Experimental approach : The case of the syntax of attributive adjectives

Juliette Thuilier

1Université Rennes 2 & LLF (CNRS - Paris Diderot)
juliette.thuilier@univ-rennes2.fr

September 28, 2013

Colloque de Syntaxe et Sémantique à Paris
CSSP 2013
Université Paris Diderot
Introduction

- In French, attributive adjectives (A) can appear both before or after the noun (N)

(1)  
\begin{align*}
  & a. \text{une agréable soirée (anteposed)} \\
  & b. \text{une soirée agréable (postposed)} \\
  & \text{a nice evening}
\end{align*}

⇒ Numerous factors involved and interacting in a complex way
⇒ Experimental approach: Multifactorial statistical modeling of corpus data and questionnaire experiment
The phenomenon
- Semantic aspects
- Lexical aspects
- Syntactic aspects
- Specific combinations of N and A

Corpus data modeling
- Methodology
- Observations
- Multifactorial statistical modeling
- Results

Speaker preferences: Correlation with the corpus model
- Methodology
- Results

Conclusion and perspectives
1. Semantic aspects

- Anteposed As tend to be subsective or intensional (= non-predicative)

  (2) a. une petite souris  
      *a small mouse*  
  
  b. un vrai complot  
      *a true plot*

- Postposed As tend to be intersective (or predicative)

  (3) un vase fragile  
      *a fragile vase*

- But there is no semantic property categorically associated with one position (Abeillé and Godard, 1999)

  ⇒ Hypothesis: the semantics do not account for the entire phenomenon (*contra* Bouchard, 1998) and the choice of the position is mainly driven by lexical properties and syntactic constraints
2. Lexical aspects

Adjectives show individual preferences that are shaped by formal properties: length, frequency and morphological complexity (Wilmet, 1981; Thuilier et al., 2010; Fox and Thuilier, 2012)

<table>
<thead>
<tr>
<th>Anteposition</th>
<th>Noun</th>
<th>Postposition</th>
</tr>
</thead>
<tbody>
<tr>
<td>short</td>
<td>frequent</td>
<td>long</td>
</tr>
<tr>
<td></td>
<td>morphologically simple</td>
<td>rare</td>
</tr>
<tr>
<td></td>
<td></td>
<td>morphologically complex</td>
</tr>
</tbody>
</table>
3. Syntactic aspects

1. A postadjectival dependent makes postposition obligatory: only categorical constraint

(4) a. une musique agréable à écouter
   b. *une agréable à écouter musique
      a nice (type of) music to hear

2. Premodified and coordinated adjectives can appear in both positions

(5) a. une très agréable soirée
    b. une soirée très agréable
       a very nice evening

(6) a. une belle et longue table
    b. une table belle et longue
       a beautiful and long table

- Observations in corpus data show that coordinated and modified As are more often postposed (Forsgren, 1978; Thuilier et al., 2012)
3. Syntactic aspects

Other dependents after the noun favor anteposition (Grevisse and Goosse, 2007)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
</table>
| 7 | a. **un récent** recueil [de textes grecs]$_{PP}$ *a recent collection of Greek texts*
    | b. **l’habituel** air [que joue Paul]$_{RC}$ *the usual tune that Paul plays*
    | c. **un étrange** animal [indomptable]$_{A}$ *a strange untameable animal*

Definite determiners favor anteposition (Forsgren, 1978)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
</table>
| 8 | a. **un éblouissant** spectacle *a this dazzling show*
    | b. **cet éblouissant** spectacle
| 9 | a. **un habituel** refrain *a her usual record*
    | b. **son habituel** refrain
| 10| a. **une très attendue** décision *a/the highly anticipated decision*
    | b. **la très attendue** décision
4. Specific combinations of N and A

Some adjective - noun pairs are strongly collocational

(11) a. un vibrant hommage
    a vibrant tribute
b. un lourd tribut
    a heavy toll

Some semantic effects are observed with specific pairs of adjective and noun

(12) a. un gros fumeur
    a heavy smoker
b. un fumeur gros
    a fat smoker
(13) a. un gros coiffeur
    a fat hairdresser

⇒ The noun that combines with the adjective has an influence on the choice of the position
1 The phenomenon
   - Semantic aspects
   - Lexical aspects
   - Syntactic aspects
   - Specific combinations of N and A

2 Corpus data modeling
   - Methodology
   - Observations
   - Multifactorial statistical modeling
   - Results

3 Speaker preferences: Correlation with the corpus model
   - Methodology
   - Results

4 Conclusion and perspectives
Corpus data modeling : methodology (1)

- The idea : to test syntactic factors found in the literature with attested data
  - Statistical modeling : logistic regression (Agresti, 2007) and mixed effect models (Gelman and Hill, 2006)
  - Assumption : these statistical tools allow one to be free from variation due to the sampling of the corpora
  - Advantage : predictive models

- Data extracted from two corpora :
  - French Treebank (FTB, Abeillé and Barrier, 2004) : newspaper corpus
    - automatic extraction of the attributive As appearing in both positions
    - leaving aside As with post-adjectival dependent
  - C-ORAL-Rom (CORAL, Cresti and Moneglia, 2005) : spontaneous speech corpus
    - manual extraction of the same As as in FTB
Corpus data modeling: methodology (2)

- Annotation of the data for 11 variables

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 COORD</td>
<td>the adjective is coordinated or not</td>
</tr>
<tr>
<td>2 MOD</td>
<td>the adjective is pre-modified or not</td>
</tr>
<tr>
<td>3 DEMDet</td>
<td>the NP is introduced by a demonstrative determiner or not</td>
</tr>
<tr>
<td>4 POSSDet</td>
<td>the NP is introduced by a possessive determiner or not</td>
</tr>
<tr>
<td>5 DEFArt</td>
<td>the NP is introduced by a definite article or not</td>
</tr>
<tr>
<td>6 PP</td>
<td>the NP contains a PP or not</td>
</tr>
<tr>
<td>7 REL</td>
<td>the NP contains a relative clause or not</td>
</tr>
<tr>
<td>8 POSTAdj</td>
<td>the NP contains a postposed or not</td>
</tr>
<tr>
<td>9 COLLOCAN</td>
<td>collocation score for A+N bigram ((\log(\chi^2)))</td>
</tr>
<tr>
<td>10 COLLOCNA</td>
<td>collocation score for N+A bigram ((\log(\chi^2)))</td>
</tr>
<tr>
<td>11 MODALITY</td>
<td>the modality is speech (s) or writing (w)</td>
</tr>
</tbody>
</table>
Corpus data modeling: methodology (3)

- Differentiation of two lemmas for 5 As
  - **seul** 'alone/single'
    (14) a. un seul homme / un homme seul
      'a single man / a lonely man'
  - **simple** 'simple/mere'
    (15) a. une simple phrase / une phrase simple
      'a mere sentence / a simple phrase'
  - **pur** 'pure'
    (16) a. un pur produit / un produit pur
      'an archetypal product / a pure product (not mixed)'
  - **propre** 'own/clean'
    (17) a. son propre pantalon / son pantalon propre
      'her own pants / her clean pants'
  - **ancien** 'former/ancient'
    (18) a. un coffre ancien / un ancien coffre
      'an ancient chest / a former chest'

- Only 5 lemmas of these 10 disambiguated lemmas appear in both positions in the data
Observations

- 6612 occurrences of attributive adjectives
  - 4494 in FTB
  - 1627 in CORAL

### Lemmas repartition in the data

<table>
<thead>
<tr>
<th></th>
<th>FTB</th>
<th>CORAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of lemmas</td>
<td>170</td>
<td>130</td>
</tr>
</tbody>
</table>

- Alternating lemmas
  - 43% in both positions in CORAL
  - Less alternation in spoken French
Statistical modeling

**Logistic regression** (Agresti, 2007)
- modeling the behavior of a binary variable as a function of several predictive variables
- adjective position = binary variable
  - postposition = 0
  - anteposition = 1

**Logistic function**
Function for which values can be interpreted as conditional probabilities

\[
\pi_{ante} = \frac{e^{\beta X}}{1 + e^{\beta X}} \tag{1}
\]

where
- \( \pi_{ante} \) = probability for the anteposition
- \( \beta \) = regression coefficients \( \alpha, \beta_0 \ldots \beta_n \)
- \( X \) = predictive variables \( X_0 \ldots X_n \)
Adjectives as source of variation in the data

- The example of three adjectives

How to capture the characteristics of each group of data?

⇒ Mixed-effects model (Gelman and Hill, 2006):
  - the idea: besides the general model (fixed-effects), each group of data receives its own coefficient (random-effect)
  - each adj ectival lemma has a coefficient capturing its individual behavior
The model

- Model built with 11 predictive variables and 1 random effect
- All the effects are significant and thus participate in predicting the position of the As
- The model predictions are very accurate
  - mean accuracy : 0.88 (10-fold cross validation)
  - mean concordance probability $C=0.947$ (10-fold cross validation)
- Low collinearity : condition number $\kappa = 8.15$
**Random effects**

<table>
<thead>
<tr>
<th>ADJECTIVE</th>
<th>Variance</th>
<th>Std.Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADJECTIVE</td>
<td>2.3938</td>
<td>1.5472</td>
</tr>
</tbody>
</table>

Number of obs : 6612, groups : ADJECTIVE, 170

**Fixed effects**

|                | Estimate | Std. Error | z value | P(>|z|)    |
|----------------|----------|------------|---------|------------|
| Intercept      | -0.782   | 0.182      | -4.304  | 1.68e-05 ***|
| DEMDET=1       | 1.226    | 0.246      | 4.99    | 5.9e-07 ***|
| POSSDET=1      | 1.185    | 0.235      | 5.04    | 4.7e-07 ***|
| DEFArt=1       | 0.370    | 0.107      | 3.47    | 0.0005 *** |
| POSTAdj=1      | 0.587    | 0.154      | 3.82    | 0.0001 *** |
| PP=1           | 0.840    | 0.104      | 8.04    | 8.6e-16 ***|
| REL=1          | 0.714    | 0.210      | 3.40    | 0.0007 *** |
| COLOCAN        | 0.378    | 0.018      | 20.52   | <2e-16 *** |
| MOD=1          | -1.957   | 0.174      | -11.26  | <2e-16 *** |
| COORD=1        | -1.266   | 0.266      | -4.76   | 1.9e-06 ***|
| COLLOCNA       | -0.443   | 0.020      | -22.12  | <2e-16 *** |
| MODALITY = w   | 0.458    | 0.121      | 3.78    | 0.0002 *** |
Results

Each coefficient associated with fixed-effects can be interpreted as the preference for a position

- positive coefficient = preference for anteposition
- negative coefficient = preference for postposition

Anteposition is favored by:

- Demonstrative, possessive and definite determiners
- Other dependent of the noun (relative, PP and postposed adjective)
- High score for AN collocation

Postposition is favored by:

- Heavy AP (coordinated or modified adjectives)
- High score for NA collocation
Results

1. Less alternation in the speech data
   ⇒ In unplanned discourse, speakers may tend to comply more often with lexical preferences

2. Lexical preferences and noun - adjective combination
   ▶ Each adjective has a more or less strong preference (random-effects)
   ▶ The noun that the adjective appears with can favor the non-preferred position:
     - *fort* ‘strong’ has a preference for anteposition (60%)
     - combined with the N *point*, postposition is strongly preferred

3. Effect of weight: long and complex APs tend to be postposed
   ▶ In accordance with the generalization that in SVO languages, the heavy constituents tend to appear last
   ▶ Significant interaction between MOD and MODALITY: the effect of the presence of a modifier is stronger in the speech data
   ⇒ In unplanned discourse, speakers may follow more often general principles like “short before long”

4. Effect of determiners
   ⇒ In anaphoric contexts, when the relation between the referent of the noun and the property denoted by the adjective is established, the anteposition is facilitated (Waugh, 1977)
1. The phenomenon
   - Semantic aspects
   - Lexical aspects
   - Syntactic aspects
   - Specific combinations of N and A

2. Corpus data modeling
   - Methodology
   - Observations
   - Multifactorial statistical modeling
   - Results

3. Speaker preferences: Correlation with the corpus model
   - Methodology
   - Results

4. Conclusion and perspectives
Speaker preferences: Correlation with the corpus model

- The corpus model estimates the probability of anteposition of each adjectival occurrence given the syntactic environment.
- A questionnaire experiment in order to test whether these probabilities are related to the preferences of the speakers.

Hypothesis

For many speakers the frequency of choice for anteposition corresponds to the probability of anteposition estimated in the corpus model.
Methodology

The questionnaire

- 29 sentences extracted from the data (FTB part)
- Sentences selected according to their probability
- Each sentence is part of a pair of sentences containing
  ▶ the original sentence
  ▶ a modified version with the sequence A N in the opposite order
- In both versions, the NP is in bold and colored letters
- The pairs and the sentences within the pairs are randomly ordered in each questionnaire

Henri Guitton a joué un rôle important dans la modernisation de l’enseignement de l’économie en France

Henri Guitton a joué un important rôle dans la modernisation de l’enseignement de l’économie en France
Methodology

The experiment

- Participants see both versions of the sentence on the screen
- They have to choose their preferred version
- The participants were contacted *via* social networks and scientific mailing lists
- 141 participants completed the questionnaire online
Results

- The proportion of choice for anteposed As significantly correlates with the probability of anteposition estimated in the corpus model
  - Pearson correlation coefficient: 0.74 (p.value = 0.0001)

⇒ Speaker preferences seem to be sensitive to the constraints used in the corpus model
⇒ The statistical modeling seems to be an appropriate way of describing and accounting for this rather complex syntactic phenomenon
The phenomenon
- Semantic aspects
- Lexical aspects
- Syntactic aspects
- Specific combinations of N and A

Corpus data modeling
- Methodology
- Observations
- Multifactorial statistical modeling
- Results

Speaker preferences: Correlation with the corpus model
- Methodology
- Results

Conclusion and perspectives
Experimental approach

- Accurate modeling of the phenomenon based on corpus data
- Questionnaire experiment showing that the probabilities estimated in the corpus model seem to be in accordance with speaker’s preferences

3 “levels of organization” involved

1. Lexical level
   - random-effects

2. Level of the combination of two lexical items
   - importance of the collocation variables

3. Syntactic level
   - constraints concerning the structure of the AP and the NP
Other Romance languages display the attributive A alternation

(19) Italian : un triste racconto / un racconto triste
    *a sad story*

(20) Romanian : o tristă poveste / o poveste tristă
    *a sad story*

(21) Portuguese : um imenso escritório / um escritório imenso
    *a huge office*

(22) Spanish : un inmenso escritorio / un escritorio inmenso
    *a huge office*

The study on French suggests that these data should be analyzed by means of experimental approaches.

Using statistical modeling would allow us to observe and quantify cross-linguistic differences and similarities (cf. Bresnan and Hay, 2008)
Thank you for your attention


