

Préférences concernant l'ordre relatif des compléments du verbe en français

Juliette Thuilier¹, Anne Abeillé² & Benoît Crabbé¹

¹Université Paris Diderot & Alpage

²Université Paris Diderot & LLF

10 septembre 2011

Colloque AFLS
8-10 septembre 2011
Nancy

Introduction

En français l'ordre des constituants postverbaux est relativement libre (Blinkenberg, 1928; Abeillé and Godard, 2000)

- le verbe a une position fixe
- les constituants postverbaux sont ordonnés librement
- sauf : les adverbes et quantifieurs légers (Abeillé and Godard, 2001) et les noms nus compléments (Abeillé and Godard, 2004) qui ont une position plus contrainte

Notre objet d'étude : ordre des compléments post-verbaux

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

- ⇒ étude des facteurs influençant l'ordre relatif des compléments des verbes ditransitifs
- ⇒ approches quantitatives et expérimentales

L'alternance dative en anglais

- (2) a. Mary gave a book to John
b. Mary gave John a book

Série de travaux : Bresnan et al. (2007); Bresnan (2007); Bresnan and Nikitina (2009); Bresnan and Ford (2010)

- **sur le plan théorique** : mise à jour du rôle de facteurs hétérogènes : longueur relative des compléments, statut discursif, caractère pronominal ou animé, classe sémantique du verbe...
 - **sur le plan méthodologique** : analyse des données de corpus grâce à la statistique inférentielle et utilisation de questionnaires psycholinguistiques
- ⇒ montre que l'on peut tirer des généralités sur des questions de préférences à partir de l'étude de corpus
- ⇒ montre que les préférences dégagées sur corpus sont en correspondance avec des préférences observées chez les locuteurs

Travaux précédents

- Modèle à vocation universelle
 - ▶ mêle facteurs linguistiques et conceptuels
 - ▶ cherche à prédire l'alignement harmonique des arguments

Tendances générales

Pour toutes les langues,

animé précède **non-animé**
pronominal précède **non-pronominal**
défini précède **indéfini**

Hawkins (1994) :

pour les langues SVO
pour les langues SOV

court précède **long**
long précède **court**

- Expériences et études quantitatives réalisées pour
 - ▶ allemand (sujet-OD et OD-OI) (Kempen and Harbusch, 2004)
 - ▶ espagnol (sujet-OD) (Prat-Sala and Branigan, 2000)
 - ▶ grec (sujet-OD) (Branigan and Feleki, 1999)
 - ▶ japonais (Tanaka et al., 2011)
 - ▶ turc, néerlandais...

Objectifs

- mettre à jour les contraintes préférentielles guidant le choix de l'ordre des compléments en français
- pouvoir comparer avec les tendances générales observées dans les autres langues

Outils

pour observer les préférences :

- corpus annoté en syntaxe : French Tree Bank (Abeillé et al., 2003; Abeillé and Barrier, 2004)
- autres corpus : Est-Républicain et ESTER
- un questionnaire psycholinguistique

pour modéliser et généraliser :

- méthodes statistiques inférentielles : la régression logistique (Agresti, 2007)

- 1 1. Etude sur corpus
 - Corpus et extraction de données
 - Modélisation

- 2 2. L'expérience psycholinguistique
 - Le questionnaire
 - Résultats

- 3 3. Conclusions

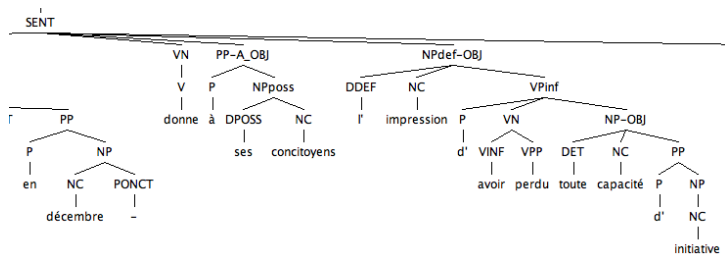
Corpus et extraction de données

Notre corpus

- phrases contenant un V suivi **uniquement** de 2 compléments
- **motif** : V SN SP ou V SP SN

French Tree Bank (FTB)

- corpus annoté syntaxiquement
- extraction automatique sur une sous-partie du corpus (12 000 phrases)
- **325 occurrences** / 146 lemmes verbaux



Est-Républicain (ER)

- version lemmatisée avec morfette (Grzegorz Chrupała and van Genabith, 2008) : 148 000 000 occurrences de mots, 662 000 lemmes
- extraction des phrases contenant le motif
 - 1 sélection manuelle des phrases
 - 2 analyse syntaxique automatique
 - 3 puis correction manuelle
- **418 occurrences** / 18 lemmes verbaux

ajouter, annoncer, assurer, devoir, dire, donner, expliquer, faire, mettre, montrer, passer, permettre, porter, prendre, réduire, rendre, trouver, vendre

ESTER

- corpus radiophoniques transcrits (60h d'enregistrement)
- même procédure que pour l'Est-Républicain
- **239 occurrences** / 23 lemmes verbaux

accorder, ajouter, annoncer, appeler, demander, dire, donner, expliquer, faire, lancer, mettre, montrer, obtenir, offrir, passer, porter, prendre, présenter, proposer, réduire, rendre, trouver

Variables

- 1 SN-SP : longueur relative du SN et du SP en nombre de mots : nombre de mots du SN - nombre du mots du SP (échelle logarithmique pour la modélisation)
- 2 PREPN : nature du SP : Preposition + nom nu
- 3 SNDEF : Caractère défini du SN
- 4 SPDEF : Caractère défini du SP
- 5 SNPRO : Caractère pronominal du SN
- 6 SPPRO : Caractère pronominal du SP
- 7 SUB : verbe et ses compléments dans une subordonnée
- 8 LEMMEVB : lemme verbal
- 9 PREP : préposition introduisant le SP (*à, de* ou autre)
- 10 SNANIM : caractère animé du SN (annoté manuellement selon les catégories de Zaenen et al. (2004))
- 11 SPANIM : caractère animé du SP (idem)

Description générale des données

- 982 exemples avec une préférence pour l'ordre SN-SP : **58%**

3 sources de variation dans les données

1 le corpus :

	corpus	FTB	ER	ESTER
ordre SN-SP	58%	68.6%	46.4%	64%
ordre SP-SN	42%	31.4%	53.6%	36%

2 la préposition :

	A	DE	Autres
ordre SN-SP	64%	28.4%	61.3%
ordre SP-SN	36%	71.6%	38.7%

3 le lemme verbal

	<i>montrer</i>	<i>donner</i>	<i>mettre</i>
ordre SN-SP	37.8%	80.2%	46.7%
ordre SP-SN	62.2%	19.8%	53.3%

- (3) a. ... **montrant** à son copain Maxime ses chefs-d'oeuvre (ER)
b. ... **montrer** aux Forgerons et aux passionnés de l'agglomération quelques bijoux de cette époque (ER)
- (4) a. ... **donne** des cours aux enfants (ER)
b. ... **donnerait** plus de poids politique à la trêve (ESTER)
- (5) a. ... **avait mis** des documents dans un coffre d'une banque à l'étranger (ER)
b. ... **mettre** en place un système (ESTER)

⇒ nouvelle variable pour distinguer les emplois d'un même lemme :
LEMMEPREP = lemme verbal + préposition (165 valeurs)

- exemple : porter+A : *porter la bonne nouvelle à sa fiancée*
porter+P : *porter l'opérette dans les petites villes*

Travail en cours : annotation en classes sémantiques selon Dubois and Dubois-Charlier (1997)

Exemple : 2 emplois de *mettre*

- 2 types d'emploi avec les prépositions "Autres"
 - ▶ locatif : *avait mis des documents **dans un coffre d'une banque à l'étranger***
 - ▶ collocationnel : *mettre **en place** un système*
- emplois capturés par 2 classes de Dubois and Dubois-Charlier (1997)
 - ▶ classe **L** (locatif)
 - ▶ classe **R** (réalisation, mise en état)
- permet d'observer des tendances très différentes

	<i>mettre</i> L	<i>mettre</i> R
ordre SN-SP	69%	30%
ordre SP-SN	31%	70%

Inférence statistique : Régression logistique (Agresti, 2007)

- permet de modéliser le comportement d'une variable binaire en fonction de variables prédictives
- ordre de 2 compléments = variable binaire
 - ordre SN SP = 0
 - ordre SP SN = 1

Fonction logistique

Fonction logistique = fonction à valeurs dans l'intervalle $[0, 1]$

$$\pi_{\text{spsn}} = \frac{e^{\beta\mathbf{X}}}{1 + e^{\beta\mathbf{X}}} \quad (1)$$

où

- π_{spsn} = probabilité de l'ordre SPSN
- β = coefficients de régression $\alpha, \beta_0 \dots \beta_n$
- \mathbf{X} = variables prédictives $X_0 \dots X_n$

Prise en compte de la variation des données dans le modèle

- comment capturer la spécificité des données selon chaque corpus, selon l'item verbal et selon la préposition ?

⇒ utilisation de modèles à **effets aléatoires**

- idée : en plus du modèle général, chaque groupe de données se voit attribuer un coefficient propre
- par ex. : chaque corpus a un coefficient propre qui reflète son comportement particulier

Le modèle

Effets aléatoires

Groupe	Variance	Ecart-Type
lemmePrep	1.76719	1.32936
corpus	0.22209	0.47127

Nbre d'obs : 982, groupes : lemmePrep, 165; corpus, 3

Effets fixes

	Coef. estimé	Erreur-Type	Valeur Z	P(> z)
Intercept	-1.356374	0.533032	-2.545	0.0109 *
SN-SP	2.904711	0.211958	13.704	<2e-16 ***
PrepNom	1.128978	0.534583	2.112	0.0347 *
SNdef	-0.182574	0.263311	-0.693	0.4881
SPdef	-0.071076	0.358507	-0.198	0.8428
SNpro	-0.732035	0.756112	-0.968	0.3330
SPpro	0.835603	0.632876	1.320	0.1867
Sub	0.387202	0.262482	1.475	0.1402
SNanim	0.001898	0.416942	0.005	0.9964
SPanim	-0.066371	0.308412	-0.215	0.8296

Importance de la longueur relative confirmée

- Longueur différente : variable qui permet de déterminer l'ordre dans **85,7%** des cas

(6) ... ajoutera [*sa touche*]_{SN} [*à l'embellissement de la commune*]_{SP} (ER)

(7) ... a présenté [*au parlementaires européens*]_{SP} [*son plan pour la présidence italienne de l'Europe qu'il exerce depuis hier*]_{SN} (ESTER)

SN plus court que le SP : **91%** de SN-SP

SN plus long que le SP : **79.6%** de SP-SN

- Longueur égale : forte préférence pour l'ordre SN-SP avec **67.7%**

La nature du SP : Preposition + Nom nu

- malgré la corrélation avec la variable de la longueur, ce facteur reste significatif
- préférence beaucoup plus forte pour l'ordre SP-SN (82%), par rapport aux autres SP de 2 mots (58%)

(8) *mettre [en péril]_{SP} [la bonne marche de l'État réunié]_{SN}* (FTB)

⇒ lien de type collocationnel entre le verbe et le SP favorise l'ordre SP-SN

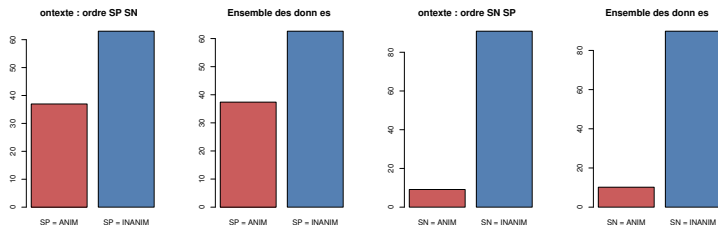
SP = Prep + Nom : **82%** de SP-SN

Autres SP : **38%** de SP-SN

Résultats

Non significativité du caractère animé, de la pronominalité et du caractère défini

- résultats différents par rapport à ce qui est observé dans les autres langues
- **pronominalité** : phénomène pas vraiment comparable aux autres langues, cliticisation massive des pronoms en français
- **caractère animé** : résultat plus surprenant, à explorer



- **caractère défini** : résultat différent dans l'expérience

1. Etude sur corpus
 - Corpus et extraction de données
 - Modélisation

2. L'expérience psycholinguistique
 - Le questionnaire
 - Résultats

3. Conclusions

2. Expérience psycholinguistique

- les données issues de corpus présentent des corrélations difficiles à contrôler
- But du questionnaire : contrôler la longueur pour pouvoir observer les autres variables

Notre questionnaire

- Consigne : choisir la continuation préférée parmi deux alternatives
- 23 items testés, pour lesquels :
 - ▶ la longueur est neutralisée (nombre de mots du SN = nombre de mots du SP)
 - ▶ ordre des items randomisé
 - ▶ ordre des alternatives randomisé
 - ▶ présence de distracteurs pour ne pas attirer l'attention des sujets sur le but du questionnaire
- 25 sujets, étudiants en Lettres (L2, Paris 7)

- phrases extraites des corpus journalistiques (72% SN-SP)
- phrases contextualisées
- préférences exprimées à l'aide d'une *échelle de Likert* : 5 niveaux
 - ▶ 1 = ordre SP SN préféré
 - ▶ 3 = pas de préférence
 - ▶ 5 = ordre SN SP préféré

Exemple

Une nuit d'été la mère de Pierre lui annonce qu'elle a arrêté la date de son mariage avec Lucie. Pierre fonce dans la nuit

A. *porter à sa fiancée la bonne nouvelle*

B. *porter la bonne nouvelle à sa fiancée*

A **B**

- préférence générale en correspondance avec le corpus : préférence moyenne pour l'ordre SN-SP = 3.5
- modélisation des données expérimentales : régression linéaire à effets aléatoires

Effets fixes

	Coef. estimé	P (> t)
Intercept	3.5585	0.0000
SNdef	-0.4914	0.0225
SPpro	-0.4465	0.5172
SNanim	0.0323	0.9964
SPanim	-0.2410	0.5263
SPdef	0.4713	0.3151
PrepNom	-0.2466	0.6938
Sub	-0.0577	0.8815

- mêmes tendances que pour le corpus : pronominalité et caractère animé non significatifs
- différence : effet de SNDEF ($p=0.02$)

1. Etude sur corpus
 - Corpus et extraction de données
 - Modélisation

2. L'expérience psycholinguistique
 - Le questionnaire
 - Résultats

3. **Conclusions**

Conclusions

- Observation et modélisation de la question de la préférence dans l'ordonnement des compléments post-verbaux
- Confirmation de l'importance de la longueur relative des compléments
- autres facteurs pertinents : PREPN et dans une moindre mesure SNDEF.
- Caractère animé et pronominalité non pertinents, contrairement à ce qui est observé dans les autres langues
- Travail à poursuivre : autres facteurs à étudier (classe sémantique du verbe, statut informationnel, effet de priming...), autres types de corpus (notamment oral spontané)

Conclusions

Méthodes et outils

- Corpus et statistique inférentielle dans le but est d'inférer des préférences générales (de langue) à partir d'observations sur un échantillon restreint (corpus)
 - Préférences étayées à l'aide d'un questionnaire psycholinguistique
- ⇒ complémentarité de l'approche quantitative sur corpus et de l'expérimentation
- Etudes sur les préférences exigent données richement annotées
- Echantillon biaisé ⇒ nécessité d'un treebank de l'oral

- Abeillé, A. and N. Barrier (2004). Enriching a french treebank. In *Proceedings of Language Ressources and Evaluation Conference (LREC)*, Lisbon.
- Abeillé, A., L. Clément, and F. Toussanel (2003). Building a treebank for french. In *Treebanks*. Dordrecht : Kluwer.
- Abeillé, A. and D. Godard (2000). French word order and lexical weight. In R. Borsley (Ed.), *The Nature and Function of Syntactic Categories (Syntax and Semantics 32)*, pp. 325–358. New-York : Academic Press.
- Abeillé, A. and D. Godard (2001). A class of lite adverbs in french. In J. Camps and C. Wiltshire (Eds.), *Romance syntax, semantics and their L2 acquisition*, pp. 9–25. Amsterdam : John Benjamins.
- Abeillé, A. and D. Godard (2004). De la légèreté en syntaxe. *Bulletin de la Société de Linguistique de Paris XCIX*(1), 69–106.
- Agresti, A. (2007). *An introduction to categorical data analysis*. Wiley interscience.
- Blinkenberg, A. (1928). *L'ordre des mots en français moderne*. Copenhague : Munskgaard.
- Branigan, H. P. and E. Feleki (1999). Conceptual accessibility and serial order in greek language production. In M. Hahn and S. C. Stoness (Eds.),

Proceedings of the 21st Conference of the Cognitive Science Society, Mahwah, pp. 96–101. Erlbaum.

Bresnan, J. (2007). Is syntactic knowledge probabilistic ? experiments with the english dative alternation. In S. Featherston and W. Sternefeld (Eds.), *Roots : Linguistics in Search of Its Evidential Base*, pp. 77–96. Berlin : Mouton de Gruyter.

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

Bresnan, J. and M. Ford (2010). Predicting syntax : processing dative constructions in american and australian varieties of english. *Language* 86(1), 168–213.

Bresnan, J. and T. Nikitina (2009). The gradience of the dative alternation. In L. Uyechi and L. H. Wee (Eds.), *Reality Exploration and Discovery : Pattern Interaction in Language and Life*, Stanford, pp. 161–184. CSLI Publications.

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

Grzegorz Chrupała, G. D. and J. van Genabith (2008). Learning morphology with morfette. In *In Proceedings of LREC 2008*, Marrakech, Morocco. ELDA/ELRA.

Hawkins, J. (1994). *A performance theory of order and constituency*. Cambridge : Cambridge University Press.

Kempen, G. and K. Harbusch (2004). A corpus study into word order variation in german subordinate clauses : Animacy affects linearization independently of grammatical function assignment. In T. Pechmann and C. Habel (Eds.), *Multidisciplinary approaches to language production*, pp. 173–181. Berlin : Mouton de Gruyter.

Prat-Sala, M. and H. P. Branigan (2000). Discourse constraints on syntactic processing in language production : A cross-linguistic study in english and spanish. *Journal of Memory and Language* 42, 168–182.

Tanaka, M., H. Branigan, and M. Pickering (2011). Conceptual influences on word order and voice in sentence production : Evidence from japanese. *Journal of Memory and Language* 65(3), 168–182.

Zaenen, A., J. Carletta, G. Garretson, J. Bresnan, A. Koontz-Garboden, T. Nikitina, M. C. O'Connor, and T. Wasow (2004). Animacy encoding in

english : why and how. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, DiscAnnotation '04, Stroudsburg, PA, USA, pp. 118–125. Association for Computational Linguistics.

Exemples de l'expérience

La Poste, d'une part, France Télécom, d'autre part, disposent désormais d'une autonomie financière.

A. *Ce changement de statut va contraindre les deux exploitants publics à plus de rigueur*

B. *Ce changement de statut va contraindre à plus de rigueur les deux exploitants publics*

A **B**

*Vendredi matin, les enfants de l'école ont parcouru les rues du village à la recherche des vieux papiers. Une bonne leçon d'écologie pour les enfants...
Demandez donc à ce petit garçon qui dès ce matin*

A. *annonçait à sa maman une journée extraordinaire !*

B. *annonçait une journée extraordinaire à sa maman !*

A **B**

Pas moins de 320 enfants ont participé à la fête de la musique. Après le blues de Manu Grilloit proposé le matin, les enfants ont ainsi présenté en public les fruits du travail musical de l'année :

A. *"C'est intéressant de montrer aux autres nos productions", explique Sarah.*

B. *"C'est intéressant de montrer nos productions aux autres", explique Sarah.*

A **B**

Comparée à la dette totale des pays en développement, celle de l'Afrique représente un cinquième environ du total. Le chiffre paraît faible. Il l'est beaucoup moins

A. *si l'on compare au produit national brut le montant de la dette.*

B. *si l'on compare le montant de la dette au produit national brut.*

A **B**