adjectives in the French NP

Gwendoline Fox, Juliette Thuilier

Predicting the Position of Attributive Adjectives in the French NP

Gwendoline Fox[1] and Juliette Thuilier[2]

[1] University of Paris 3 - Sorbonne Nouvelle (ILPGA) and EA 1483
[2] Univ Paris Diderot, Sorbonne Paris Cité, ALPAGE, UMR-I 001 INRIA

1 Introduction

French displays the possibility of both pre-nominal and post-nominal ordering of adjectives within the noun phrase (NP).

(1) un **magnifique** tableau / un tableau **magnifique**
a magnificent painting / a painting magnificient
"a magnificent painting"

While all adjectives may alternate in position, the choice between both orders is not as free as suggested in (1):

(2) a. un **beau** tableau / ??un tableau **beau**
a nice painting / a painting nice
b. un très **beau** tableau / un tableau très **beau**
a very nice painting / a painting very nice
"a (very) nice painting"
c. *un **beau** à couper le souffle tableau / un tableau **beau** à couper le souffle
a nice to cut the breath painting / a painting nice to cut the breath
"a breathtakingly beautiful painting"

The examples in (2) show that the positionning of attributive adjectives is a complex phenomenon: unlike *magnifique*, the adjective *beau* cannot be placed freely when it is the only element of the adjectival phrase (AP). It is strongly preferred in anteposition (2-a). The addition of the pre-adjectival adverb *très* gives more flexibility and equally allows both orders (2-b) whereas the use of a post-adjectival modifier constrains the placement to postposition (2-c).

The phenomenon of adjective alternation has been widely studied in French linguistics ([1], [2], [3], [4], [5], [6], [7] among others). Many constraints were proposed on different dimensions of the language: phonology, morphology, syntax, semantics[3], discourse and pragmatics. Only one of them is categorical in the

[3] In some cases, alternation leads to meaning differences for the adjective (see for instance [1], [3], [7]). The decision between the different possible accounts of how these differences could be generated is beyond the scope of this article. We thus leave aside these semantic considerations and focus here on the form of the adjective.

sense that it imposes a specific position to any attributive adjective: the presence of a post-adjectival complement (3) or modifier (2-c) only allows postposition of the adjective.

(3) un homme **fier** de son fils / *un **fier** de son fils homme
a man proud of his son / a proud of his son man
"a man proud of his son"

The other constraints participating in the alternation between anteposition and postposition are not categorical. For instance, as noted in the corpus studies of [2] and [3], length is a preferential constraint: short adjectives tend to be anteposed to the noun. The sequence "un magnifique tableau" in (1) illustrates that this rule can be violated whether one considers the length of the adjective alone: *magnifique* has 3 syllables, or the relative length between the adjective and the noun ($3 > 2$).

Although the above-mentioned works have enabled to identify the constraints playing a role in the placement of adjectives, most are based on introspection and only examine a few of these constraints. It is thus very difficult to evaluate the actual impact of each of them in usage, and therefore to estimate their respective weight in the speaker's choice for one position over the other. This paper aims to get a better grasp of the general picture of the phenomenon. To do so, we present along the same lines as [8], [9] and [10], a quantitative study, based on two corpora: the French Tree Bank (henceforth FTB) and the Est-Républicain corpus (henceforth ER). We propose a regression model based on interpretable constraints and compare the prediction capacities of different subsets in order to determine what kind of informations are the most reliable to account for the placement of adjectives.

The paper is organised as follows. We present in section 2 the methodological aspects of our study : constitution of the datatable and presentation of the statistical model. In section 3, we describe the variables derived from the constraints found in the literature. Section 4 is dedicated to the comparison of the models based on the different subsets of variables and to the interpretation of the results.

2 Methodology

Building the Datatable The first step of this work is to collect the data concerning adjectives and capture the constraints found in the literature. The study is based on the functionally annotated subset of the FTB corpus [11][4], which contains 12,351 sentences, 24,098 word types and 385,458 tokens. It is, for the moment, the only existing treebank for French. We extracted all the occurrences

[4] This subset corresponds to the part that was manually corrected.

of attributive adjectives from this corpus[5], and filtered out numeral adjectives[6], adjectives appearing in dates[7], abbreviations[8] and incorrectly annotated occurrences. We also discarded the 438 adjectives occurring with a post-adjectival dependent since postposition is obligatory in this case, regardless of the values of other constraints that we consider (see (2-c) and (3)). The remaining adjectives constitute the basis of the datatable, to which we have added information on the position of each adjective with respect to the noun it modifies, and 10 other variables that we describe in section 3.

Three variables of our study are based on frequency counts: FREQ, COLLOCANT and COLLOCPOST. They were extracted from the ER corpus for more reliable counts. The raw corpus contains 147,934,722 tokens, and is available on the ATILF website[9]. It was tagged and lemmatized with the *Morfette* system [12] adapted for French. We used ER for these constraints because it is around 380 times larger than FTB. We therefore consider that frequency in ER is a better estimator of the probability of use of an adjective. Also, we use here a log transformed value of the frequency to reduce the range of values of this variable. More precisely, the three variables take the following value: $log(frequency\ in\ ER + 1)$, in order to avoid a null value in case an adjectival lemma or noun-adjective combination is absent from ER.

Presentation of the Datatable The datatable contains 14,804 occurrences corresponding to 1,920 adjectival lemmas. 4,227 (28.6%) tokens appear in anteposition, and 10,577 (71.4%) in postposition. Table 1 shows that the adjectival lemmas displaying occurrences in both positions represent only 9.5% of all lemmas, yet these few lemmas correspond to 5,473 occurrences, i.e. 37.0% of the datatable, which means that very few adjectives actually alternate in usage but they are highly frequent.

Note that among the alternating adjectives (occurring in both positions), the ratio between anteposed and postposed occurrences is the reverse from that of all adjectives: there are 3,727 anteposed (68,1%) and 1,746 postposed (31,9%) adjectives. Alternating adjectives thus show a preference for anteposition. The general pattern is therefore that postposed adjectives tend to be infrequent lemmas occurring only in postposition, whereas alternating adjectives tend to be frequent and to prefer anteposition.

Statistical inference and Logistic Regression We used logistic regression models [13] to estimate the distribution of adjective positions using the variables from the datatable. Formally, a logistic regression is a function for which values

[5] We identified attributive adjectives using the following pattern in the treebank: an adjective occuring with a nominal head within a NP is an attributive adjective.

[6] Cardinal numerals such as *trois* 'three', *vingt* 'twenty', *soixante* 'sixty'... are sometimes annotated as adjectives in the FTB.

[7] Examples of dates containing adjectives: "$[13]_{ADJ}\ [mars]_N$", "$[lundi]_N\ [31]_{ADJ}$".

[8] Nouns or adjectives are viewed as abbreviations if their last letter is a capital letter.

[9] *http://www.cnrtl.fr/corpus/estrepublicain/*

	anteposed	*postposed*	*both positions*	*Overall*
number of lemmas	125	1613	182	1920
	6.5%	*84.0%*	*9.5%*	*100%*
tokens	500	8831	5473	14804
	3.4%	*59.7%*	*37.0%*	*100%*

Table 1. Distribution of adjectival lemmas and tokens according to position

can be interpreted as conditional probabilities. Its analytical form is as follows:

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

where, in our case, π_{ante} is the probability for the adjective to be anteposed and β corresponds to the abbreviation of the sequence of regression coefficients α, $\beta_0 ... \beta_{\text{n}}$, respectively associated with the predicting variables $X_0 ... X_{\text{n}}$. Given a scatter plot, the calculation of regression consists in the maximum likelihood estimation of α and β_{i} parameters for each variable in a *logit* space.

This type of modelling consists in the combining of several explicative variables (binary or continuous) to predict the behaviour of a single binary variable, here the position of the adjective. More precisely, we estimate the probability of anteposition as a function of 10 variables. Given one adjectival occurrence and the value of the 10 explanatory variables attributed to this occurrence, the model gives the probability of anteposition of the occurrence. Here, the model predicts postposition if the probabilty is below 0.5, and anteposition if it is higher or equal to 0.5. The accuracy gives the proportion of data that is correctly predicted according to this threshold. However, this measure does not evaluate completely satisfactorily the predictive power of the model because the threshold is arbitrary and does not account for the fact that a probability of 0.55 is different from a probability of 0.95. We therefore use an additional measure: the area under the ROC curve (AUC) [14], [15]. This measure gives the discrimination capacity of the model for all the pairs of opposite responses. A model with an AUC probability close to 0.5 indicates random predictions, and a value of 1, perfect prediction. It is usually considered that a model with an AUC value equal or above 0.8 has some utility in predicting the value of the dependent variable [14, p. 247].

The methodology of this paper consists in the comparison of models based on different constraint clusters, in order to evaluate their respective relevance. The comparisons take as reference a baseline model that does not contain any explanatory variables and systematically predicts postposition. Its accuracy is of 71.4% ($\sigma = 0.019$), which corresponds to the proportion of postposed adjectives in the datatable. Moreover, for the baseline model, AUC = 0.5, given that this model does not discriminate anteposition and postposition.

3 Variables

The variables we use in our logistic regression models are derived from the constraints found in the literature on attributive adjectives in French. They are summarized in table 2. Each model is based on different sets of constraints according to specific properties. The first set (COORD and ADV) concerns the syntactic environment of the adjective, the second is based on the lexical properties of the adjectival item (DERIVED, NATIO and INDEF), the third one on constraints linked to cognitive processing (ADJ-LENGTH, AP-LENGTH and FREQ). Finally, the fourth group examines collocationnal effects of the Noun - Adjective combination (COLLOCANT and COLLOCPOST).

Variables	Types	Description
COORD	*bool*	adjective in coordination or not
ADV	*bool*	adjective with pre-modifying adverb or not
DERIVED	*bool*	derived adjective or not
NATIO	*bool*	adjective of nationality or not
INDEF	*bool*	indefinite adjective or not
ADJ-LENGTH	*real*	length of the adjective in syllables (log scale)
AP-LENGTH	*real*	length of the AP in syllables (log scale)
FREQ	*real*	adjective frequency in the ER corpus (log scale)
COLLOCANT	*real*	score for the adjective-noun bigram (log scale)
COLLOCPOST	*real*	score for the noun adjective bigram (log scale)

Table 2. Summary table of variables and their values (*bool* = boolean and *real* = real number)

3.1 Syntactic variables

The variables based on syntactic properties rely on the idea that the internal structure of the AP has an influence on the placement of adjectives. As seen in the introduction, one of them, i.e. the presence of a post-adjectival dependent, is categorical and is therefore not integrated in our study. It suggests however that the syntactic environment of the adjective may have an important role in its positioning. We thus propose here two other constraints related to different internal structures within the AP.

Coordination (COORD) In a competence account of attributive position like in [6], the position of coordinated adjectives is not restricted, as can be seen in example (4) (from [6]).

(4) une **belle** et **longue** table / une table **belle** et **longue**
a beautiful and long table / a table beautiful and long
"a long and beautiful table"

However, 94.6% of coordinated adjectival occurrences (i.e. 758 occurrences) are postposed in our data. Usage-based data thus indicate that coordination is a factor that strongly favours postposition.

Presence of a Pre-Adjectival Adverb (ADV) The constraint is the same as for coordination if one considers the adverbial category on a general level: the presence of a pre-adjectival modifier does not restrict the position of the modified adjective (example (5)).

(5) une très **longue** table / une table très **longue**
a very long table / a table very long
"a very long table"

On a more specific level, [6] point out that adjectives can be postposed with any adverb whereas only a small set of adverbs allows anteposition. This is confirmed in our datatable: 11 types of adverb[10] are observed with anteposed adjectives, while 119 different types appear with adjectives in postposition. Furthermore, the adverbs found with adjectives in anteposition are not specific to this position, they also appear with postposed occurrences. From a general quantitative point of view, 74.9% of the premodified adjectival occurrences are in postposition.

3.2 Cognitive Processing Variables

Length and frequency of occurrence are constraints that have cross-linguistically been observed to play a role in different types of phenomena, amongst which the adjective alternation. These constraints are usually related to processing ease (see for instance [16], [17] and [18]). We present the functioning of the constraints in what follows and leave the interpretation in terms of cognition to the discussion of the models performance.

Length Numerous works on word order use the notion of length: for attributive adjectives in French [2], [3], for word [19], [20] and constituent [21], [16], [9], [10] alternation in other languages. The main idea is expressed by the principle *short comes first*, i.e. short elements tend to appear first. Here, we consider length in terms of number of syllables and we introduce two variables: length of the adjective (ADJ-LENGTH) and length of the adjectival phrase (AP) (AP-LENGTH)[11].

[10] The 11 adverbs are: 'encore' *again*, 'désormais' *from now on*, 'moins' *less*, 'peu' *not much*, 'plus' *more*, 'si' *so*, 'tout' *very*, 'très' *very*, 'trop' *too*, 'bien' *well*, 'aussi' *also*.

[11] We obtain the number of syllables using the speech synthesis software ELITE [22]. It counts the number of syllables for every token, taking into account the actual form of the adjective (feminine versus masculine, for instance) as well as the possible effects of sandhi phenomena, like the *liaison* phenomenon. The length associated to each adjectival type corresponds to the mean of all its tokens length. For both variables, we use the log transformed value of the length in order to reduce the effect of outliers.

Lemma Frequency (FREQ) In his corpus study, [3] observes that frequency is correlated with the position of the adjective: pre-nominal adjectives tend to be frequent whereas post-nominal adjectives tend to be rare. According to the author, this distribution has historical grounds. In Old French, the general pattern was the reverse of that of Modern French: adjectives were generally placed before the noun, as in English. The evolution to the preference for postposition in Modern French did not affect the most frequent adjectives because their association to anteposition was too robust to reverse the pattern. Note that this hypothesis is not particular to French, nor to the adjective alternation, see for instance the summary in [23, ch.11] of several studies that make the same observation of conservatism linked to frequency.

3.3 Lexical Variables

Most reference grammars state that adjectives are mainly placed according to their lexical properties. These properties can concern different aspects of language. We propose to examine the relevance of lexical information with the example of three classes, each based on one particular aspect: morphology (DERIVED), semantics (NATIO), and syntactic behaviour (INDEF).

Derived Adjectives (DERIVED) Adjectives may be derived from other parts-of-speech: for instance, certain verbal forms can be used as adjectives (past participles, present participle) or the adjective is obtained by suffixation, to a verbal basis: *-ible* 'faillible' (*faillible) / -able* 'faisable' (*doable*) */if* 'attractif' (*attractive*), or to a noun ('métallique' (*made of metal*), 'scolaire' (*academic*), 'présidentiel' (*presidential*)). These adjectives are described as prefering postposition but anteposition is also possible as shown in example (6).

(6) notre **charmante** voisine est beaucoup trop bavarde
our charming neighbour is lot too talkative
'our charming neighbour is too talkative'

In our datatable, the adjectives derived from another part-of-speech (noun or verb) are collected using the software of derivational morphological analysis DERIF [24]. Our data confirms the strong preference for postposition within this class (91.3%).

Semantic Classes It is usually said that objective adjectives, i.e. adjectives for which the semantic content is perceptible or can be infered from direct observation, are postposed. Objective adjectives are classified into sub-groups like form, colour, physical property, nationality, technical terms... In order to estimate the relevance of lexical classes according to semantic properties, we test the predictive capacity of adjectives denoting nationality[12] (NATIO). In theory, these adjectives strongly tend to be postposed, but they may also occur in anteposition (example (7)):

[12] Using the dictionnary PROLEXBASE [25].

(7) cette très **italienne** invasion de l'Albanie
this very Italian invasion of Albania
"this very Italian invasion of Albania" (in a typical Italian fashion way)[26, p. 142]

The strong preference for postposition of these adjectives is confirmed by our data: only one pre-nominal occurrence is observed (example (8)).

(8) la très **britannique** banque d' affaires et de marché
the very British bank of affairs and of market
"the very British merchant bank"

Indefinite Adjectives (INDEF) A relatively closed set of adjectives are special in the fact that their syntacitc properties show a hybrid behaviour between determiners and adjectives. On the one hand, indefinite adjectives may introduce and actualise the noun, like determiners. On the other hand, they may co-occur with a determiner and can be placed in post-nominal position, even though they favour anteposition (89% in our data). These latter properties are specific to attributive adjectives. The adjectives identified as indefinite in the datatable are: 'tel' (*such*), 'autre' (*other*), 'certain' (*some/sure*), 'quelques' (*few*), 'divers' (*various*), 'différent' (*different*), 'maint' (*numerous*), 'nul' (*null/lousy*), 'quelconque' (*any/ordinary*), 'même' (*same/itself*).

3.4 Collocation Variables

It is well known that the nature of some Adjective-Noun combinations is strongly collocational in French [27]. This implies that the position of attributive adjectives in French should also be influenced by collocational effects. Collocations are here defined according to [28, p. 151]. Adjective-Noun collocations may be non-compositional sequences as well as more compositional ones. The sequence 'libre échange' (lit. *free exchange*) is an example of the former: it refers to a specific economical system, not to exchange in general. As an illustration of the latter case, the meaning of 'majeure partie' (*major part*) is predictable from the meaning of the two components. It is nevertheless a collocation because the order between the elements is fixed by convention of use. As mentioned in section 2, the collocation score in our datatable is based on the frequency of Adjective-Noun (COLLOCANT) and Noun-Adjective (COLLOCPOST) bigrams in the ER corpus.

4 Prediction Model of Attributive Adjective Position

The prediction model is built with all the variables described in part 3 and maximized with a backward elimination procedure based on the AIC criterion [29][13]. The ADJ-LENGTH constraint's contribution to the model is not significant

[13] Forward selection procedure gives the same results for this particular model.

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

$$\begin{aligned} X\beta = & -2.14 && \\ & -1.07 \ \text{COORD} = 1 && *** \\ & -1.30 \ \text{AP-LENGTH} && *** \\ & +0.29 \ \text{FREQ} && *** \\ & -0.50 \ \text{DERIVED} = 1 && *** \\ & +0.91 \ \text{INDEF} = 1 && *** \\ & -4.58 \ \text{NATIO} = 1 && *** \\ & -0.75 \ \text{ADV} = 1 && *** \\ & +1.28 \ \text{COLLOCANT} && *** \\ & -1.24 \ \text{COLLOCPOST} && *** \end{aligned}$$

Fig. 1. Formula of prediction model, significant effects are coded *** p<0.001, ** p<0.01, * p<0.1

according to the procedure. It was thus eliminated. The model is presented in figure 1[14].

The coefficients combined with each variable are estimated from the distribution of the variables in our datatable. In the case of boolean variables, these coefficients are multiplied by 1 when the predictor is true, and by 0 when it is false. As for the numerical variables (AP-LENGTH, FREQ, COLLOCANT, COLLOCPOST), their participation to the models consists in the multiplication of the coefficient by the numerical value of the variable itself. In this model, all the variables have positive values, so we can straightforwardly interpret the sign of the coefficients: positive coefficients indicate that the variables prefer anteposition, whereas negative coefficients show that the variables favour postposition. As we expected, the variables COORD, AP-LENGTH, DERIVED, NATIO, ADV and COLLOCPOST tend to favour postposition, whereas FREQ, INDEF and COLLOCANT vote for anteposition

Compared to the *baseline model* performances (accuracy of 71.4% and AUC = 0.5), this model has significantly better predictive capacities. The prediction performances associated with the procedure of decision are presented in table 3. One can see that the *global model* correctly predicts the position of 92.6% of the datatable. Moreover, the concordance probability is AUC = 0.969 (σ = 0.003), which indicates that the model predictions are very accurate. To have a graphical idea of the goodness of fit of the model, the plot in figure 2 gives the relation between the observed proportions and the corresponding mean expected

[14] The condition number of the model is $\kappa = 13.35$. It indicates that the collinearity of the model is moderate [30]. When the predictors of a regression model are collinear, the interpretation of the contribution of each predictor can rise problems. Given that we do not interpret the values of the coefficients, but only to the sign of these coefficients, the moderate collinearity of our data does not affect the validity of our results.

probability for the model[15]. It shows that the fit is very good for probabilities under 0.5, and not quite as good for higher probabilities.

		Predicted position		% Correct
		P	A	
observed	P	10222	355	96.6%
position	A	748	3479	82.3%

Overall accuracy: 92.6% ($\sigma = 0.008$)

Table 3. Classification table for *prediction model*

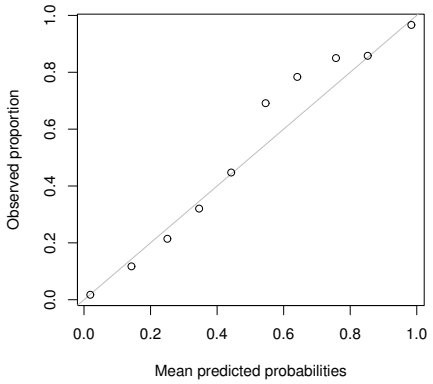

Fig. 2. Observed proportions of anteposition and the corresponding mean predicted probabilities for the prediction model (the line represents a perfect fit).

4.1 Comparison of models with different sets of constraints

In order to compare the effect of different constraint clusters, we propose 4 prediction models based on different groups of variables: a *Syntactic model* containing COORD and ADV; a *Lexical property model* with NATIO, INDEF and DERIVED; a *Frequency-Length model* containing the variables AP-LENGTH and FREQ and a *Collocation model* containing COLLOCANT and COLLOCPOST.

[15] We compute the mean probability of success of ten equally sized bins of probabilities $(0 - 0.1, 0.1 - 0.2, 0.2 - 0.3 \ldots)$ and we compare this mean with the proportion of observed success in the data.

Syntactic Model (COORD and ADV). The comparison based on accuracy shows that the effect of the syntactic constraints is insignificant when they are not combined with other constraints. The *Syntactic model* accuracy is 71.4% ($\sigma = 0.02$), and the classification table in 4 shows that this model cannot predict anteposition. The value of the concordance probability (AUC = 0.534, $\sigma = 0.008$) confirms that the predictive power of these variables is very poor. This lack of predictive power can be partly explained by the fact that these two variables are relevant for a very small set of data: ADV and COORD represent respectively 5.2% and 5.4% of all the data. In addition, both constraints favour postposition, which is already the default position predicted by the baseline model. This means that these constraints can only be relevant when other constraints are also taken into account.

		Predicted position		% Correct
		P	A	
observed	P	10574	3	99.9%
position	A	4227	0	0%

Overall accuracy: 71.4% ($\sigma = 0.02$)

Table 4. Classification table for *Syntactic model*

Lexical Properties Model (NATIO, INDEF and DERIVED) Lexical properties are relevant when they are not combined with the other constraints (*Lexical properties model* accuracy = 74.7% and AUC = 0.717). However, the table in 5 indicates that the lexical properties that we used do not predict satisfactorily anteposition: only 12.9% of anteposed adjectives are correctly accounted for. This is mainly due to the fact that only one of the variables (INDEF) favours the prenominal position. Nevertheless, one can see that these three constraints alone enable the model to consider anteposition, which was not the case with syntactic constraints. This observation suggests that speakers may be sensitive to this type of information and encourages us to extend the lexical classification for all the adjectives of the datatable, in particular those that favour anteposition, in order to improve our modelling.

		Predicted position		% Correct
		P	A	
observed	P	10506	71	99.3%
position	A	3681	546	12.9%

Overall accuracy: 74.7% ($\sigma = 0.02$)

Table 5. Classification table for *Lexical properties model*

Frequency-Length Model (AP-LENGTH and FREQ). The variables of length and frequency have an important predictive power (accuracy 81.7% ($\sigma = 0.009$), AUC = 0.869 ($\sigma = 0.010$)). In particular, the predictions for anteposition are much higher than observed with the two preceding models (65%). These two constraints may thus play an important role in the placement of adjectives. As expected, the model tends to predicts anteposition for short and frequent adjectives, and postposition when the adjective is longer and/or less frequent[16].

		Predicted position		% Correct
		P	A	
observed	P	9334	1243	88.2%
position	A	1475	2752	65.1%

Overall accuracy: 81.7% ($\sigma = 0.009$)

Table 6. Classification table for *Frequency-Length model*

Collocation Model (COLLOCANT and COLLOCPOST). The *Collocation model* shows that the frequency of bigrams represents the best predictor. The *Collocation model* accuracy is of 89.9% ($\sigma = 0.013$) and the AUC value increases up to 0.940 ($\sigma = 0.006$). This result suggests that the order of the adjective-noun sequence depends highly on the nature of both the noun and the adjective, and on the frequency with which these elements appear in a specific order. It thus appears here again that frequency is a good predictor for the placement of adjectives. However, the fact that this model is more performant than the previous one seems to show that frequency is a better predictor when it takes into account more information than the adjective isolated from its context of appearance.

		Predicted position		% Correct
		P	A	
observed	P	10327	250	97.6%
position	A	1249	2978	70.5%

Overall accuracy: 89.9% ($\sigma = 0.007$)

Table 7. Classification table for *Collocations model*

[16] Note that frequencies are biased by the journalistic nature of corpora: adjectives of nationality are frequent despite the fact that they are postposed in most cases. Nevertheless, the variable NATIO of the global prediction model votes for postposition, which neutralizes the frequency effect.

4.2 Discussion

The models presented above show that the constraints playing a significant role in the adjective alternation are information specific to the adjectival item and to its context of use, rather than constraints based on a more general and abstract level. These specific informations relate to different aspects of language. On the one hand, the constraints tested in the lexical model (NATIO, INDEF and DERIVED) concern inherent linguistic properties. On the other hand, length of the AP, frequency of the adjective, and collocational effects are more related to the way language is processed and used: how speakers place the AP according to its linear constitution during discourse, and how they retrieve the units (or sequences) in accordance with their past experience of these elements. The importance of the predictive power of the second set of specific constraints suggests that adjective alternation may be best accounted for in terms of cognitive approaches to language.

As mentioned in the description of the length variable (sec. 3.2), the tendency to place short elements first is not specific to adnominal adjectives in French. It is also observed in other works for various phenomena in other languages. The general preference for such a placement is explained by the fact that it eases the on-line processing of the structure within which the element occurs: for example, [21], [16], [9], [10] , who study different constituent to constituent ordering phenomena, state that anteposition of the short element helps to faster plan/recognise the overall structure of the immediately dominating constituent. This idea can be applied to the Adjective-Noun combination. A short AP in anteposition leads to a faster production/reception of the Head-Noun in comparison with a longer one, and hence to a faster access to 1) the complete internal constitution of the AP, 2) more information concerning the structure of the NP. The significant contribution of length in the prediction of adjective alternation may thus be viewed as another support for an explanation in terms of processing ease.

In a similar perspective, Usage-based models (see for example [31], [32], [33]) consider that frequency plays an important role in the constitution of the speaker's linguistic knowledge. These approaches view linguistic knowdledge as mental representations based on the storage of instances of language encountered by the speaker. This means that speakers store isolated words like it is traditionally assumed for the constitution of the lexicon, but also that they memorize information about specificities related to their context of appearance. In our study, the assumption is that speakers would have mental representations corresponding to the adjective, and representations of specific ordered Adjective-Noun sequences in which it appears. Furthermore, these models consider that every occurrence of an instance affects the corresponding mental representations. Of particular interest here, [18] notes that the repetition of a language instance strengthens its representation and makes its execution more fluent. The instance also becomes more entrenched in the morpho-syntactic structure in which it usually appears, which leads to more resistance for a change of structure. Applied to adjectives, this means that a highly frequent item (or sequence) is highly ac-

cessible, and thus easy to process. When the item is tested in isolation (FREQ), if processing ease plays an important role in the placement of adjectives as it was suggested for length, one would expect highly frequent adjectives to favour anteposition. Concerning collocationnal effects (COLLOCANT and COLLOCPOST), the prediction of this approach is that a collocationnal sequence would be reproduced in the same morpho-syntactic configuration, i.e the order between the Adjective and the Noun should be maintained as it is usually encountered. As it was observed for length, the good results of the models involving frequency constraints are in accordance with theses assumptions and can be seen as a support for this type of approach.

5 Conclusion

We examined in this article the question of the alternation of attributive adjectives in French using quantitative methods applied to corpora. One can draw several conclusions from the logistic regression models that we proposed. First the satisfactory results of the general model show that a good part of the modelling can be done on the basis of the form without considering the semantics due to position. The importance of the form is also outlined by the fact that the constraints identified as having some relevance when isolated from the others are all based on a knowledge linked to the specificities of the item, or to the specific context in which it appears. Nevertheless, the prediction performances may be improved by taking more semantics into account: adding information for other lexical classes, including semantics, should naturally enhance the model. Furthermore, the importance of collocational effects suggests that semantics should also be considered on a specific relational level between the noun and the adjective. It thus raises the question on how to capture and formalise semantic relations in a quantitative study. Finally, the results of our study show that the best models are based on length and frequency, and collocational effects. This confirms the role of the nature of the items involved. It also suggests that usage may have an important role in the construction of linguistic knowledge, and hence in the placement of adjectives.

To conclude, the model proposed in this article is restricted to the journalistic genre. A future perspective in our work would be to extend this study to other genres, in particular spoken data, in order to test the relevance of our conclusions for French more generally. Furthermore, a comparison between the probabilities given by our model and speakers' preferences on the basis of experiments would enable us to see if future psycholinguistic work will confirm our hypothesis that the effects of usage statistics on adjective position in French is mediated by cognitive processes whereby linguistic representations are directly sensitive to the statistics of language use experienced by language users.

Acknowledgements We would like to express our gratitude to Benoît Crabbé for his support and advise in the work that lead to this paper, and to the anonymous reviewers for their valuable comments.

References

1. Waugh, L.R.: A semantic analysis of word order : Position of the Adjective in French. E. J. Brill, Leiden (1977)
2. Forsgren, M.: La place de l'adjectif épithète en français contemporain, étude quantitative et sémantique. Almqvist & Wilksell, Stockholm (1978)
3. Wilmet, M.: La place de l'épithète qualificative en français contemporain : étude grammaticale et stylistique. Revue de linguistique romane **45** (1981) 17–73
4. Delbecque, N.: Word order as a reflexion of alternate conceptual construals in french and spanish. similarities and divergences in adjective position. Cognitive Linguistics **1** (1990) 349–416
5. Nølke, H.: Où placer l'adjectif épithète? focalisation et modularité. Langue française **111** (1996) 38–57
6. Abeillé, A., Godard, D.: La position de l'adjectif èpithète en français : le poids des mots. Recherches linguistiques de Vincennes **28** (1999) 9–32
7. Noailly, M.: L'adjectif en français. Ophrys (1999)
8. Arnold, J.E., Wasow, T., Losongco, A., Ginstrom, R.: Heaviness vs. newness: the effects of structural complexity and discourse status on constituent ordering. Language **76(1)** (2000) 28–55
9. Rosenbach, A.: Animacy versus weight as determinants of grammatical variation in english. Language **81(3)** (2005) 613–644
10. Bresnan, J., Cueni, A., Nikitina, T., Baayen., H.: Predicting the dative alternation. In Boume, G., Kraemer, I., Zwarts, J., eds.: Cognitive Foundations of Interpretation. Royal Netherlands Academy of Science, Amsterdam (2007)
11. Abeillé, A., Barrier, N.: Enriching a french treebank. In: Proceedings of Language Ressources and Evaluation Conference (Lrec), Lisbon (2004)
12. Grzegorz Chrupala, G.D., van Genabith, J.: Learning morphology with morfette. In Calzolari, N., *et al*, eds.: Proceedings of LREC'08, Morocco, ELRA (2008)
13. Agresti, A.: An introduction to categorical data analysis. Wiley interscience (2007)
14. Harrell, F.E.: Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. Springer series in statistics. Springer (2001)
15. Bresnan, J., Ford, M.: Predicting syntax: Processing dative constructions in american and australian varieties of english. Language **86**(1) (2010) 186–213
16. Wasow, T.: Postverbal behavior. CSLI publications (2002)
17. Hawkins, J.: Efficiency and Complexity in Grammars. Oxford University Press (2004)
18. Bybee, J., Mcclelland, J.L.: Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. The Linguistic Review **22** (2005) 381–410
19. Cooper, W.E., Ross, J.R.: World order. In: Papers from the Parasession on Functionalism. Chicago Linguistic Society (1975) 63–111
20. Benor, S.B., Levy, R.: The chicken or the egg? a probabilistic analysis of english binomials. Language **82(2)** (2006) 28–55
21. Hawkins, J.: The relative order of prepositional phrases in english: Going beyond manner-place-time. Language Variation and Change **11** (2000) 231–266
22. Beaufort, R., Ruelle, A.: elite : système de synthèse de la parole à orientation linguistique. In: Actes des XXVI journées d'études sur la parole, Dinard (2006)
23. Croft, W., Cruse, D.A.: Cognitive Linguistics. Cambridge University Press (2004)

24. Namer, F.: Acquisition automatique de sens à partir d'opérations morphologiques en français : étude de cas. In: Traitement Automatique de la Langue Naturelle (TALN). (2002)
25. Tran, M., Maurel, D.: Prolexbase : Un dictionnaire relationnel multilingue de noms propres. Traitement automatique des langues **47**(3) (2006) 115–139
26. Bouchard, D.: The distribution and interpretation of adjectives in french: a consequence of bare phrase structure. Probus **10**(2) (1998) 139–183
27. Gross, G.: Les expressions figées en français: noms composés et autres locutions. Ophrys, Paris (1996)
28. Manning, C.D., Schütze, H.: Foundations of statistical natural language processing. MIT Press, Cambridge (1999)
29. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control **19**(6) (1974) 716–723
30. Belsley, D.A., Kuh, E., Welsch, R.E.: Regression diagnostics: identifying influential data and sources of collinearity. Wiley, New-York (1980)
31. Bybee, J.: The emergent lexicon. In Gruber, M.C., Higgins, D., Olson, K.S., Wysocki, T., eds.: CLS 34: The panels, University of Chicago, Chicago Linguistic Society (1998)
32. Croft, W.: Radical Construction Grammar. Oxford University Press (2001)
33. Goldberg, A.: Constructions at Work: the nature of generalization in language. Oxford University Press (2006)