Lemme verbal et classe sémantique dans l'ordonnancement des compléments postverbaux

Juliette Thuilier

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100172

Lemme verbal et classe sémantique dans l'ordonnancement des compléments postverbaux

Juliette Thuilier

Univ Paris Diderot, Sorbonne Paris Cité, ALPAGE, UMR-I 001 INRIA
jthuilier@linguist.jussieu.fr

1 Introduction

Dans cet article, nous étudions l'ordre des compléments postverbaux en français, à travers un corpus de 956 phrases extraites du French Tree Bank, de l'Est-Républicain et d'ESTER. Nous montrons plus particulièrement que l'item verbal ainsi que sa classe sémantique influencent significativement l'ordre choisi.

L'ordre des constituants postverbaux en français est relativement libre (Blinkenberg, 1928 ; Abeillé & Godard, 2000). Le verbe a une position fixe et les constituants sont ordonnés librement après lui. Certains travaux ont montré que les adverbes et quantifieurs légers (Abeillé & Godard, 2001) et les noms nus compléments (Abeillé & Godard, 2004) ont une position plus contrainte dans la zone postverbale. Dans le cadre du travail présenté ici, notre objet d'étude est restreint à l'ordre des compléments du verbe, dans les cas où il n'y a pas d'autres éléments dans la zone postverbale, comme dans l'exemple (1).

(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].

Blinkenberg (1928) estime que l'ordre Objet Direct - Objet Indirect (OD-OI) est l'ordre par défaut en français, mais qu'il existe des facteurs favorisant l'ordre inverse. Lorsque l'OD et l'OI sont respectivement réalisés sous la forme d'un SN et d'un SP, l'ordre OI-OD est favorisé par un OD plus long que l'OI, par un OI répété ou déjà connu ainsi que par la possibilité d'éviter une ambiguïté au niveau du rattachement du SP, comme dans l'exemple (2b).

(2) a. Il sut parler [le langage qui convenait] [à la France] (rattachement du SP ambigu)

b. Il sut parler [à la France] [le langage qui convenait]

Berrendonner (1987) reprend, sous une autre formulation, les facteurs dégagés par Blinkenberg (1928) et postule que ces facteurs, parfois contradictoires, entrent en conflit et que « *les locuteurs ne choisissent pas de façon purement aléatoire, mais plutôt selon des stratégies stables, régulières et formalisables* » (Berrendonner, 1987 : 18). À la suite de Blinkenberg (1928) et Berrendonner (1987), nous faisons l'hypothèse que l'ordre des compléments postverbaux est un phénomène multifactoriel et que l'influence des différents facteurs est formalisable.

Le type de phénomène que nous nous proposons d'étudier permet d'observer le rôle de contraintes préférentielles, contraintes qui ne déterminent pas la grammaticalité d'une séquence, mais influent sur le choix d'un ordre ou d'une structure. Ces contraintes préférentielles ont été l'objet de plusieurs travaux concernant l'anglais, notamment l'alternance dative (Gries 2003, Bresnan et al. 2007 ; Bresnan & Hay, 2008 ; Bresnan & Ford 2010), l'ordre verbe particule (Wasow, 2002), l'alternance génitive (Rosenbach, 2005). Ces études ont permis de mettre à jour le rôle de facteurs hétérogènes, tels que la longueur des constituants, le statut discursif, la pronominalité, le caractère animé des référents, la classe sémantique du verbe… Ces travaux ont également mis en oeuvre des méthodes statistiques permettant de tirer des généralités sur des questions de préférences à partir de l'étude de corpus. De plus, à partir d'expériences psycholinguistiques, ces travaux ont montré que les préférences observées en corpus sont en correspondance avec des préférences chez les locuteurs. Nous nous plaçons dans la lignée de ces travaux

Article available at http://www.shs-conferences.org or http://dx.doi.org/10.1051/shsconf/20120100172

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100172

et nous présentons une étude basée sur données extraites de corpus pour lesquelles nous proposons une modélisation statistique inspirée de Bresnan et al. (2007) et Bresnan & Ford (2010).

Dans cet article, nous nous intéressons particulièrement à l'influence du lemme verbal ainsi qu'à celle de sa classe sémantique sur l'ordonnancement des compléments postverbaux. Sur la base d'observations de corpus, Schmitt (1987a, b) a introduit l'idée selon laquelle certains verbes présentent un ordre déterminé par la sémantique, indépendamment des critères de longueur ou de structure informationnelle. Parmi ce type de verbes, il rassemble *transposer*, *traduire*, *remplacer*, *troquer*, *substituer*, *comparer*, *faire* (dans la construction « *faire de quelque chose quelque chose* »)... Nous reprenons cette idée et nous l'élargissons à l'ensemble des verbes sous-catégorisant deux compléments. Cependant, nous nous écartons de Schmitt (1987a, b) dans la mesure où nous estimons que, pour tous les verbes, les deux ordres sont possibles, c'est-à-dire grammaticaux, même si certains verbes favorisent très nettement un ordre par rapport à l'autre.

L'article est organisé de la façon suivante. Dans la partie 2, nous présenterons les données utilisées, les variables prises en compte ainsi que la méthode de modélisation (régression logistique à effets mixtes). La partie 3 sera consacrée à l'étude de l'influence du lemme verbal sur l'ordre des compléments à travers des observations dans nos données et la prise en compte du lemme verbal dans la modélisation. Dans la partie 4, c'est le lemme verbal associé à sa classe sémantique annotée en contexte qui sera examiné. Comme pour le lemme verbal seul, nous exposerons les observations dans les données ainsi que les effets de la prise en compte de la classe sémantique dans la modélisation. Dans la partie 5, nous évoquerons deux pistes permettant d'expliquer le rôle du lemme verbal et de la classe sémantique dans le phénomène étudié. Enfin, la partie 6 sera consacrée à la conclusion et aux perspectives.

2 La méthode

Pour étudier le rôle du lemme verbal et de sa classe sémantique dans l'ordonnancement des compléments en français, nous utilisons des données issues de trois corpus : la sous-partie annotée en fonctions du French Tree Bank (FTB, Abeillé & Barrier 2004), le corpus de l'Est-républicain (ER) disponible sur le site du CNRTL[i] et le corpus ESTER distribué par ELRA. Nous avons procédé à l'analyse de ces données en utilisant une méthode reposant sur les statistiques inférentielles : la régression logistique (Agresti, 2007). En tant que méthode de statistique inférentielle, cette dernière permet de généraliser au-delà de l'échantillon étudié. De plus, elle est conçue pour la modélisation d'une variable binaire, ce qui nous intéresse dans la mesure où nous cherchons à décrire et modéliser l'ordre relatif du SN et du SP. Par ailleurs, c'est une méthode particulièrement adaptée à la modélisation des phénomènes de langue car elle n'implique pas d'hypothèses de normalité sur les données[ii], contrairement à l'analyse discriminante des données, utilisée par exemple, dans Gries (2003). Enfin, contrairement à la théorie de l'optimalité (Prince & Smolensky, 2004; Legendre et al., 2001), ce type d'approche permet de rendre compte des effets de *gang-up* (Jäger & Rosenbach, 2006), à savoir des effets où plusieurs contraintes de faible poids s'unissent pour contredire une contrainte unique de poids important.

2.1 Extraction des données sur corpus

Le FTB est un corpus constitué d'articles du Monde (de 1989 à 1993) annotés syntaxiquement. La sous-partie utilisée contient 12 000 phrases. À partir de l'annotation en constituants et en fonctions, nous avons extrait automatiquement les motifs V SN SP et V SP SN, avec SN et SP, compléments de V. Nous avons obtenu 325 phrases contenant 148 lemmes verbaux différents. Nous avons ensuite complété ces données avec des phrases extraites des corpus ER et ESTER. ER est un corpus journalistique pour lequel nous avons utilisé une version lemmatisée[iii] (Seddah et al. 2012) contenant 148 millions de mots et 662 000 lemmes. ESTER est un corpus radiophonique transcrit contenant l'équivalent de 60h d'enregistrement. Pour ces deux corpus, nous avons sélectionné manuellement les phrases contenant les motifs en utilisant les lemmes les plus fréquents des données extraites du FTB. Ces phrases ont fait l'objet d'une analyse syntaxique automatique, puis d'une correction manuelle. Ainsi, nous avons 378 phrases et 18 lemmes

pour ER ainsi que 239 phrases et 23 lemmes pour ESTER. Parmi les 956 phrases, on observe 577 phrases avec l'ordre V SN SP, soit 60.4% des données. Les données extraites des corpus sont organisées sous forme d'une table de données dans laquelle nous avons listé les phrases et l'ordre attesté, représenté par la variable ORDRE et qui peut avoir deux valeurs : SN-SP ou SP-SN. Nous avons enrichi cette table de données en y ajoutant des variables pouvant influencer l'ordre relatif des compléments. Ainsi, nous avons testé le rôle de la longueur du SN et du SP, du caractère défini du SN et du SP, de la pronominalité du SN et du SP, du caractère animé[iv] du SN et du SP ainsi que les effets de figement entre le SP et le verbe[v].

2.2 Les variables

Dans cette partie, nous présentons les variables que nous avons étudiées à partir des données de corpus et à l'aide de questionnaires psycholinguistiques, en dehors de celles concernant l'item verbal et sa classe sémantique. Nous montrons notamment que les seuls facteurs qui apparaissent comme significatifs sont la longueur des constituants et le caractère figé de la séquence V SP (Thuilier et al., 2011b).

2.2.1 La pronominalité

La pronominalité n'apparaît pas significative, contrairement à ce qui est observé en anglais. On peut supposer que c'est en raison de la cliticisation massive des pronoms que ce facteur n'intervient pas en français. En effet, dans nos données, seuls 2.0% des SN et 3.6% des SP sont pronominaux.

2.2.2 Le caractère animé

Le caractère animé des référents n'a pas d'effet significatif sur l'ordre des compléments d'après les données que nous étudions. Ces observations ont été confirmées au moyen d'un questionnaire psycholinguistique (Thuilier et al., 2011a). Ce résultat semble indiquer que le français se distingue des langues comme l'anglais ou l'allemand pour lesquelles on observe l'influence du caractère animé sur l'ordonnancement des compléments verbaux (pour l'anglais, Bresnan et al, 2007 ; pour l'allemand, Kempen & Harbusch, 2004).

2.2.3 Le caractère défini

Dans nos données, le caractère défini du SN et du SP n'apparaît pas comme un facteur significatif, comme le montrent les proportions présentées dans le tableau 1.

	Total	SN défini	SP défini	SN indéfini & SP défini
ordre SN-SP	60.4%	59.7%	64%	63.1%

Tableau 1 : Proportions d'ordre SN-SP selon le caractère défini du SN et du SP

La proportion d'ordre SN-SP reste quasi équivalente, que le SN et le SP soient définis ou non. D'un point de vue statistique, les différences observées ne sont pas significatives, ce qui indique que le caractère défini n'est pas un facteur influençant l'ordre des compléments dans nos données. Cet élément va à l'encontre de l'affirmation de Berrendonner (1987) selon laquelle il existe une tendance à placer les compléments dans un ordre qui « *va du défini à l'indéfini* ». La dernière colonne du tableau 1 montre bien que l'hypothèse de Berrendonner (1987) n'est pas vérifiée : la proportion de SN-SP ne diminue pas quand le SN est défini et le SP indéfini. Si l'on considère que le défini et l'indéfini constituent une approximation des référents respectivement accessibles et inaccessibles, les données dont nous disposons suggèrent que le contraste accessible/inaccessible n'est pas pertinent pour rendre compte de l'ordre des compléments.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100172

2.2.4 Le caractère donné du référent

Dans le même ordre d'idée, on pourrait supposer que les compléments postverbaux ont tendance à être ordonnés selon le statut des référents, en faisant apparaître les référents donnés puis les référents nouveaux, comme le suggère par exemple Berrendonner (1987 : 10). Les données dont on dispose ne sont pas annotées pour l'opposition donné/nouveau. Nous avons donc tenté d'évaluer l'impact de ce facteur au moyen d'un questionnaire psycholinguistique. Ce questionnaire visait à tester l'effet du caractère donné ou nouveau du SP sur la préférence pour l'ordre SP-SN. Le questionnaire était composé de 16 phrases, pour lesquelles les autres contraintes susceptibles d'intervenir dans la préférence pour un ordre avaient été contrôlées : le SN et le SP étaient inanimés, de même longueur (en nombre de mots), le SN était toujours nouveau et indéfini, et les verbes utilisés étaient contrôlés. Seule la nouveauté du SP variait selon les phrases et leur contexte. L'hypothèse testée était la suivante : toutes choses égales par ailleurs, dans les cas où le SN est nouveau et que le SP est donné, la préférence pour l'ordre SP-SN est plus forte que dans les cas où les référents du SN et du SP sont tous deux nouveaux. Les sujets devaient juger l'acceptabilité des deux ordres possibles sur une échelle allant de 1 (= pas acceptable) à 10 (= complètement acceptable). Voici un exemple de phrase à juger, extraite du questionnaire :

> De nombreuses questions se posent à propos de la situation économique du pays. Il faut que les candidats maintenant
> donnent à ces questions des réponses appropriées. 1 ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ 10
> donnent des réponses appropriées à ces questions. 1 ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ 10

L'ordre des items était randomisé et la moitié du questionnaire était constituée de phrases non pertinentes pour notre problème (distracteurs). Le questionnaire a été rempli par 24 sujets de langue maternelle française (étudiants en L2 de l'Université Paris 4). L'analyse des jugements recueillis indique que l'ordre (SN-SP ou SP-SN) a un effet significatif sur les jugements des sujets. Par contraste, les résultats indiquent que ni le caractère donné du SP, ni l'interaction entre le caractère donné du SP et l'ordre n'a un effet sur les jugements des sujets. Les graphiques présentés dans la figure 1 représentent les jugements des sujets en fonction de l'ordre des compléments, sur le graphique de gauche, et en fonction du statut du SP, sur le graphique de droite.

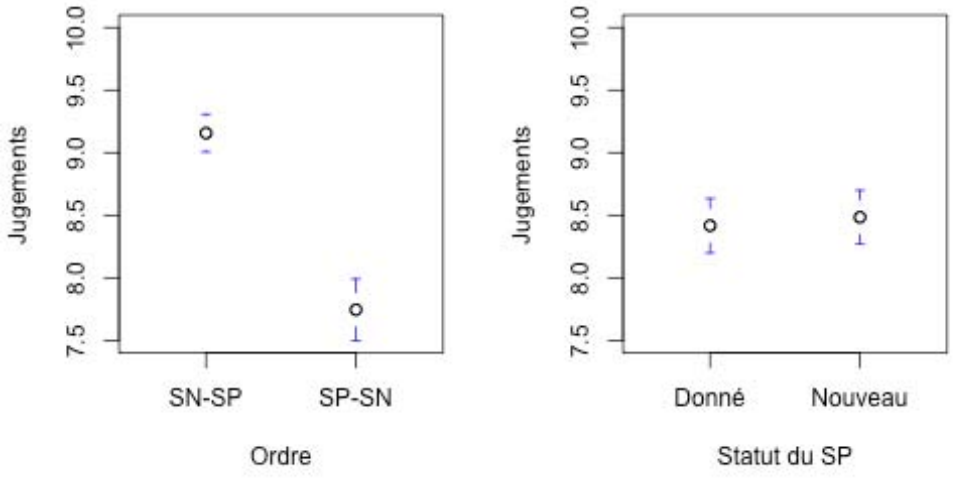

Figure 1 : Graphique de gauche : moyenne des jugements des sujets en fonction de l'ordre des compléments postverbaux ; graphique de droite : moyenne des jugements en fonction du caractère donné ou nouveau du référent du SP.

Sur les graphiques, chaque point indique la moyenne des jugements selon la condition indiquée en abscisse. Pour chaque moyenne, 95% des réponses recueillies se situent entre les extrémités de la barre verticale bleue. On observe que l'ordre SN-SP est jugé significativement plus acceptable que l'ordre SP-SN, tandis que les moyennes des jugements ne sont pas significativement différentes lorsque l'on fait varier le caractère donné ou nouveau du référent du SP. Les jugements recueillis grâce à ce questionnaire ne donnent pas de preuves que l'opposition entre donné et nouveau n'a pas d'influence sur l'ordre, mais, selon nous, elle permet de minimiser son importance. Les jugements des locuteurs n'étant absolument pas affectés par le caractère donné du SP, il paraît envisageable de considérer que l'effet de cette dimension n'est pas central dans les préférences concernant l'ordre des compléments postverbaux.

En nous appuyant sur l'observation de l'absence d'effet du caractère défini dans nos données ainsi que sur les résultats de notre questionnaire, nous estimons que les facteurs relatifs à la structure informationnelle ne sont pas centraux, à la différence de la longueur ou du caractère figé de la séquence V SP. Nous faisons l'hypothèse qu'il est possible d'observer le rôle du lemme verbal et de sa classe sémantique, sans prendre en compte des informations relatives à la structure informationnelle.

2.2.5 La longueur du SN et du SP

Dans les phénomènes touchant à l'ordre des mots, la longueur et la complexité des constituants sont considérées comme des facteurs pertinents. Dans les langues SVO, le principe général est que le constituant le plus court et le moins complexe apparaît avant le constituant le plus long et le plus complexe. Wasow (1997, 2002) a montré, à partir d'études sur des corpus de l'anglais, que les mesures en termes de nombre de mots et celles en termes de nombre de nœuds syntaxiques ou syntagmatiques dans l'arbre, sont quasi-équivalentes pour expliquer l'ordre des constituants. Étant donné que le nombre de mots est très peu coûteux à obtenir et que le nombre de nœuds dépend de l'analyse choisie pour l'annotation en constituants, nous avons opté pour une mesure unique : le nombre de mots. En utilisant cette mesure, on observe que 75,6% des données se conforment au principe *court avant long*, soit pour 69,5% des phrases présentant l'ordre SN-SP, le SP est plus long que le SN, et pour 82,3% des phrases ayant l'ordre SP-SN, le SN est plus long que le SP.

Pour capturer ce phénomène, nous utilisons une variable LOG(SN)-LOG(SP) qui correspond à la différence de longueur du SN et du SP. La longueur en mots du SN et la longueur en mots du SP étant distribuées selon une loi d'allure exponentielle, nous appliquons une transformation logarithmique à ces données. Nous avons donc : LOG(SN)-LOG(SP) = log(longueur en mots du SN) – log(longueur en mots du SP). Lorsque cette variable est positive, le SN est plus long que le SP, et inversement, lorsqu'elle est négative, c'est le SP qui est plus long que le SN.

2.2.6 Expression verbale figée

Dans certaines phrases, nous observons que le SP et le verbe forment une séquence figée non connexe. Il s'agit d'exemples tels que *mettre en valeur*, *mettre à disposition*, *prendre en charge*… Nous avons conservé ces séquences car le SP peut apparaître avant ou après le SN, comme le montrent les exemples (3) et (4) tirés de nos données.

(3) Le groupe Mada est l'une des formations musicales modernes qui essaie de **mettre en relief** cet instrument. (ER)

(4) Quoi de mieux qu'un lustre pour **mettre** l'orgueil **en lumière**. (ER)

Le lien sémantique particulier qui unit le verbe et le SP, semble favoriser l'ordre SP-SN. Nous observons par exemple que parmi les 269 SP qui ont une longueur égale à deux mots, ceux qui sont figés se présentent à 84.1% dans l'ordre SP-SN, alors que les autres le sont à 56% seulement. L'ensemble des SP figés est repéré dans nos données à l'aide de la variable SPFIG, qui est égale à 1 pour les SP figés et à 0 pour les autres SP. Les SP figés représentent 8.6% de nos données, soit 82 phrases.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100172

2.2.7 Le corpus

L'ordre des compléments du verbe est différent selon les 3 corpus. On observe 67.8% d'ordre SN-SP dans le FTB, 64% dans ESTER et 51.3% dans ER. Pour le problème qui nous intéresse, il semble que le corpus d'où est issue la phrase, est une source de variation. La répartition de ces données s'explique en partie par la proportion de SP figés que contient chaque corpus : 13.8% dans l'ER, 9.2% dans ESTER et 2.4% dans le FTB. Cependant, la proportion de SP figés n'explique pas l'ensemble de la variation entre corpus. On peut supposer que cette variation est due à une différence de registre : ER présente un style plus informel que FTB et ESTER. Pour confirmer une telle hypothèse, il faudrait observer d'autres corpus, notamment de l'oral en situation informelle. Nous prenons donc en compte le corpus d'origine à l'aide de la variable CORPUS, qui a trois valeurs possibles : FTB, ESTER ou ER.

2.3 Modélisation

Ce travail ne vise pas simplement à observer et à décrire le comportement de la variable ORDRE dans nos données, il vise également à tirer des généralités sur le phénomène pour le genre de discours traité (journalistique et radiophonique), voire pour la langue elle-même. Pour atteindre ce but, nous adoptons des méthodes de modélisation qui sont largement utilisées dans les sciences humaines (Howell, 1998) et qui permettent d'inférer les propriétés d'une population (le discours journalistique ou la langue) à partir d'un échantillon (nos données). Pour le problème d'ordre qui nous intéresse, nous utilisons la régression logistique à effets mixtes (Agresti 2007, Gelman & Hill 2006). Cette méthode, utilisée notamment par Bresnan et al. (2007) pour modéliser l'alternance dative en anglais, permet de prédire le comportement d'une variable binaire à partir de plusieurs variables prédictrices. L'ordre relatif du SN et du SP peut se représenter sous la forme d'une variable binaire : ordre SN-SP = 0 et ordre SP-SN = 1. Le résultat de la régression logistique s'interprète, dans notre cas, comme la probabilité que l'ordre soit SP-SN, étant donné toutes les variables prédictrices, autrement dit P(Y=SP-SN|X), où X représente l'ensemble des variables prédictrices. Ainsi, pour une phrase donnée, si P(Y=SP-SN|X) < 0.5, le modèle prédit l'ordre SN-SP, et si P(SP-SN|X) > 0.5, le modèle prédit l'ordre SP-SN. Le calcul de cette probabilité repose sur la formule mathématique suivante :

$$P(Y = SPSN|X) = \frac{e^{\beta X}}{1 + e^{\beta X}}$$

où X représente l'ensemble des variables prédictrices et β les coefficients associés à chaque variable prédictrice.

Comme nous l'avons expliqué dans la partie 2.2, les contraintes de pronominalité, caractère défini et caractère animé ne sont pas significatives pour la description de l'ordre des compléments postverbaux en français[vi]. Nous construisons donc le modèle à partir de la combinaison des trois variables pertinentes décrites précédemment : LOG(SN)-LOG(SP), SPFIG et CORPUS. Les deux premières variables sont des variables à effets fixes, c'est-à-dire des variables explicatives conditionnant la probabilité P(Y=SP-SN|X). D'après ce que nous avons vu dans la partie précédente, la nature du corpus est une source de variation dans les données. On pourrait imaginer construire un modèle pour chaque corpus. De cette façon, on aurait trois modèles différents qui reflèteraient le comportement de la variable ORDRE dans chaque corpus. En procédant de la sorte, on perdrait une partie de la généralisation, car les modèles proposés ne seraient valables que pour chaque corpus. Nous souhaitons obtenir une modélisation permettant la généralisation sur l'ensemble de nos données tout en tenant compte du corpus. Pour cela, nous traitons la variable CORPUS comme un effet aléatoire, c'est-à-dire comme une variable qui forme des groupes dans les données. Chaque valeur de la variable CORPUS (ER, FTB, ESTER) se voit alors attribuer un coefficient propre qui capture son comportement particulier. Avec cette approche, on fait l'hypothèse que l'effet d'un corpus particulier provient d'une distribution normale de moyenne 0 et d'écart-type σ, N(0, σ). Dans la formule, on note C_i la variable aléatoire CORPUS, où i peut avoir la valeur ER, FTB ou ESTER. Le modèle de base est le suivant [vii] :

Modèle 1

$$P(Y = SPSN|X)_i = \frac{e^{\beta X + C_i}}{1 + e^{\beta X + C_i}}$$

où $\boldsymbol{\beta X}$ =

et où $\boldsymbol{C_i} \sim N(0, 0.35)$

-0,98

+ 2,53 LOG(SN)-LOG(SP)

+1,08 SPFIG

L'interprétation des coefficients associés aux effets fixes n'est pas la même pour les deux variables car la variable SPFIG n'a que des valeurs positives (0 ou 1), alors que la variable LOG(SN)-LOG(SP) peut être positive ou négative. Le coefficient positif associé à la variable SPFIG indique que le variable vote pour l'ordre SP-SN. Le vote de la variable LOG(SN)-LOG(SP) dépend de son propre signe : si LOG(SN)-LOG(SP) est positif, le vote se fait pour l'ordre SP-SN, tandis que si LOG(SN)-LOG(SP) est négatif, le vote est pour l'ordre SN-SP. Cela correspond à ce que l'on attendait : lorsque le SN est plus long que le SP, on prédit l'ordre SP-SN et lorsque c'est le SP qui est plus long que le SN, on s'attend à avoir l'ordre SN-SP.

3 Lemme verbal

Le lemme verbal sous-catégorisant les deux compléments que nous étudions a une influence sur la variable ORDRE. C'est ce que nous allons montrer à partir de nos données, d'abord par des observations, puis de façon quantitative, en intégrant la variable LEMMEVERBAL au modèle de régression logistique.

3.1 Observations

La table de données contient 151[viii] lemmes verbaux différents, dont 114 ont une fréquence inférieure ou égale à 3 dans nos données. Nous avons sélectionné quelques verbes ayant une fréquence plus importante pour observer le comportement de la variable ORDRE. Le tableau 2 présente les pourcentages d'ordre SN-SP et SP-SN en fonction de 6 verbes avec leur nombre d'occurrences entre parenthèses. On observe que les verbes *ajouter* et *montrer* montrent une préférence pour l'ordre SP-SN, alors que *trouver* et *donner* favorisent l'ordre SN-SP. Enfin, les verbes *mettre* et *faire* présentent une tendance moins marquée avec des proportions autour de 47% pour l'ordre SN-SP.

	ajouter (25 occ.)	*montrer* (74)	*mettre* (124)	*faire* (62)	*trouver* (29)	*donner* (91)
SN-SP	28%	37,8%	46,8%	46,8%	69%	78%
SP-SN	72%	62,2%	53,2%	53,2%	31%	22%

Tableau 2: La variable ORDRE selon six lemmes verbaux.

Ces données semblent donc indiquer que le verbe influe sur l'ordre relatif des compléments postverbaux.

3.2 Modélisation

Nous introduisons la variable LEMMEVERBAL dans notre modèle de base. De la même façon que la variable CORPUS, la variable LEMMEVERBAL est traitée comme un effet aléatoire. Ainsi, nous estimons que les données sont groupées autour de chaque lemme verbal. Mettre les lemmes verbaux en effets aléatoires revient en quelque sorte à lexicaliser le modèle : en plus de l'intercept général et des effets fixes, chaque lemme se voit attribuer un coefficient propre qui rend compte de son comportement

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100172

particulier. On note L_j la variable aléatoire LEMMEVERBAL (j prend pour valeur le lemme verbal concerné).

Le modèle construit sur notre table de données est le suivant :

Modèle 2

$$P(Y = SPSN|X)_{ij} = \frac{e^{\beta X + C_i + L_j}}{1 + e^{\beta X + C_i + L_j}}$$

où $\boldsymbol{\beta X}$ =

-1.20
+ 2.77 LOG(SN)-LOG(SP)
+1,37 SPFIG

et où $\boldsymbol{C_i} \sim N(0, 0.32)$

$\boldsymbol{L_j} \sim N(0, 1.02)$

Les effets fixes gardent des coefficients orientés de la même façon que dans le modèle 1. Cela signifie que les variables continuent de voter pour le même ordre lorsque la variable LEMMEVERBAL est prise en compte. Chaque lemme verbal se voit associer un intercept qui indique son biais : un intercept positif indique une préférence pour l'ordre SP-SN et un intercept négatif pour l'ordre SN-SP. Une valeur très proche de 0, comme pour le verbe *ramener* (tableau 3), indique que le lemme n'a qu'une très faible préférence. Par manque de place, nous ne reproduisons, dans le tableau 3, qu'un échantillon des lemmes verbaux accompagnés de leur intercept.

ramener	-0.0192	*demander*	+0.0896	*vendre*	+1.0270
redonner	-0.6668	*annoncer*	+0.2302	*trouver*	-0.3000
céder	-0.4433	*assurer*	-0.2952	*porter*	-0.9465
laisser	-1.1649	*expliquer*	+1.1656	*prendre*	-0.0020
présenter	+0.5512	*accorder*	-0.1664	*passer*	-1.8508
apporter	-0.3110	*devoir*	+0.0724	*faire*	+1.6674
dire	+0.4624	*réduire*	+0.6689	*montrer*	+0.8743
obtenir	+0.6023	*ajouter*	+1.3717	*donner*	-0.2350
offrir	-0.1049	*rendre*	+0.4849	*mettre*	-0.1074

Tableau 3 : Intercepts aléatoires pour la variable LEMMEVERBAL dans le modèle 2.

Il existe une méthode qui permet de comparer deux modèles et notamment de vérifier si le modèle le plus complexe apporte quelque chose à la modélisation de la variable qui nous intéresse. Dans notre cas, nous cherchons à vérifier si le modèle 2, qui comporte une variable de plus que le modèle 1, apporte quelque chose de significatif à la modélisation de la variable ORDRE. Pour cela, on utilise un test de rapport de vraisemblance[ix]. Ainsi, nous observons que la prise en compte du lemme verbal comme effet aléatoire améliore significativement la vraisemblance des données : chi2(1) = 27.342, P(> chi2) = 1.705e-07. Cela signifie que la probabilité d'occurrence des données est plus élevée selon l'hypothèse représentée par le modèle 2 que selon celle représentée par le modèle 1.

Les observations sur six lemmes, effectuées dans la partie précédente, sont donc confirmées par la modélisation : le lemme verbal influe sur le choix de l'ordre des compléments, et ce, indépendamment des variables LOG(SN)-LOG(SP), SPFIG et CORPUS. Cependant, il faut noter que la préférence en termes de pourcentage dans les données ne correspond pas toujours à la préférence évaluée grâce à l'intercept

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100172

aléatoire dans le modèle 2. Par exemple, d'un côté, le verbe *donner* se rencontre à 78% des données avec l'ordre SN-SP, ce qui semble indiquer une forte préférence pour l'ordre SN-SP. De l'autre côté, l'intercept aléatoire de ce verbe dans le modèle 2 est d'environ -0.24, ce qui signifie que le biais estimé de *donner* est bien en faveur de l'ordre SN-SP, mais de façon relativement faible. L'intercept aléatoire est une valeur plus fiable de la préférence de chaque lemme verbal car le modèle dans lequel cette valeur est estimée prend en compte les autres variables influençant l'ordre des compléments. Cependant, cette valeur est dépendante du modèle dans lequel elle est calculée : ainsi nous pouvons comparer les intercepts aléatoires des verbes à l'intérieur du modèle 2, mais nous ne pouvons pas les comparer avec des intercepts aléatoires qui auraient été calculées dans un autre modèle.

3.3 Bilan

Le modèle 2 a permis de mettre en évidence le rôle de la variable LEMMEVERBAL. De plus, cette modélisation a permis d'estimer le biais que chaque verbe impose sur l'ordre des compléments, grâce à la valeur de l'intercept aléatoire associé à chaque verbe. L'utilisation de cette valeur comme mesure de préférence du verbe est particulièrement pertinente dans la mesure où elle n'est pas biaisée par les facteurs généraux qui sont capturés par les variables à effet fixe.

4 La sémantique du verbe

Dans la lignée des travaux de Bresnan et al. (2007), nous examinons le rôle de la classe sémantique du verbe sur l'ordre des compléments. Pour l'alternance dative, les travaux de Gries (2003) et Bresnan et al. (2007) ont montré que le sens du verbe a une influence sur le choix entre la construction à double objet et la construction à SP-datif. Pour capturer le sens du verbe, Bresnan et al. (2007) ont combiné les lemmes verbaux avec 6 classes sémantiques de façon à créer des sous-classes pour les différents emplois de chaque verbe. Nous avons procédé de la même façon en utilisant le classement sémantique du dictionnaire *Les verbes du français* (LVF, Dubois & Dubois-Charlier, 1997). Dans le LVF, les verbes sont classés selon 5 niveaux qui mêlent des critères syntaxiques et sémantiques : classes génériques, classes sémantico-syntaxiques, sous-classes syntaxiques, sous-types syntaxiques et constructions. Dans le cadre de cet article, nous nous intéressons uniquement au niveau de classification le plus général, c'est-à-dire les classes génériques. Il existe 14 classes génériques qui sont détaillées dans le tableau 4.

C : communication	L : locatif	S : saisir, serrer, posséder
D : don, privation	M : mouvement sur place	T : transformation, changement
E : entrée, sortie	N : munir, démunir	U : union, réunion
F : frapper, toucher	P : verbes psychologiques	X : verbes auxiliaires
H : états physiques et comportement	R : réalisation, mise en état	

Tableau 4 : Les classes génériques du dictionnaire *Les Verbes du Français*.

4.1 Annotation en classes sémantiques

Pour travailler sur la sémantique des verbes, nous avons procédé à une annotation des données selon les trois premiers niveaux de classification du LVF : classes génériques, classes sémantico-syntaxiques et sous-classes syntaxiques. La tâche d'annotation[x] consistait à donner les 3 niveaux de classification pour chaque occurrence de verbe dans son contexte d'origine.

Trois annotateurs ont réalisé cette tâche. Pour mesurer l'accord inter-annotateur, nous utilisons le Multi-pi de Fleiss (Artsein & Poesio, 2008), aussi connu sous le nom de Kappa de Carletta (Carletta, 1996). En ce qui concerne les classes génériques, Multi-pi = 0.84 (k = 3, N=1346). En général, on estime qu'un taux

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100172

supérieur à 0.8 indique un bon accord inter-annotateur. Nous pouvons donc considérer que les données annotées concernant les classes génériques sont fiables.

4.2 Observations

Dans le tableau 5, pour chaque classe générique, nous présentons le nombre d'occurrences appartenant à cette classe ainsi que le pourcentage d'ordre SN-SP.

	occurrences	ordre SN-SP		occurrences	ordre SN-SP
C	174	46.6%	N	6	50%
D	283	70.3%	P	15	100%
E	100	80%	R	99	52.5%
F	4	75%	S	21	71.4%
H	34	44.1%	T	46	28.3%
L	107	64.5%	U	44	52.3%
M	23	39.1%			

Tableau 5 : Nombre d'occurrences et pourcentage d'ordre SN-SP dans les données selon les classes génériques.

Ces données générales montrent que les classes *D*, *E*, *L*, *S* et *P* semblent avoir une préférence pour l'ordre SN-SP, alors que les classes *M* et *T* semblent plutôt favoriser l'ordre SP-SN. On pourrait supposer qu'un sens prédispose en général à un ordre particulier. Cependant, si l'on examine les verbes qui composent chaque classe, on observe qu'il y a de la variation selon les lemmes. Par exemple, dans la classe *D*, le verbe *donner* se présente à 76.2% des cas avec l'ordre SN-SP, tandis que le verbe *vendre* présente seulement 40% d'ordre SN-SP. Le verbe *devoir*, quant à lui, se combine à 94.4% avec l'ordre SN-SP. De même, la classe *T* est composée de 6 lemmes verbaux : *compléter*, *échanger*, *ériger*, *faire*, *remplacer*, *transformer*. Les occurrences du verbe *faire* représentent 76.1% des données de cette classe. Or, ce verbe n'apparaît que dans 8.6% des cas dans l'ordre SN-SP, alors que les autres verbes ne présentent que peu d'occurrences qui se combinent très majoritairement avec l'ordre SN-SP. C'est donc la surreprésentation du lemme *faire* dans cette classe qui aboutit à une préférence pour l'ordre SP-SN de la classe *T*. Ainsi, étant donné la variation dans chaque classe, il semble que le niveau pertinent pour étudier l'ordre des compléments postverbaux reste le lemme verbal. Néanmoins, les classes sémantiques peuvent être utilisées pour désambiguïser les emplois d'un même lemme.

Afin de capturer les emplois de chaque verbe, nous avons créé une nouvelle variable, LEMMESEM. Cette variable est constituée de la concaténation du lemme verbal et de la classe sémantique. La table de données contient 151 lemmes et la variable LEMMESEM a 185 valeurs. Cela signifie que seule une minorité de verbes présente différents emplois dans nos données. Ces verbes sont, en général, les plus fréquents. Nous présentons, dans le tableau 6, le cas de trois verbes pour lesquels les différents emplois semblent correspondre à des différences de préférence pour l'ordre des compléments. D'après ce tableau, l'emploi du verbe *mettre* dans le sens locatif (*mettre L*, exemple 5) est plus fréquemment observé avec l'ordre SN-SP que l'emploi avec le sens de don (*mettre D*, exemple 6). Enfin, l'emploi, avec le sens de réalisation (*mettre R*, exemple 7), présente une forte préférence pour l'ordre SP-SN. Les classes génériques permettent de différencier deux emplois du verbe *faire* : *faire R* (exemple 8) et *faire T* (exemple 9). Ces différences d'emploi sont en correspondance avec des préférences opposées pour l'ordre des compléments. Enfin, le verbe *réduire* présente deux emplois dans les données (*réduire E*, exemple 10 et *réduire M*, exemple 11). L'emploi capturé par la classe générique *M* préfère à plus de 90%

l'ordre SN-SP, alors que les emplois appartenant à la classe E s'observent à 41.7% seulement avec l'ordre SN-SP.

LemmeSem	ordre SN-SP	LemmeSem	ordre SN-SP	LemmeSem	ordre SN-SP
mettre D	45.5%	*faire R*	96.3%	*réduire E*	90.9%
mettre L	69.4%	*faire T*	8.6%	*réduire M*	41.7%
mettre R	29.7%				

Tableau 6 : Proportions d'ordre SN-SP pour 7 valeurs de la variable LEMMESEM.

(5) *mettre L* : Alors c'est sûr que là, il y a déjà une première empreinte qui est très négative parce que on on va **mettre** l'en l' enfant dans une position initiale de méfiance par rapport à les autres. (ESTER)

(6) *mettre D* : Ehud Barak veut sortir la négociation de l'impasse et **mettre** son action au service de l'espoir. (ER)

(7) *mettre R* : C'est l'AVDAM […] qui **met** en oeuvre ce dispositif et organise 170 prestations annuelles dans les établissements du département. (ER)

(8) *faire R* : Avec son excellent français, il se charge de **faire** la traduction à ses coéquipiers. (ER)

(9) *faire T* : Finalement accepté moyennant des aménagements, il **fait** du groupe français le numéro un mondial en équipements de transmissions. (FTB)

(10) *réduire E* : L'auteur s'ébahit d'un "monde qui **réduit** l'espace international à la liste d'un annuaire téléphonique". (FTB)

(11) *réduire M* : La semaine dernière, le président de l'OPEP à Vienne estimait que l'épidémie de pneumopathie **réduirait** la demande mondiale de 300 mille barils de pétrole par jour en Asie. (ESTER)

Il est intéressant de noter que l'association du lemme verbal avec la préposition introduisant le SP donne, pour une part importante des verbes, les mêmes résultats que la variable LEMMESEM. En d'autres termes, le sens dans lequel est employé un lemme verbal ditransitif, est très souvent capturé par la préposition utilisée. Cependant, on observe que le découpage en classes sémantiques permet une meilleure désambiguïsation. Prenons l'exemple des verbes *mettre* et *prendre*. Nous représentons la préposition à l'aide de trois valeurs « A », « De » et « Autres ».

	occurrences	ordre SN-SP		occurrences	ordre SN-SP
prendre A	7	100%	*prendre S*	7	85.7%
prendre Autres	30	30%	*prendre L*	8	37.5%
			prendre H	22	31.8%
mettre A	8	62%	*mettre L*	49	69.4%
mettre Autres	115	46%	*mettre D*	11	45.5%
			mettre R	64	29.7%

Tableau 7 : Comparaison de l'association du lemme avec la préposition avec la variable LEMMESEM.

Dans le tableau 7, le verbe *prendre* présente deux emplois selon la préposition et trois emplois selon les classes sémantiques. Les classes nous permettent de voir que lorsque le verbe est employé avec le sens de

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100172

« saisir, posséder » (*prendre S*, exemple 12), l'ordre SN-SP est largement favorisé, alors qu'avec le sens de « comportement » (*prendre H*, exemple 14) ou de locatif au sens abstrait (*prendre L*, exemple 13), c'est l'ordre SP-SN qui est préféré. De même, le verbe *mettre* ne présente que deux emplois selon les prépositions, alors que les classes sémantiques nous permettent d'observer trois emplois : « locatif » (*mettre L*, exemple 5), « don » (*mettre D*, exemple 6) et « réalisation » (*mettre R*, exemple 7).

(12) *prendre S* : Les habitants de Waldstetten "qui depuis le temps sont devenus des amis", **prendront** une part active à la fête des pains. (ER)

(13) *prendre L*^{xi} : Ainsi Sarkozy a -t-il **pris** François Hollande pour cible spécifique (ER)

(14) *prendre H* : J'ai jamais **pris** le rôle de star au sérieux (ESTER)

Ainsi, même si la préposition semble être une bonne approximation pour désambiguïser les emplois d'un lemme verbal, les classes sémantiques apparaissent plus pertinentes et plus précises. De la même façon que pour les lemmes verbaux, les observations faites dans le tableau 6 semblent indiquer que les classes sémantiques, associées aux lemmes verbaux, ont une influence sur l'ordre des compléments.

4.3 Modélisation

Afin de vérifier la validité de ces observations, nous présentons le modèle 3 qui comporte la variable LEMMESEM comme effet aléatoire (notée S_k, avec k la valeur concernée de la variable LEMMESEM). Le modèle construit sur la table de données est le suivant :

Modèle 3

$$P(Y = SPSN|X)_{ik} = \frac{e^{\beta X + C_i + S_k}}{1 + e^{\beta X + C_i + S_k}}$$

où $\boldsymbol{\beta X}$ =

-1.29
+ 2.80 LOG(SN)-LOG(SP)
+1,45 SPFIG

et où $\boldsymbol{C_i} \sim N(0, 0.33)$

$\boldsymbol{S_k} \sim N(0, 1.22)$

Les effets fixes du modèle 3 sont quasi équivalents à ceux du modèle 2, ce qui signifie que le vote de ces variables va dans le même sens. Ce modèle contient un intercept pour chaque lemme verbal associé à une classe générique. Dans le tableau 8, nous donnons les intercepts aléatoires pour les verbes que nous avons étudiés dans la partie 4.2.

mettre D	+0.8775	*faire R*	-0.7138	*réduire E*	-0,2895
mettre L	-0.6564	*faire T*	+2,9521	*réduire M*	1.0095
mettre R	+0.1354				

Tableau 8 : Intercepts aléatoires pour la variable LEMMESEM dans le modèle 3.

Comme nous l'avons expliqué pour les intercepts aléatoires des lemmes verbaux, ces intercepts donnent une idée plus juste des préférences des emplois des lemmes verbaux, car l'effet des variables LOG(SN)-LOG(SP) et SPFIG est contrôlé. Par exemple, l'intercept de *réduire M* est d'environ +1.01, ce qui indique une préférence marquée pour l'ordre SP-SN, alors qu'en termes de pourcentage (cf. tableau 6) la préférence était légère (58.3% d'ordre SP-SN).

Pour comparer le modèle 2 (avec LEMMEVERBAL en effet aléatoire) et le modèle 3 (avec LEMMESEM en effet aléatoire), nous ne pouvons pas utiliser le test de rapport de vraisemblance car les deux modèles ne sont pas imbriqués l'un dans l'autre (ils se différencient par deux variables). Nous nous appuyons alors

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100172

sur la précision et la qualité de chaque modèle pour tenter de déterminer lequel de ces deux modèles représente la meilleure modélisation de la variable ORDRE. Nous mesurons l'exactitude des modèles avec une évaluation 100 passes. Nous divisons 100 fois notre table de données en un ensemble d'entraînement de 946 phrases et en un ensemble test de 10 phrases. Nous construisons un modèle à partir de chaque ensemble d'entraînement puis nous calculons les probabilités prédites par ce modèle sur l'ensemble test. L'exactitude est estimée à partir de la probabilité de concordance, C[xii]. Comme Bresnan & Ford (2010) l'expliquent, « la probabilité de concordance est utilisée pour mesurer à quel point un modèle est capable de discriminer des paires de réponses opposées ». Pour interpréter la valeur C, on considère qu'une valeur de C = 0.5 indique des prédictions aléatoires et une valeur de C = 1 indique des prédictions parfaites. Enfin, une valeur supérieure à environ C = 0.8 a quelque utilité dans la prédiction des réponses (Harrell, 2001 : 247). La moyenne de probabilité de concordance sur les 100 ensembles test est C = 0.933 pour le modèle 2 et C = 0.941 pour le modèle 3. Cette mesure indique que l'exactitude des modèles 2 et 3 est très bonne. On observe également que l'exactitude du modèle 3 est légèrement meilleure que celle du modèle 2.

Nous comparons également les modèles en utilisant une représentation graphique de l'adéquation des données avec les prédictions. Pour cela, nous groupons les probabilités prédites selon 10 sous-intervalles égaux sur l'intervalle [0,1]. La probabilité moyenne dans chaque sous-intervalle est comparée aux proportions observées de réponse 1 (c'est-à-dire de réponse ordre SP-SN) dans le même sous-intervalle. Les graphes de la figure 2 représentent les probabilités moyennes, prédites par les modèles 2 et 3, en fonction des proportions observées dans la table de données. Les deux graphes montrent une très bonne qualité de l'ajustement. Le modèle 2 présente un ajustement moins bon pour les probabilités groupées autour de 0.4, 0.5 et 0.6. Or ces probabilités sont critiques dans la mesure où 0.5 est le seuil distinguant les deux résultats possibles pour la variable ORDRE : SN-SP ou SP-SN. Le modèle 3 présente un très bon ajustement pour l'ensemble des probabilités, même celles autour de 0.5. Ainsi, la valeur de la probabilité de concordance C ainsi que le graphe de l'ajustement des données, semblent indiquer que le modèle 3 constitue une meilleure modélisation de la variable ORDRE que le modèle 2.

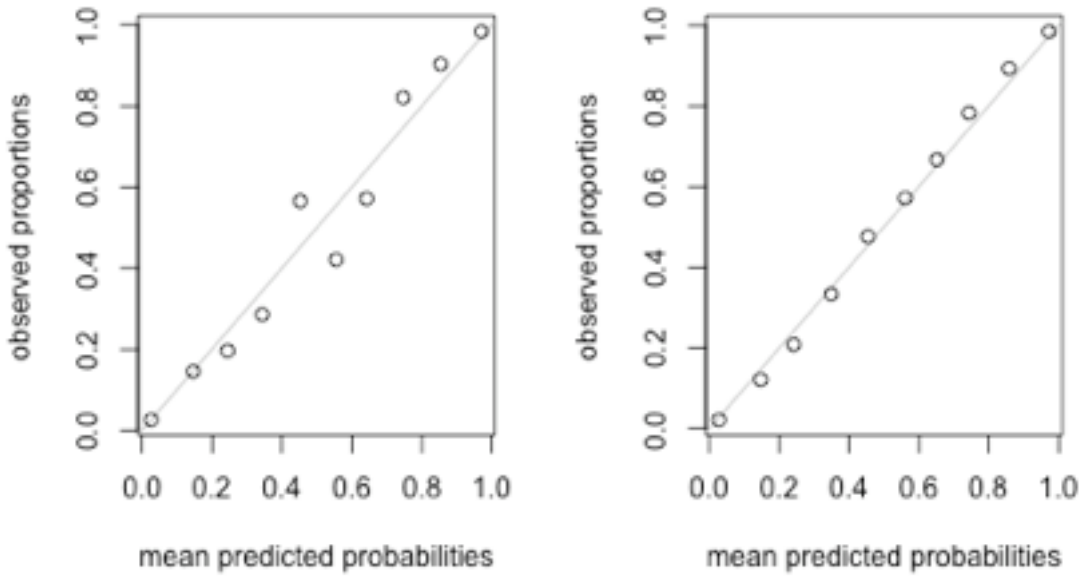

Figure 2 : Ajustement des observations groupées et des probabilités prédites moyennes pour le modèle 2 (à gauche) et le modèle 3 (à droite).

En montrant que le modèle 3 est un modèle avec une meilleure exactitude et un meilleur ajustement aux données, nous avons confirmé les observations selon lesquelles différents emplois d'un même lemme peuvent entraîner des préférences différentes dans l'ordonnancement des compléments postverbaux. On

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100172

peut supposer que le fait que le modèle 3 n'a une exactitude que légèrement supérieure à celle de modèle 2 est une conséquence du peu de lemmes qui présentent plusieurs emplois sur l'ensemble des lemmes de nos données. En effet, on peut penser que, pour la plupart des verbes, le lemme seul suffit puisqu'il n'y a qu'un seul emploi. Le modèle 3 améliore donc seulement la modélisation des données pour les verbes à plusieurs emplois.

4.4 Bilan

Nous avons montré, avec le modèle 3, que le type d'emploi du verbe joue un rôle dans le choix de l'ordre des compléments. Le modèle que nous avons construit présente un très bon ajustement aux données (cf. Figure 2). Il permet de connaître l'ordre des compléments pour des phrases inconnues avec une exactitude de C = 0.941[xiii].

5 Discussion

Grâce au travail sur les données des corpus FTB, ER et ESTER, nous avons montré que le verbe, associé à sa classe sémantique, permet de mieux modéliser le comportement de la variable ORDRE. Comme nous l'avons expliqué dans la partie 4.3, l'un des intérêts de la modélisation que nous proposons, est de pouvoir évaluer la préférence de chaque verbe pour un ordre, indépendamment de la longueur et du caractère figé de la séquence V SP. Dans cette partie, nous évoquons deux pistes d'analyse des intercepts associés aux différentes valeurs de l'effet aléatoire LEMMESEM : la première met en lien les préférences de certains verbes avec les rôles sémantiques des arguments ; la seconde repose sur l'idée que certains verbes favorisent un ordre sous l'influence de leur fréquence d'apparition dans des constructions où l'ordre des compléments est quasiment déterminé (Stallings et al. 1998).

Nous avons montré, avec le modèle 3, que le verbe désambiguïsé par sa classe verbale a une influence sur l'ordre des compléments. Cette conclusion va dans le sens des observations de Schmitt (1987a, b). Les verbes étudiés par cet auteur sont en grande partie regroupés dans la classe *T* (transformation, changement). Dans le tableau 9, nous reproduisons les coefficients attribués aux lemmes inclus dans cette classe. Les intercepts de la quasi-totalité des verbes de cette classe sont négatifs, ce qui indique une préférence pour l'ordre SN-SP. Seul le verbe *faire T* a un intercept positif marquant une forte préférence pour l'ordre SP-SN. Le comportement des verbes de la classe *T* n'est pas homogène du point de vue syntaxique, puisque ces verbes n'influent pas de la même façon sur l'ordonnancement des syntagmes compléments.

faire T (35)	+2.9521	*ériger T* (1)	-0.2435	*remplacer T* (4)	-0.4214
échanger T (2)	-0.3850	*compléter T* (1)	-0.0045	*transformer T* (3)	-0.1348

Tableau 9 : Intercepts aléatoires associés aux verbes de la classe *T* (entre parenthèses, figure le nombre d'occurrences dans les données).

Pour ces verbes, on peut avancer une autre hypothèse : l'ordonnancement des compléments est sensible aux rôles sémantiques que leur attribue le verbe, autrement dit à la structure argumentale du verbe. Hormis le verbe *compléter T*, les verbes *T* présents dans nos données peuvent être décrits comme mettant en jeu un thème et une cible et comme exprimant le passage du thème vers la cible, soit par la transformation (*faire T*, *ériger T*, *transformer T*), soit par la substitution (*échanger T*, *remplacer T*). Pour la majorité de ces verbes, l'appariement entre rôles sémantiques et fonctions syntaxiques est le suivant : le thème est apparié à la fonction OD et la cible à la fonction oblique. Seul le verbe *faire T* se distingue de ce point de vue : pour ce dernier, c'est la cible qui est appariée à l'OD tandis que le thème est apparié au complément oblique. Si on admet que l'ordre des compléments reflète l'ordre <[thème] [cible]>, on peut alors unifier le comportement des verbes individuels, indépendamment de leur cadre de sous-catégorisation. Ce sont alors les rôles sémantiques imposés par ces verbes qui expliquent les préférences

observées. Une telle hypothèse implique que l'ordre des mots peut être sensible aux rôles sémantiques assignés par le verbe, indépendamment de l'appariement entre rôles sémantiques et fonctions syntaxiques. Cette piste doit être approfondie en précisant le type de prédicat concerné, en élargissant les observations à un plus grand nombre de verbes et en augmentant le nombre d'observations pour chacun des verbes.

Pour d'autres verbes, notamment ceux de la classe *C* (communication), on observe une correspondance entre les préférences des verbes pour l'ordre SP-SN et la réalisation potentielle de l'objet direct sous la forme d'une complétive (COMPL) ou d'une infinitive (INF). Nous reprenons ici les travaux de Stallings et al. (1998) qui traitent de la question du *Heavy NP Shift* (HNPS) en anglais. Sous le nom de HNPS (Kimball, 1973), on désigne les cas où un SN objet direct apparaît à la fin du SV et est séparé du verbe par du matériel linguistique, par exemple un SP, comme en (15).

(15) […] Snowball had **found** [in the harness-room]SP [an old green tablecloth of Mrs Jones's]SN (G. Orwell, *Animal Farm*)

En s'appuyant sur un travail expérimental en psycholinguistique, Stallings et al. montrent que, au-delà de la longueur et de la complexité du SN, la fréquence d'apparition d'un verbe dans des constructions où l'objet direct n'est pas adjacent au verbe, influence significativement la production de HNPS pour ce verbe. L'idée est que, dans le savoir langagier des locuteurs, un verbe est associé aux fréquences relatives des différentes constructions dans lesquelles il est instancié et que ces fréquences ont une influence sur l'ordre attesté à chaque nouvelle occurrence du verbe. Par exemple, le HNPS tend à être favorisé par un verbe qui apparaît fréquemment dans une construction où l'objet direct est réalisé comme une COMPL et apparaît donc facilement en fin de syntagme verbal. Stallings et al. posent une hypothèse qu'ils nomment *Verb disposition hypothesis* et qu'ils formulent de la façon suivante : « *individudal verbs carry with them informations on the history of their participation in shifted structures and […] this history influences the likelihood of their allowing HNPS* » (Stallings et al. 1998 : 396). Nous nous inspirons de cette hypothèse pour tenter d'interpréter la préférence de certains verbes de la classe *C* (communication). Le tableau 10 présente les intercepts associés aux verbes *C* qui ont une fréquence supérieure à 5 dans nos données.

expliquer C(15)	+1.4807	*présenter C*(7)	+0.6202	*demander C*(12)	+0.0882
proposer C(6)	+0.7906	*annoncer C*(14)	+0.2200		
montrer C(74)	+0.7595	*dire C*(8)	+0.1688		

Tableau 10 : Intercepts aléatoires associés aux verbes de la classe *C* ayant une fréquence supérieure à 5 (entre parenthèses, figure le nombre d'occurrences dans les données).

Ces verbes ont tous un intercept aléatoire positif qui indique une préférence pour l'ordre SP-SN. De plus, l'ensemble de ces verbes, excepté *présenter C*, autorise la réalisation de leur objet direct sous la forme d'une COMPL (ex. 16, 18, 19) ou d'une INF (ex. 17, 20)[xiv].

(16) […] **expliquer** à ces mêmes PME, qui gémissent sous le poids de leurs frais financiers, [que plus les affaires vont mal, plus il faut relever les taux d'intérêt] (FTB)

(17) […] **proposait** à Pierre Grimblat, le PDG de la société de production Hamster, [de tourner un "Kojak" à la française] (FTB)

(18) […] **montrer** à mes deux grands-pères (qui sont accordéonistes eux aussi) [que j'étais capable de jouer aussi bien qu'eux] (ER)

(19) […] **ont annoncé** à Michel Humbert, maire, présent aux côtés de Christiane Laval et de Jean-Claude Thomachot, ses adjoints, [qu' ils souhaitaient la relève] (ER)

(20) […] je **dirai** à mon entourage [de venir la prochaine fois] (ER)

Dans les corpus FTB, ER, et ESTER, nous avons extrait 303 phrases contenant un verbe suivi d'une COMPL ou d'une INF ainsi que d'un SP sous-catégorisé par le verbe[xv]. L'ordre attesté parmi ces 303

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100172

phrases est à 99,7% l'ordre SP-COMPL/INF. On peut donc faire l'hypothèse que, lorsque l'objet direct de ces verbes est réalisé sous la forme d'une COMPL ou d'une INF, ces verbes sont quasiment toujours séparés de leur objet direct par le SP complément. Ainsi, une partie des verbes de communication sont fréquemment séparés de leur objet direct par le SP complément dans les constructions à COMPL ou INF. La préférence de ces verbes pour l'ordre SP-SN, dans les cas où l'objet direct est réalisé sous la forme d'un SN, peut s'expliquer par la *Verb disposition hypothesis* adaptée au français : « chaque verbe porte avec lui des informations sur l'histoire de sa participation à des structures où l'objet direct est séparé du verbe par un SP complément et cette histoire influence la probabilité pour ce verbe d'apparaître dans une construction où le SN objet direct est séparé du verbe par un SP complément (ordre SP-SN) ». En d'autres termes, nous faisons l'hypothèse que la préférence d'une partie des verbes de la classe *C* pour l'ordre SP-SN est partiellement due à la fréquence d'occurrences de ces mêmes verbes dans des constructions où, réalisé sous une forme COMPL ou INF, l'objet direct est très fréquemment séparé du verbe. Cette hypothèse doit être vérifiée sur un éventail plus large de verbes et en s'appuyant plus précisément sur la fréquence d'apparition de ces verbes dans des contextes où la complétive objet direct est séparée du verbe par un SP complément.

Les pistes d'interprétation quant au rôle du verbe dans l'ordre des compléments postverbaux suggèrent que les facteurs sont hétérogènes et interviennent à différents niveaux : au niveau de la relation sémantique unissant les arguments du verbe aussi bien qu'au niveau de la construction syntaxique dans laquelle le verbe peut être instancié.

6 Conclusion et perspectives

Dans cet article, nous avons étudié l'ordre des compléments postverbaux en français à partir de données extraites de trois corpus (FTB, ER, ESTER). Nous avons montré que, au-delà de la longueur relative des compléments (LOG(SN)-LOG(SP)) et du lien sémantique unissant le verbe et le SP (SPFIG), le lemme verbal ainsi que son emploi influencent l'ordre attesté. Nous avons confirmé les intuitions de Schmitt (1987a, b) et nous les avons étendues à l'ensemble des verbes sous-catégorisant deux compléments postverbaux. La valeur ajoutée du présent travail est de montrer que l'influence du lemme verbal est statistiquement significative et ce, lorsque les autres contraintes qui interviennent de façon massive dans l'ordonnancement des compléments, sont prises en compte (LOG(SN)-LOG(SP) et SPFIG). Nous proposons deux pistes pour expliquer l'effet du lemme verbal sur l'ordonnancement des compléments postverbaux. Premièrement, en ce qui concerne les verbes de la classe *Communication*, nous faisons l'hypothèse d'un effet de la fréquence d'apparition du verbe dans des contextes où l'objet est séparé du verbe par un SP complément. Deuxièmement, pour les verbes de la classe sémantique *Transformation*, nous suggérons une généralisation basée sur les rôles sémantiques plutôt que sur les catégories. Sur le plan méthodologique, ce travail met en œuvre des méthodes peu répandues en syntaxe et qui constituent un outil de plus à la disposition des linguistes pour décrire et modéliser les phénomènes langagiers. Nous espérons, à la suite des linguistes anglo-saxons, avoir montré l'utilité et l'apport de telles méthodes pour traiter de problèmes impliquant plusieurs contraintes préférentielles.

Un des buts de la régression logistique est de permettre la généralisation au-delà de l'échantillon que représente le corpus. Notre échantillon est composé de textes journalistiques (en considérant que ESTER est une forme de discours journalistique). On peut faire l'hypothèse que les propriétés dégagées sur nos données sont valables pour le discours journalistique. Ainsi, dans ce type de discours, la longueur et le caractère figé du SP sont des facteurs centraux pour le choix de l'ordre des compléments. Leur rôle peut être quantifié grâce aux coefficients qui ont été attribués à ces variables dans le modèle 3. De plus, l'influence du lemme verbal et de son emploi a également été mise à jour et formalisée grâce aux intercepts aléatoires du modèle 3.

Nous voyons deux perspectives principales à ce travail. Premièrement, les données utilisées étant issues de corpus journalistique et radiophonique, nous n'avons pas un échantillon représentatif de la langue. Il faut donc agrandir le corpus en intégrant d'autres genres, notamment de l'oral spontané. Les conclusions que nous obtiendrons pourront alors être envisagées comme portant sur la langue. Deuxièmement, ce type

d'étude sur corpus a pour objectif de décrire et de formaliser des phénomènes de langue. Cependant, en nous appuyant seulement sur des données de corpus, nous ne pouvons pas savoir si les éléments dégagés font partie des connaissances de la langue des locuteurs. Pour vérifier une telle hypothèse, ce travail doit être complété par une expérience psycholinguistique testant le rôle du lemme verbal et de son emploi dans le choix de l'ordre des compléments chez les locuteurs.

Références bibliographiques

Abeillé A., Clément L. & Toussenel F. (2003) Building a treebank for french. In *Treebanks*. Dordrecht : Kluwer.

Abeillé A. & Godard D. (2000) French word order and lexical weight. In R. Borsley (Eds), *The Nature and Function of Syntactic Categories* (Syntax and Semantics 32), p. 325‑358. New-York : Academic Press.

Abeillé A. & Godard D. (2001) A class of lite adverbs in french. In J. Camps & C. Wiltshire (Eds), *Romance syntax, semantics and their L2 acquisition*, p. 9‑25. Amsterdam : John Benjamins. 1

Abeillé A. & Godard D. (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.

Artstein R. & Poesio M. (2008) Inter-coder agreement for computational linguistics. *Computational Linguistic*, 34, 555–596.

Berrendonner, A. (1987) L'ordre des mots et ses fonctions. *Travaux de linguistique,* **14/15**, 9–19.

Blinkenberg A. (1928) *L'ordre des mots en français moderne*. Copenhague : Munskgaard.

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.

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

Bresnan J. & Hay J. (2008) Gradient grammar : An effect of animacy on the syntax of give in New Zealand and American english. *Lingua*, 118(2), 245‑259.

Carletta J. (1996) Assessing agreement on classification tasks : the kappa statistic. *Computational Linguistic*, 22(2), 249–254.

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

Gelman A. & Hill J. (2006) *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gries S. T. (2003) Towards a corpus-based identification of prototypical instances of constructions. *Annual Review of Cognitive Linguistics*, 1, 1–27.

Harrell, F. E. (2001) *Regression modeling strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer-Verlag New York Inc.

Hawkins J. (2000) The relative order of prepositional phrases in english : Going beyond manner-place-time. *Language Variation and Change*, **11**, 231–266.

Howell, D.C. (1998) *Méthodes statistiques en sciences humaines*. Paris: De Boeck Université.

Jäger G. & Rosenbach A. (2006) The winner takes it all - almost. cumulativity in grammatical variation. *Linguistics*, **44**(5), 937–971.

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

Kimball, J. (1973) Seven principles of surface structure parsing in natural language. *Cognition,* **2,** 15–47. CA: Academic Press.

Legendre G., Grimshaw J. & Vikner S. (2001) *Optimality-theoretic syntax*. MIT Press.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100172

Prince A. & Smolensky P. (2004) *Optimality Theory : Constraint Interaction in Generative Grammar*. Blackwell publishers.

Rosenbach A. (2005) Animacy versus weight as determinants of grammatical variation in english. *Language*, **81(3)**, 613–644.

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.

Seddah, D., M. Candito, B. Crabbé, & E. Henestroza Anguiano. (2012) Ubiquitous usage of a french large corpus : Processing the Est-Républicain corpus. In *Proceedings of Language Ressources and Evaluation Conference, (LREC 2012)*. Istanbul.

Stallings, L. M., M. C. MacDonald, & 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.

Thuilier, J., A. Abeillé & B. Crabbé. (2011a) Do animate arguments come first? In *Proceedings of Architectures and Mechanisms for Language Process (AMLAP 2011)*. Paris.

Thuilier, J., A. Abeillé & B. Crabbé. (2011b) Préférences concernant l'ordre relatif des compléments du verbe en français. Colloque AFLS. Nancy.

Thuilier, J. (2012) Semantic annotation of french corpora : animacy and verb semantic classes. In *Proceedings of Language Ressources and Evaluation Conference, (LREC 2012)*. Istanbul.

Wasow T. (1997) Remarks on grammatical weight. *Language Variation and Change*, **9**, 81–105.

Wasow T. (2002) *Postverbal behavior*. CSLI publications.

Zaenen A., Carletta J., Garretson G., Bresnan J., Koontz-Garboden A., Nikitina T., O'Connor M. C. & Wasow T. (2004) Animacy encoding in english : why and how. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation, DiscAnnotation '04*, p. 118–125, Stroudsburg, PA, USA : Association for Computational Linguistics.

[i] http://www.cnrtl.fr/corpus/estrepublicain/

[ii] Ces hypothèses de normalité ne sont généralement pas satisfaites par les données portant sur le langage naturel.

[iii] La version lemmatisée du corpus de l'Est-Républicain est disponible librement sur http://alpage.inria.fr/estrepu/

[iv] Le caractère animé des référents a été annoté manuellement sur l'ensemble des phrases du corpus, à partir des catégories de Zaenen et al. (2004).

[v] Nous n'avons pas testé l'influence de l'ambiguïté potentielle de rattachement du SP. Cependant, les travaux de Wasow (2002) tendent à montrer qu'en anglais, les locuteurs n'utilisent pas les différents ordonnancements possibles comme stratégies d'évitement de l'ambiguïté de rattachement du SP.

[vi] Il faut noter que si les variables concernant la pronominalité, le caractère défini et le caractère animé du SN et du SP sont introduites dans le modèle de régression logistique, elles apparaissent comme étant non-significatives pour la modélisation de la variable ORDRE (cf. Thuilier et al. 2011b).

[vii] Les coefficients associés aux variables sont obtenus par maximisation de la vraisemblance que le modèle donne aux données. Le détail des calculs est expliqué dans Agresti (2007) et Gelman & Hill (2006). Nous avons effectué ces calculs avec le logiciel R (http://www.r-project.org/).

[viii] La table de données contient 151 lemmes verbaux différents car seuls 3 verbes de ER et ESTER n'apparaissent pas dans FTB : *annoncer*, *dire* et *montrer*.

[ix] Le test du rapport de vraisemblance pour la comparaison de modèles n'est possible que lorsque les deux modèles ne diffèrent que par une seule variable. C'est bien le cas des modèles 1 et 2 : le modèle 2 contient la variable LEMMEVERBAL alors que le modèle 1 ne la contient pas.

[x] La tâche d'annotation a été réalisée sur un corpus plus important que celui utilisé ici. Ce corpus comportait 1346 phrases. Les taux d'accord inter-annotateur reportés sont les taux calculés pour les 1346 phrases. Pour plus de précisions sur l'annotation des verbes en classes sémantiques, se reporter à Thuilier (2012).

[xi] Sens locatif abstrait : « placer quelqu'un dans telle situation, telle catégorie ».

[xii] La probabilité de concordance correspond à l'aire sous la courbe ROC.

[xiii] Le modèle 3 ne présente qu'un léger surapprentissage sur les données. Pour l'ensemble des données C=0.956, ce qui signifie que les prédictions sur des données inconnues ne fait perdre que 0.01 à l'exactitude.

[xiv] Les données du lexique Treelex confirment nos données. TreeLex est un lexique de sous-catégorisation automatiquement extrait du FTB, disponible à l'adresse suivante : http://erssab.u-bordeaux3.fr/spip.php?article150 Dans ce lexique, les verbes *expliquer*, *proposer*, *montrer*, *annoncer*, *dire* et *demander* apparaissent dans des cadres de sous-catégorisation qui comportent un objet direct ainsi qu'un objet indirect, et où l'objet direct est réalisé sous la forme d'une complétive (SSub) ou d'une infinitive (VPinf).

[xv] Les verbes que nous avons recueillis dans les corpus FTB, ESTER et ER, et qui apparaissent avec un complément prépositionnel et une complétive ou infinitive sont : *permettre*, *demander*, *montrer*, *dire*, *annoncer*, *expliquer*, *exiger*, *proposer*, *recommander*, *interdire*, *valoir*, *revenir*, *rester*, *remémorer*, *ordonner*, *indiquer*, *imposer*, *déclarer*, *conseiller*.