Contraintes préférentielles et ordre des mots en français

Juliette Thuilier

Université Paris Diderot & ALPAGE (INRIA — Paris Diderot)

Soutenance de thèse — 28 septembre 2012

Préférences syntaxiques

Exemple des phénomènes d'ordre en français

- Ordre défini catégoriquement
  ex. ordre relatif du déterminant et du nom

(1) a. *les progrès accomplis par les enfants*
b. * *les progrès accomplis par enfants les*

- Ordre choisi selon des préférences
  ex. ordre relatif des compléments du verbe

(2) a. *montrer* [*au public, essentiellement composé de parents,*] [*les progrès accomplis par les enfants*]. (Est-Républicain)
b. *montrer* [*les progrès accomplis par les enfants*] [*au public, essentiellement composé de parents*].

⇒ Le choix d'un ordre dans un certain nombre de phénomènes relève d'une préférence syntaxique

Hypothèses et contraintes préférentielles

2 Hypothèses centrales

1. Préférence syntaxique = combinaison de contraintes préférentielles
2. Contraintes préférentielles = propriétés spécifiques à la langue, dont la syntaxe doit rendre compte

Caractéristiques des contraintes préférentielles

- Affectent l'acceptabilité d'une phrase, et non sa grammaticalité
- Relèvent l'interface entre syntaxe et d'autres dimensions : prosodie, phonologie, sémantique, structure informationnelle
- Caractère préférentiel varie d'une langue à l'autre : une contrainte peut être catégorique dans une langue et préférentielle dans une autre

Problématique méthodologique

Comment étudier les contraintes préférentielles ?

- Leur étude échappe aux méthodes de recherche traditionnellement utilisées en syntaxe
  - Introspection et jugement de grammaticalité non-contrôlé
- Deux méthodes envisagées
  - Analyse statistique de données de corpus
  - Expérimentation linguistique

$\Rightarrow$ Dans ma thèse, je montre qu'il est fructueux d'employer ces méthodes pour analyser deux phénomènes linguistiques du français :

  - Position de l'adjectif épithète par rapport au nom
  - Ordre des compléments du verbe ditransitif dans le domaine postverbal

Objectifs généraux de la thèse

- Sur le plan méthodologique :
  - Utiliser l'analyse de données de corpus pour décrire et modéliser des phénomènes linguistiques
  - Utiliser des expériences pour confirmer / infirmer les résultats sur corpus
- Sur le plan linguistique
  - Position de l'adjectif :
    - **Confrontation** des caractéristiques du phénomène repérées dans la littérature et des données de corpus
    - Proposition d'une **modélisation statistique interprétable** et rendant compte au mieux du phénomène
  - Ordre des compléments verbaux :
    - **Étude exploratoire** de ce phénomène, à partir de données de corpus et de questionnaires élicitant les jugements d'acceptabilité
    - Étude permettant de **comparer** les facteurs significatifs en français avec les tendances générales observées dans d'autres langues

Plan de la thèse

6 chapitres répartis en trois grandes parties

- Situation du problème
  1. Les contraintes préférentielles
  2. Méthodes et outils
- Les adjectifs épithètes en français
  3. Le problème de la position de l'adjectif épithète (État de l'art)
  4. Analyse de données de corpus
- Les compléments postverbaux
  5. L'ordre des dépendants du verbe (État de l'art)
  6. Analyse de données de corpus

Données de corpus

Corpus = la principale source de données, mais 2 limitations

1. Généralisation
   - Comment passer d'observations sur des données de corpus à des propriétés du système de la langue ?
   - Inférence statistique : permet d'inférer les propriétés d'une population à partir d'un échantillon,
     - Ici la population = le système de la langue, l'échantillon = le corpus
   - Mais tributaire de la représentativité de l'échantillon
2. Représentativité
   - Dans l'idéal, l'échantillon devrait être représentatif de la population
   - Dans le cas de la langue, on ne connaît pas la composition de la population générale
   - Pour pallier ce problème :
     - Utilisation de sources de données diverses :
       2 corpus journalistiques : *French Treebank* et Est-Républicain
       2 corpus d'oral : ESTER et C-ORAL-ROM
     - Résultats sur corpus complétés par des questionnaires élicitant le jugement des locuteurs

Inférence statistique et analyse de données

- Régression logistique (modèles à effets mixtes)
  - Modélisation du comportement d'une variable binaire en fonction de variables prédictrices
    - Variable binaire = ordre relatif des compléments ou position de l'adjectif
    - Variables prédictrices = contraintes préférentielles
  - Modèle prédictif : permet de prédire le phénomène sur des exemples nouveaux
  - Modèle à effets mixtes : principalement pour prendre en compte le comportement spécifique de chaque item lexical

$\Rightarrow$ Construction de modèles dans lesquels

- les différentes variables sont interprétables
- les généralisations peuvent être confrontées à des données nouvelles

Étude corrélationnelle et expérience

- Étude corrélationnelle
  - Hypothèse de départ : les préférences détectées en corpus sont en correspondance avec une forme de savoir des locuteurs
  - Objectif : tester la corrélation entre les prédictions du modèle et les préférences des locuteurs
  - Réalisation : questionnaire dans lequel chaque sujet choisit une phrase dans une paire de phrases (l'une des phrases étant extraite des données analysées)
- Approche expérimentale : élicitation de jugements d'acceptabilité
  - Objectif : étude du comportement d'une contrainte préférentielle par le biais des jugements d'acceptabilité
  - Réalisation : questionnaire dans lequel chaque sujet donne un jugement d'acceptabilité pour des phrases dans lesquelles l'intégralité des variables est contrôlée, sauf la variable étudiée qui, elle, est manipulée
  - Analyse statistique pour déterminer si la variable étudiée a un effet significatif sur les jugements

La position de l'adjectif épithète

(3) a. Antéposition : *une agréable soirée*
b. Postposition : *une soirée agréable*

- Nombreux facteurs intervenant dans le choix de la position aux niveaux phonologique, morphologique, lexical, syntaxique, sémantique et discursif
- Hypothèses
  1. Il existe une possibilité générale d'alternance de position pour les mots appartenant à la catégorie 'Adjectif'
  2. Cette alternance est guidée par un ensemble de contraintes préférentielles
- Démarche
  - Validation des contraintes rencontrées dans la littérature à partir de données de corpus
  - Modélisation statistique du phénomène

Description des données

- Extraction de l'ensemble des adjectifs épithètes accompagnant une tête nominale dans le *French Treebank*
  - 13933 occurrences réparties en 1750 lemmes
  - 72.7% de postposition
  - Seuls 171 lemmes adjectivaux alternent de position dans les données
- 28 variables prédictrices étudiées :

longueur de l'adjectif, longueur du SADJ, fréquence de l'adjectif, caractère morphologiquement construit de l'adjectif, classes sémantiques de l'adjectif (intensionnel, évaluatif, indéfini, nationalité, couleur), adjectif en coordination, adjectif modifié, présence d'autres dépendants dans le SN (adjectif antéposé, adjectif postposé, SP, relative), le type de déterminant (défini, démonstratif, possessif, pas de déterminant), fonction du SN (sujet, objet, attribut du sujet, objet non-direct, modifieur), force d'association de la séquence Adjectif-Nom, force d'association de la séquence Nom-Adjectif, hiatus potentiel en postposition, adjectif masculin singulier à consonne latente en présence d'un nom à initiale vocalique.

Conclusions sur le plan linguistique

1. Niveau lexical
   - Caractéristiques formelles

Antéposition	NOM	**Postposition**
court		long
fréquent		rare
simple		construit

   - Caractéristiques lexico-sémantiques : intensionnel, indéfini et évaluatifs favorisent l'antéposition

   ⇒ L'identité du lemme a une place centrale dans le choix de la position, il existe une préférence lexicale

2. Combinaison du nom et de l'adjectif
   - Le nom modifié par l'adjectif influence la position
   - Plus l'ordre Nom$_i$ - Adjectif$_i$ est fréquent, plus la séquence Nom$_i$ - Adjectif$_i$ tend à être produite dans cet ordre

Conclusions sur le plan linguistique

3. Structure du SAdj
   - Plus le SAdj est complexe (coordination et modifieur), plus il a tendance à être postposé
4. Structure du SN
   - Le type de déterminant : défini, possessif et démonstratif favorisent l'antéposition
   - La présence d'autres éléments dans le SN : autre adjectif ou SP favorisent l'antéposition

$\Rightarrow$ Le modèle proposé permet de rendre compte de l'effet combiné des différentes contraintes préférentielles

- Exactitude 100 passes : $\mu = 0.869$ ($\sigma = 0.02$)
- Aire sous la courbe ROC : AUC $= 0.935$

Confrontation des résultats à d'autres sources de données

- Étude corrélationnelle
  - Questionnaire recueillant les préférences de 141 locuteurs natifs sur 29 paires de phrases
  - Corrélation entre proportion d'antéposition prédite par le modèle et proportion d'antéposition préférée par les locuteurs : $\rho = 0.74$ (p$= 3.7 \times 10^{-06}$)
- Position de 171 adjectifs dans des corpus d'autres genres
  - Corpus oraux : ESTER et C-ORAL-ROM
  - Corpus littéraire : 80 oeuvres littéraires du XX^{e} (Wilmet, 1980)

  ⇒ Proportions générales d'antéposition équivalentes à celle du genre journalistique

⇒ Faisceau de résultats convergeant :
  - Analyse de données de corpus (*French Treebank*)
  - Proportions observées dans d'autres genres de corpus
  - Analyse de préférences des locuteurs

Perspectives

- Étendre l'analyse de données à d'autres genres de corpus pour
  - évaluer la généralité du modèle proposé
  - vérifier que les genres de discours ont un impact limité
- Approfondir la corrélation entre observations en corpus et préférences des locuteurs dans l'objectif de vérifier l'idée selon laquelle les résultats obtenus sur corpus sont en correspondance avec une forme de savoir langagier
- Utiliser des expériences autres que l'élicitation de préférence, en compréhension et en production, afin d'évaluer à quels niveaux ces contraintes interviennent

Une étude exploratoire

Ordre SN-SP (4a) ou SP-SN (4b)

(4) a. montrer [*au public, essentiellement composé de parents,*] [*les progrès accomplis par les enfants*]. (Est-Républicain)

b. montrer [*les progrès accomplis par les enfants*] [*au public, essentiellement composé de parents*].

- Phénomène peu étudié en français
- Étude des contraintes rencontrées pour ce type de phénomènes dans d'autres langues
  - Pronominalité
  - Poids (différence de longueur du SN et du SP en mots)
  - Caractère figé de la séquence V SP
  - Caractère animé des référents
  - Caractère défini du SN et du SP
  - Lemme verbal et sa classe sémantique

Description des données

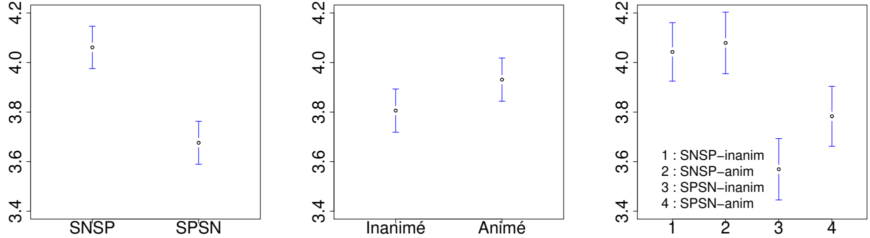
- Les données de corpus
  - Écrit journalistique (*French Treebank* et Est-Républicain), oral (ESTER et C-ORAL-ROM)
  - Patrons extraits : V SN SP ou V SP SN, avec SN et SP sous-catégorisés par V
  - 1434 occurrences réparties en 182 lemmes verbaux
  - 70.4% d'ordre SNSP
  - Données manuellement annotées pour le caractère animé des référents et les classes sémantiques des verbes (Dubois and Dubois-Charlier, 1997)
- Modélisation
  - Variables ayant une influence significative sur l'ordre : poids et lemme verbal associé à sa classe sémantique
  - Autres variables non-significatives dans les données

Biais verbaux

- Le lemme verbal, désambiguïsé grâce à sa classe sémantique en contexte, influe sur l'ordre des compléments
- Deux explications proposées concernant ces biais verbaux
  1. Niveau syntaxique : reprise de l'hypothèse de Stallings et al. (1998) (*Verb disposition hypothesis*)
     - Verbes avec plusieurs réalisations possibles de l'objet direct (SN, subordonnée, infinitive) tendent à préférer l'ordre SP-SN
     - D'après les biais verbaux estimés par le modèle, les verbes de communication sous-catégorisant une subordonnée ou une infinitive vérifient cette hypothèse (*confirmer*, *dicter*, *expliquer*, *montrer*, *proposer*, *préconiser*, *apprendre*, *demander*)
  2. Niveau sémantique : reprise du travail de Schmitt (1987a,b)
     - Verbes exprimant une relation entre un patient et un état résultant : *compléter*, *convertir*, *ériger*, *transformer*, *faire*
     - La relation établie entre les deux arguments par le verbe a un impact sur l'ordre des compléments
     - Soit R(β, γ) la relation qu'instaure le prédicat R entre ses deux arguments β et γ, si β est affecté par R(β, γ) et que γ correspond à l'état résultant de β, alors les arguments ont tendance à apparaître linéairement dans l'ordre β γ.

Caractère animé

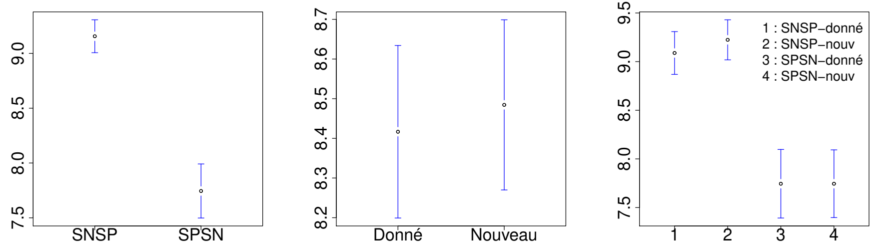
- Tendance générale à travers les langues : *animé avant inanimé*
- Données de corpus ne confirment pas cette tendance
- Questionnaire d'élicitation du jugement d'acceptabilité :
  - Hypothèse : les locuteurs préfèrent l'ordre *animé avant inanimé* lorsque les autres variables sont neutralisées
  - Jugements de 38 locuteurs natifs sur 16 phrases
  - Résultats : pas d'effet significatif du caractère animé sur l'acceptabilité des phrases



⇒ Pas d'argument pour dire que le français suit la tendance générale

Opposition *donné/nouveau*

- Tendance générale à travers les langues : *donné avant nouveau*
- Données de corpus annotées selon les catégories de Prince (1981) ne confirment pas cette tendance
- Questionnaire d'élicitation du jugement d'acceptabilité :
  - Hypothèse : les locuteurs préfèrent l'ordre *donné avant nouveau* lorsque les autres variables sont neutralisées
  - Jugements de 28 locuteurs natifs sur 16 phrases
  - Résultats : pas d'effet significatif du caractère donné ou nouveau sur l'acceptabilité des phrases



⇒ Pas d'argument pour dire que le français suit la tendance générale

Conclusion et perspectives

- Rôle du poids et de la singularité du lemme verbal
- Caractère animé et opposition *donné*/*nouveau* : pas de preuve que le français suit les tendances générales observées dans d'autres langues

- Perspectives
  - Modélisation à améliorer en incluant de nouvelles variables (structure informationnelle, lien sémantique entre verbe et compléments. . .)
  - Approfondir l'analyse des biais verbaux, notamment en augmentant le nombre de données par lemme
  - Introduire des contraintes relatives au traitement cognitif : éviter l'ambiguïté, effet de persistance syntaxique (*priming*)
  - Confrontation des résultats sur corpus avec des études corrélationnelles et des expériences psycholinguistiques permettant de vérifier si les résultats obtenus sur corpus sont en correspondance avec une forme de savoir langagier

Conclusion générale

- Sur le plan linguistique
  - Description et modélisation du choix de la position de l'adjectif
  - Étude exploratoire sur l'ordre des compléments postverbaux
  - Mise en évidence de l'importance de la singularité lexicale au niveau de l'ordre entre les mots et au niveau de l'ordre entre les syntagmes
  - Approche non-réductionniste de l'ordre : la linéarisation des constituants est guidée par une variété de facteurs
- Sur le plan méthodologique
  - Méthodes permettant d'étudier et de modéliser des problèmes de préférences en syntaxe
  - Méthodologie comme procédure de validation (adjectifs) et comme procédure de découverte (compléments du verbe)
  - Modèles formels permettant de capter deux dimensions en tension
    - les spécificités lexicales
    - les principes généraux de l'organisation de la langue

Dubois, J. and F. Dubois-Charlier (1997). *Les verbes français*. Paris : Larousse-Bordas.

Prince, E. F. (1981). Toward a taxonomy of given-new information. In P. Cole (Ed.), *Radical Pragmatics*, pp. 223–256. New York : Academic Press.

Schmitt, C. (1987a). À propos de l'impact de la sémantique sur la séquence des compléments d'objets en français moderne. *Travaux de linguistique et de littérature 25*(1), 283–298.

Schmitt, C. (1987b). Sémantique et prédétermination de l'ordre des mots en français contemporain. *Travaux de linguistique 14/15*, 21–31.

Stallings, L. M., M. C. MacDonald, and P. G. O'Seaghdha (1998). Phrasal ordering constraints in sentence production : Phrase length and verb disposition in heavy- NP shift. *Journal of Memory and Language 39*(3), 392–417.

Wilmet, M. (1980). Antéposition et postposition de l'épithète qualificative en français contemporain : matériaux. *Travaux de linguistique 7*, 179–201.