Influence du genre applicatif sur la réalisation des extractions en dialogue oral : constantes et variations
Task influence on word-order phenomena in spoken dialogue: consistencies and variations

Jean-Yves Antoine
Université François Rabelais de Tours, LI & Lab-STICC-CNRS, Lorient.
Jeanne Villaneau
Université Européenne de Bretagne, VALORIA, Lorient.
Jérôme Goulian
Université Pierre-Mendès France, LIG, Grenoble.

Résumé
Cet article présente une étude de corpus portant sur les variations d'ordre linéaire en français parlé spontané. Nous avons étudié plusieurs corpus de dialogue finalisé correspondant à différentes tâches applicatives afin d'évaluer l'influence du contexte discursif sur ces phénomènes. Nous insistons dans un premier temps sur l'intérêt d'études de corpus pour orienter les recherches en Traitement Automatique des Langues. Nous présentons ensuite notre méthodologie d'analyse ainsi que les principaux résultats de l'étude. Ceux-ci montrent que la tâche et le rôle du locuteur dans l'interaction n'ont pas d'influence significative sur la réalisation des dislocations orales, alors que le degré d'interactivité joue au contraire sur leur fréquence. Ces variations d'ordonnancement respectent toutefois de fortes régularités imposées par le système de la langue. Aussi concluons-nous que le français parlé spontané reste une langue à ordre SVO fixe.

Mots-clés :inversions ; dislocation ; variation de l'ordre linéaire ; parole spontanée ; linguistique de corpus ;

Summary
This paper presents a corpus study on word order variations (WOV) in spontaneous spoken French. We have studied several corpus of spoken dialogue dedicated to different tasks to assess the influence of the discourse context on WOVs. At first, we show how the contribution of pilot corpus studies should benefit to Natural Language Processing researches. Then, we present our methodology and the main results of this study. In particular, we observe that the task and the role of the speaker have no influence on WOVs, while the frequency of WOVs is on the contrary highly influenced by the degree of interactivity of the dialogues. These WOVs respect some noticeable structural regularities which are imposed by French ordering constraints. This is why we conclude that conversational spoken French must be still considered as a language with a rigid SVO ordering.

Keywords: word order variations; spontaneous speech; corpus linguistics.

1. Ordre linéaire et variations en linguistique et ingénierie des langues

La question de l'ordre variable des mots est un sujet d'investigation concernant à la fois la théorie linguistique et le Traitement Automatique des Langues (TAL). Elle a ainsi constitué à la suite des travaux de Tesnière (1959) un des arguments des grammaires de dépendances face aux grammaires de constituants chomskyennes. Les linguistes intéressés par la question ont pour habitude de situer les systèmes langagiers entre les langues à ordre variable (russe, finnois, tchèque par exemple) et les langues à ordre fixe parmi lesquelles on compte le mandarin, l'anglais ou le français (Covington 1990). En règle générale, on observe que plus

une langue a une morphologie pauvre, plus elle est à ordre fixe. Le français est ainsi considéré comme un langage à ordre fixe SVO, pour Sujet-Verbe-Objet(s). Lorsqu'on étudie une langue à ordre fixe, on observe toutefois que l'ordre canonique attendu *a priori* n'est pas toujours respecté. Considérons l'exemple suivant :

(1)*Le poulet je l'ai terminé hier.*

Ici, la mise en avant de l'objet *le poulet* enfreint l'ordre linéaire SVO attendu en français. L'énoncé sera pourtant accepté par tout locuteur francophone. D'une manière générale, on peut distinguer deux niveaux de variabilité de l'ordre des mots dans un énoncé (Holan *et al*. 2000,Hudson 2000) :

- d'une part, une variabilité faible se traduisant par le déplacement d'un constituant complet. Ce mouvement n'induit aucune discontinuité dans la structure de dépendance de l'énoncé. C'est par exemple le cas du mouvement du groupe prépositionnel *pour Rio Sao Paulo* dans l'exemple :

(2) *bon sinon* ***pour Rio Sao Paulo*** *je pense qu'il y a pas mal de vols* (AF.II.8.C68)

- d'autre part une variabilité forte qui se traduit par un relâchement des contraintes de continuités autorisant la production d'énoncés qualifiés de **non-projectifs** (c'est-à-dire présentant un arbre syntaxique dont certaines branches se croisent). Cette variabilité forte peut être observée à l'oral en français comme le montre l'exemple (3), où la position de l'adverbe *maintenant* casse la continuité entre la subordonnée relative *qui est nouveau* et le groupe nominal complexe *un tarif encore plus intéressant sur Londres* qui constitue son antécédent:

(3) *vous savez on a* ***un tarif encore plus intéressant sur Londres*** *maintenant* ***qui est nouveau*** (AF.II.33.O17)

De même, dans l'exemple ci-dessous, l'extraction à droite[1] de la proposition *je crois* casse la continuité de la complétive *se presser pour réserver* :

(4) *Mais il faut quand même assez se presser* ***je crois*** *pour réserver* (AF.I.67.C20)

La distinction entre variabilité forte et faible est primordiale en ingénierie des langues. A titre d'exemple, certains formalismes comme les grammaires de lien, qui ont pourtant montré leur robustesse sur l'anglais, ne sont pas en mesure de modéliser des énoncés discontinus (Sleator et Temperley, 1991). A l'opposé, le traitement des structures syntaxiques discontinues (qui concerne de nombreuses langues à ordre variable) constitue un des fondements d'un formalisme tel que les grammaires syntagmatiques liées par la tête (Pollard et Sag, 94). Il a aussi été étudié (Rambow et Joshi, 94) dans le cadre des grammaires d'arbres adjoints (TAG).

Pour autant, si la question de la variabilité d'ordonnancement linéaire et ses implications sur le traitement automatique des langues ont bien été étudiées sur le langage écrit, aucun travail n'a été mené à notre connaissance sur le français parlé spontané (parole conversationnelle) avec une telle visée. A l'heure où les applications de l'ingénierie des langues s'ouvrent désormais largement à la modalité orale[2], il est important d'avoir une connaissance plus approfondie de ces phénomènes qui ont potentiellement un impact sur l'analyse syntaxique automatique. Dans cet article, nous nous proposons donc de caractériser en corpus les variations d'ordre linéaire en français parlé conversationnel. Ce type d'étude doit intervenir en amont de toute modélisation informatique, dans l'objectif d'éclairer et orienter cette dernière. Nous parlons donc d'étude sur corpus pilote. Afin d'atteindre une certaine généricité de conclusion, il est essentiel que les données étudiées englobent des situations interactionnelles diversifiées. D'où l'importance de recourir à un corpus hétérogène dans le cadre d'une méthodologie d'analyse variationniste.

---

[1]Du point de vue de l'intention du locuteur, ce détachement peut être interprété comme une incise. À un niveau purement syntaxique, un parseur traitera de toute manière cet exemple comme un énoncé avec extraction non projective.

[2] Voir par exemple la récente campagne d'évaluation Ester2 portant entre autres sur la détection des entités nommées dans les flux de parole conversationnelle radio ou télédiffusées (Galliano *et al*. 2009).

2. Etude sur corpus hétérogène des extractions et autres variations d'ordre linéaire en français parlé spontané

Cet article présente ainsi une étude en corpus des variations d'ordre linéaire en français parlé spontané. On sait qu'il n'existe pas de frontière claire entre oral et écrit, et que cette transition doit plutôt être appréhendée sous la forme d'un continuum de genres (Biber, 1988 ; Bilger et Blanche-Benveniste, 1999). Nos travaux concernent cependant un genre de dialogue oral excessivement spontané qui ne peut être identifié à aucun genre écrit. Il nous semble nécessaire de procéder à une analyse détaillée de la variabilité dans ce registre de langue avant d'espérer lui appliquer éventuellement les conclusions de travaux ayant porté essentiellement sur le français écrit jusqu'à présent.

Plusieurs études (Gadet, 1989 ; Blanche-Benveniste et al. 1990) nous ont donné une connaissance approfondie sur la variabilité de l'ordre des mots en français parlé. Ainsi, les différents procédés qui président aux phénomènes d'extraction (inversion, double-marquage, présentatif…) sont bien identifiés. Ces études linguistiques ont toutefois porté principalement sur des récits ou des interviews et bien plus rarement sur une parole conversationnelle hautement interactive. Surtout, ces travaux présentent un caractère descriptif et explicatif qui ne rend pas compte de l'importance quantitative de ces phénomènes. Dans la perspective de modélisation informatique (TALN) qui est la notre, une telle étude quantitative, menée sur des corpus représentatifs, est pourtant nécessaire.

La notion de corpus représentatif pose ici problème. Les phénomènes de variations d'ordre linéaire relèvent de procédés complexes qui ne peuvent être appréhendés par des techniques d'analyse quantitative automatique (textométrie). On doit en conséquence s'en remettre à une annotation manuelle des phénomènes observés. Compte tenu du coût humain du processus d'annotation, il n'est pas envisageable de considérer des corpus de grande dimension: l'étude que nous présentons dans cet article porte ainsi sur seulement 100 000 mots (cf. § 3.1). La représentativité du corpus ne peut donc reposer sur la masse de donnée étudiée, mais au contraire sur l'hétérogénéité du corpus: en regroupant des situations d'interactions variées, suivant des dimensions bien choisies, on peut ainsi espérer obtenir des conclusions présentant une certaine généricité (et donc une certaine représentativité) sur une quantité d'observables relativement modérée. C'est cette approche variationniste sur corpus hétérogène que nous allons illustrer dans cet article.

3. Méthodologie

3.1. Corpus hétérogène : dimensions pour une étude variationniste

Notre étude porte sur les variations d'ordre linéaire. Dans cette perspective, nous avons tout d'abord cherché à caractériser a priori les facteurs susceptibles d'influer, d'une situation interactionnelle à l'autre, sur la réalisation de ces phénomènes. Trois dimensions ont retenu notre attention :

- le media concerné par l'interaction: le dialogue se réalise-t-il de manière directe (en face à face) ou est-il médié par un dispositif de communication tel que, par exemple, le téléphone. On peut supposer que l'interaction directe, qui permet aux interlocuteurs de mesurer directement les effets de leurs prises de parole, conduit à des réalisations linguistiques différentes de celles rencontrées dans une communication téléphonique,
- le degré d'interactivité, estimé par la fréquence des interruptions et des chevauchements entre les tours de parole des interlocuteurs. On peut là encore estimer qu'un dialogue très interactif conduira à des procédés particuliers,
- enfin, le degré de finalisation de la tâche. Dans une perspective de dialogue homme-machine finalisé, nous nous intéressons à des dialogues dirigés par la

réalisation d'une tâche donnée. Il semble que plus une tâche est finalisée (donc moins complexe), plus peuvent apparaître des procédés de topicalisation, des ellipses etc. qui ont potentiellement une influence sur l'ordonnancement linéaire dans les tours de parole.

Ces trois dimensions permettent de caractériser des situations interactionnelles variées. On peut alors procéder à la constitution d'un corpus hétérogène regroupant plusieurs sous-corpus qui se différencient précisément par ces facteurs de variabilité. L'étude quantitative des extractions sur les sous-corpus peut alors fournir des conclusions générales qui seront supposées représentatives si elles sont vérifiées sur tous les corpus. A l'opposé, l'observation d'éventuelles variations entre les sous-corpus permet de caractériser les dimensions interactives qui influent sur la réalisation des phénomènes étudiés. C'est ici tout l'intérêt méthodologique d'une étude menée sur un corpus hétérogène[3].

3.2. Corpus étudiés

Notre étude différentielle a été menée sur quatre corpus de dialogue oral spontané portant sur trois tâches applicatives:

- réservation et renseignement aériens (corpus Air France)
- information touristique (corpus Murol et OTG)
- accueil standard téléphonique (corpus UBS)

Ces corpus se différencient suivant les dimensions interactives caractérisées précédemment (tableau 1). Notons toutefois que cette étude ne concerne qu'un registre : l'oral spontané en situation de dialogue finalisé vers une tâche précise.

- Le corpus **Air France** (AF), recueilli par Marie-Annick Morel à l'Université de la Sorbonne Nouvelle, puis retravaillé par Pierre Nerzic, réunit des conversations téléphoniques entre un centre de réservation aérienne et des clients qui peuvent être des particuliers ou des personnels d'agence de voyage. Le degré de finalisation de cette application est élevé. L'interactivité est celle d'une conversation téléphonique professionnelle : le dialogue reste contenu, l'hôtesse étant tenue à une certaine réserve.
- Le corpus **Murol** du laboratoire CLIPS-IMAG (désormais LIG) réunit un ensemble de conversations téléphoniques simulées entre deux compères jouant respectivement le rôle d'un touriste et d'un employé d'office de tourisme. La conversation peut porter sur des problèmes de localisation dans une ville ou de recherche variée d'activités pour un séjour touristique. Le domaine de la tâche est donc moins finalisé que dans le corpus précédent. Nous sommes en présence de dialogues assez longs se caractérisant par une interactivité très marquée (chevauchements très fréquents, par exemple).
- Le corpus **OTG** (Office du Tourisme de Grenoble) a été enregistré par le CLIPS-IMAG et transcrit par le laboratoire VALORIA (Nicolas et al. 2002). Il regroupe 315 dialogues réels entre des touristes francophones et le personnel d'accueil d'un office du tourisme. Le corpus est distribué librement sur le site *Parole Publique* (www.info.univ-tours.fr/~antoine/parole_publique/). Le degré de finalisation est équivalent à celui du corpus Murol dont il partage le cadre applicatif. On observe par contre que l'interactivité est plus faible. En particulier, la fréquence des chevauchements est significativement plus basse. Comme pour le corpus Air France, il semble qu'ici le personnel d'accueil semble tenu à une réserve et une écoute que l'on ne retrouve pas sur le corpus Murol, qui était simulé.

[3] Une étude de pertinence statistique des résultats obtenus sur les différents sous-corpus du corpus hétérogène étudié peut renforcer cette représentativité des résultats

- Le corpus **UBS** (Université de Bretagne Sud) a été collecté et transcrit par le laboratoire VALORIA de l'université éponyme. Il réunit 40 conversations téléphoniques réelles entre des individus (étudiants, parents mais aussi personnels de l'université même) et les réceptionnistes de cette université. Les dialogues concernent des sujets variés comme la simple demande de mise en communication avec un membre du personnel, mais aussi des questions pédagogiques complexes. Le dialogue est modérément finalisé, compte tenu de la variété des motivations des personnes qui appellent le standard. Le degré reste modéré, le personne d'accueil adoptant le plus souvent une réserve et une écoute proche de celle du corpus Air France. L'interactivité peut toutefois être plus élevée dans le cas où l'appelant est un collègue de l'université. Les dialogues correspondant sont toutefois minoritaires. Ce corpus est également distribué librement sur le site *Parole Publique*.

Corpus	Durée	Nb. de dialogues	Nb. de tours de parole	Nb. de mots	Media	Tâche	Interactivité	Type de dialogue
Air France	n.c.	103	5 149	49 700	Téléphone	Réservation aérienne	Modérée	Très finalisé
Murol	n.c.	9	1 078	13 500	Téléphone	Information touristique	Elevée	Finalisé
OTG	2 heures	315	n.c.	25 000	Direct	Information touristique	Modérée	Finalisé
UBS	1 heure	40	n.c.	10 000	Téléphone	Standard	Assez modérée	Peu finalisé

Tableau 1 − Description des corpus

3.3. Annotation du corpus

Nous avons procédé à un recensement des extractions sur les quatre corpus. Chaque variation a été annotée manuellement par trois experts suivant une procédure de validation croisée. L'annotation a consisté à enrichir chaque observation par quatre caractéristiques :

- Direction – Sommes-nous en présence d'un élément antéposé (5b) ou postposé (5c) par rapport à l'ordre SVO attendu en français (5a) :

(5a) *Je rencontrerai Alice demain*
(5b) *Alice je la rencontrerai demain*
(5c) *Je la rencontrerai demain Alice*

- Type – D'un point de vue structurel, il est d'usage de distinguer quatre types de variations (Gadet 1989) par rapport à l'ordre canonique attendu (6a). Les *inversions* correspondent au simple déplacement d'un élément dans l'énoncé (6b) alors que les *dislocations* (encore appelées *double marquage*) se traduisent par l'utilisation d'un pronom qui rappelle l'élément extrait à la position attendue. Si l'on prend l'exemple de l'énoncé (6c), on observe que l'élément extrait *Alice* est rappelé par relation anaphorique par le pronom *la*. Celui-ci est bien situé à la position attendue, puisque nous sommes en présence d'un clitique toujours positionné avant le verbe en français. Ici, la variation d'ordre linéaire est lexicalement marquée.

(6a) *Je rencontrerai Alice demain*
(6b) ***Demain*** *je rencontrerai Alice*
(6c) **Alice** je **la** rencontrerai demain

A l'opposé, les *présentatifs* reposent sur un marquage syntaxique : un élément initial (*c'est* ou les introductifs construits avec le verbe avoir : *il y a* / *j'ai* / *on a* etc.) introduit explicitement la partie de l'énoncé détachée et est suivi d'une subordonnée introduite par *qui* ou *que*. Les énoncés clivés (7a) ou pseudo-clivés (7b) constituent une forme particulière de présentatifs.
(7a) ***C'est Alice que*** *je rencontrerai demain*
(7b) ***Celle que*** *je rencontrerai demain* ***c'est Alice***

Finalement, le dernier type de variation est appelé énoncé binaire. Dans ce cas, le tour de parole est divisé en plusieurs fragments qui ne partagent plus aucune relation syntaxique entre eux :
(8) *Mon vélo le rouge la roue arrière elle est crevée*

- Fonction syntaxique de l'élément déplacé – Quatre catégories de fonction ont été distinguées : *sujet* (9a), *argument* de la valence du verbe (9b), *modifieur* qui correspond globalement aux compléments non sous-catégorisés du verbe (9c) et enfin *compléments de phrase* (9d) tels que définis par (Blanche-Benveniste, 1992).

(9a) ***Jean il*** *est parti*
(9b) ***Le gâteau*** *il* ***l'****a mangé*
(9c) ***Le lundi*** *je ne travaille pas*
(9d) *L'avion* ***évidemment*** *sera plus coûteux*

- Discontinuité – Enfin, l'annotation précise si la variation induit ou non une discontinuité dans la structure de l'énoncé. On retrouve ici la définition de variation forte ou faible vue précédemment: l'énoncé (2) correspond ainsi à une variation faible, à la différence des énoncés (3), (4) ou (8) où on observe une variation forte.

4. Résultats

Ce paragraphe présente les principales observations issues de cette étude de corpus. Dans un premier temps, nous allons étudier ces résultats dans toute leur généralité. Nous reviendrons ensuite sur les données rendant compte des variabilités observées entre chaque corpus.

Corpus	Interactivité des dialogues	Degré de finalisation	Media	Fréquence Moyenne	Ecart-Type	Minimum	Maximum
Air France	modérée	Très finalisé	Téléphone	**13.6 %**	10.5 %	0.0 %	30.8 %
Murol	élevée	Finalisé	Téléphone	**25.6 %**	10.2 %	10.2 %	37.5 %
OTG	modérée	Finalisé	Direct	**13.5 %**	11.7 %	0.0 %	50.0 %
UBS	Assez modérée	Peu finalisé	Téléphone	**12.2 %**	7.1 %	0.0 %	22.1%

Tableau 2. Fréquence d'apparition des extractions (% de tours de parole présentant une extraction). L'écart-type, le minimum et le maximum sont estimés sur la base d'une distribution par dialogue.

4.1. Importance relative des phénomènes d'extraction

Le tableau 2 présente la fréquence d'occurrence moyenne des extractions sur nos quatre corpus. Il montre que ces phénomènes peuvent être très répandus en français oral spontané. Les variations affectent ainsi de 12,2 à 25,6% des tours de parole, suivant le corpus considéré.

Cette fréquence d'apparition est variable: on observe ainsi que certains dialogues ne présentent aucune variation d'ordre linéaire (colonne *minimum* du tableau 2) alors que d'autres présentent une extraction dans un tiers, voire la moitié, des tours de parole (colonne *maximum*).

Ces différences de fréquence d'usage se retrouvent entre les corpus étudiés. En particulier, le corpus *Murol* présente un usage plus fréquent des extractions qui est statistiquement significatif. Au vu des caractéristiques de chaque corpus (cf. tableau 2 et tableau 1, § 3.1.), cette variabilité ne peut s'expliquer que par le plus fort degré d'interactivité du corpus. En effet, ni le type de média ni le degré de finalité ne peuvent expliquer cette variation : le téléphone (et donc la communication indirecte) se retrouvent dans deux autres corpus et le degré de finalisation du dialogue est identique à celui du corpus OTG, qui présente pourtant un taux de variations d'ordonnancement bien plus faible. Il existe ainsi une corrélation positive entre la fréquence des variations d'ordre linéaire et l'interactivité du dialogue. Cette observation est assez intuitive: il n'est pas étonnant qu'un dialogue plus interactif donne lieu à l'utilisation plus fréquente de procédés permettant une topicalisation ou un effet d'insistance.

4.2. Sens de l'extraction

Le tableau 3 présente la répartition par corpus des variations suivant leur direction (antépositions vs. postpositions). Les résultats sont homogènes et montrent que les antépositions dominent très largement (82,5 à 89,3% des cas suivant le corpus).

Corpus	Antéposition	Postposition	Ecart-Type
Air France	82.5 %	17.5 %	20.4 %
Murol	85.5 %	14.5 %	8.7 %
OTG	87.9 %	12.1 %	16.9 %
Accueil UBS	89.3 %	10.7 %	17.7 %

Tableau 3 − Distribution des variations d'ordre linéaire en fonction de leur direction

Ce résultat recoupe les observations de (Blanche-Beveniste 1998) et (Gadet 1989) selon lesquelles l'antéposition est le mode de réalisation privilégié des topicalisations en français parlé spontané. Par ailleurs, (Pekarek-Doehler 2001) a montré que les antépositions sont également utilisées pour renforcer l'interaction en reprenant un élément prononcé dans le tour de parole précédent.

4.3. Fonctions syntaxiques

Le tableau 4 présente la distribution, corpus par corpus, des éléments extraits en fonction de leur fonction syntaxique. A première vue, il est difficile de tirer des conclusions génériques. On observe que, suivant le corpus considéré, les sujets représentent de 25,4 à 42,5% des éléments extraits tandis que les arguments de valence représentent de 5,3% à 15,3% des cas. Les modifieurs concernent de 21,4% à 27,4% des extractions alors que la variabilité est bien plus grande pour les compléments phrastiques qui se retrouvent dans 20,3% à 45,8% des cas.

Corpus	Sujet	Argument de valence	Modifieur	Complément de phrase
Air France	30.7 %	12.0 %	27.4 %	30.0 %
Murol	25.4 %	5.3 %	23.5 %	45.8 %
OTG	42.5%	11.8%	25.4%	20.3%
Accueil UBS	34.4%	15.3%	21.4%	29.0%

Tableau 4 – Distribution des extractions par classes de fonctions syntaxiques

Certaines régularités distributionnelles peuvent toutefois être relevées par delà cette variabilité inter-corpus. Tout d'abord, la fonction *Sujet* est celle qui donne lieu au plus grand nombre d'extractions, à l'exception du corpus *Murol* où elle arrive derrière les compléments phrastiques. Une analyse qualitative montre que la prépondérance des compléments de phrases extraits dans ce corpus est due à la mise en avant de marqueurs de discours (*à ce moment là, maintenant, donc* …) que les locuteurs utilisent pour garder la maîtrise de l'interaction. Ces marqueurs se situent souvent entre l'explétif et le complément de phrase réel. C'est toutefois ce dernier rôle que leur assignerait un parseur syntaxique, raison pour laquelle nous avons décidé de les considérer comme tels ici.

Si l'on met de côté cette fonction qui demanderait certainement une analyse plus fine, la distribution des extractions suivant la fonction de l'élément extrait reste la même sur tous les corpus : les sujets extraits sont les plus nombreux suivis par les modifieurs et enfin, avec une fréquence d'apparition bien moindre, par les arguments de valence du verbe. Cet ordre *Sujet > Modifieur > Argument* est suffisamment marqué pour relever la prééminence du système de la langue sur les productions orales spontanées. Une analyse détaillée de chaque fonction syntaxique va nous le montrer.

Sujets – Etudions tout d'abord la prééminence des sujets extraits. Nous avons remarqué précédemment que la grande majorité des extractions est réalisée sous la forme d'une antéposition. Cette situation se retrouve pour les sujets (cf. tableau 5). Dans ce cas, le sujet antéposé reste en amont du verbe, ce qui ne modifie pas l'ordre SVO attendu. Par ailleurs, cette antéposition est réalisée souvent par un double marquage, comme dans l'exemple (10)

(10) *l**e Cargo il** est là là où il y a la la la demoiselle* [OTG]

Corpus	Sujet	Toutes fonctions confondues
Air France	81.6 %	82.5 %
Murol	90,6 %	85.5 %
OTG	87.7 %	87.9 %
Accueil UBS	91.1 %	89.3 %

Tableau 5 − Proportion d'antépositions dans les extractions de *Sujet* en fonction de leur direction

Dans cet exemple, le pronom de reprise est à la position canonique du sujet et l'extraction conduit à un schéma *S'SVO* incluant l'ordre *SVO* standard. Au final, les extractions d'éléments *Sujet* n'enfreignent généralement pas le schéma d'ordonnancement SVO privilégié par le français. Cela explique, de notre point de vue, la facilité avec laquelle les locuteurs semblent pouvoir réaliser de telles extractions en français oral spontané.

Modifieurs – Les modifieurs donnent fréquemment lieu à extraction puisqu'ils représentent de 21.4% à 27.5% des variations d'ordonnancement linéaire. Etant donné que les modifieurs ne sont pas concernés par l'ordre SVO canonique, leur extraction n'est pas contrainte: le locuteur peut dès lors les réaliser assez librement, ce qui explique la fréquence importante de leurs extractions.

Arguments – Il semble au contraire que l'ordre SVO soit un frein à l'extraction d'un argument de valence, comme le montre notre étude quantitative (entre 5.3% et 15.3% des extractions dans le tableau 5). Ici, en effet, le déplacement d'un argument O va casser l'ordre canonique comme dans l'exemple (11a) qui nous conduit à une séquence OSV :

(11a) ***la 'Science en fête'** non non on **l'**a pas reçu* [OTG]

(11b) ***la 'Science en fête'** non non on a pas reçu*

Preuve de cette résistance, cette variation donne lieu généralement à un double-marquage : un pronom anaphorique est utilisé pour rappeler l'argument extrait à la position où il était attendu. L'inversion simple de l'argument, comme dans l'exemple artificiel (11b) ne

permet pas ce rappel. Nous verrons plus loin que l'usage de procédé est précisément beaucoup plus rare.

Compléments de phrase – Revenons pour terminer sur la catégorie multiforme des compléments phrastiques. Comme les modifieurs, cette fonction syntaxique n'est pas directement concernée par l'ordre SVO. Un locuteur francophone est toutefois à même de porter un jugement sur la ou les positions préférentielle(s) de ces éléments. Par exemple, dans l'énoncé (12a), la conjonction *donc* est attendue de manière préférentielle en début d'énoncé. A l'opposé, son positionnement entre l'objet et le modifieur du verbe, comme dans l'énoncé (12b) est clairement identifié comme une variation de l'ordre attendu révélant l'intention discursive du locuteur.

(12a)*j'ai donc passé ma licence à l'UBS*

(12b)*j'ai passé ma maîtrise* ***donc*** *à l'UBS* [UBS]

Cette extraction préserve l'ordre SVO puisque l'énoncé (12b) répond au schéma linéaire SVOA'A où A' est le complément de phrase et A le modifieur. Cette indépendance vis à vis de l'ordre canonique explique là encore la fréquence élevée d'extraction des compléments de phrase.

Nous avons remarqué plus haut que cette fonction syntaxique est celle qui présente la plus forte variabilité en termes de fréquence d'extraction d'un corpus à l'autre (de 20,3% à 45,8% des extractions dans le tableau 4). Englobant la catégorie des marqueurs du discours, les compléments de phrase sont souvent l'objet d'une extraction à fins de contrôler de l'interaction (maintien du canal de communication, marque d'approbation ou de désaccord). Il n'est donc pas étonnant que le corpus le plus interactif (Murol) soit celui où leurs extractions sont les plus fréquentes.

4.4. Régularités structurelles : fonction syntaxique et type de procédé

Au cours du paragraphe précédent, nous avons noté que certaines fonctions syntaxiques semblent donner lieu à l'utilisation privilégiée d'un procédé d'extraction particulier. Aussi avons-nous procédé à une analyse des corrélations entre fonction syntaxique et type structural d'extraction (inversion, double-marquage, présentatif et énoncés binaires) pour vérifier si les contraintes imposées par l'ordre canonique SVO ne guidaient pas ces choix de procédé.

Sujets – Le tableau 6 présente la distribution des procédés rencontrés pour l'extraction des sujets. Il distingue :

a) les dislocations et les présentatfs, qui sont lexicalement ou syntaxiquement marquées,
b) les inversions et énoncés binaires pour lesquels aucun indice ne témoigne du déplacement.

Corpus	Murol	Air France	OTG	UBS
Dislocations + présentatif	100 %	95.7 %	97,9 %	100 %
Inversion + énoncé binaire	0 %	4.3 %	2,1 %	0 %

Tableau 6 – Distribution des extractions de sujets en fonction du type de procédé

On observe sur tous les corpus une utilisation quasi-exclusive (supérieure à 95% des cas) des procédés d'extraction marqués linguistiquement. Cette prédominance s'explique par une influence de l'ordre SVO standard. En effet, le déplacement d'un élément de l'énoncé est un des procédés de thématisation les plus fréquents en langue naturelle. Dans le cas des langues à ordre variable, il revient le plus souvent à mettre en première position de l'énoncé l'élément topicalisé. Dans une langue à ordre SVO fixe comme le français, cette mise en avant n'est pas possible pour le sujet, qui occupe déjà cette position. D'où le recours à une antéposition marquée. Considérons les exemples suivants :

(13a) *il l'avait remplacée*

(13b) **c'est lui qui** l'avait remplacée [Accueil UBS]

(14a) *la dame veut quelques renseignements*
(14b) ***la dame elle*** *veut quelques renseignements* [Accueil UBS]

Dans les deux cas, l'utilisation d'un présentatif par clivage (13b) ou d'une dislocation avec double-marquage (14b) conduit à une structure de la forme[4] S'SVO où S' représente l'élément extrait. Ainsi, dans les deux cas, le procédé retenu permet de respecter l'ordre SVO tout en mettant en avant le thème du discours. Les post-positions sont réalisées également à l'aide de procédés marqués. La volonté de respecter l'ordre SVO y est également manifeste: un pronom tient la place S de l'élément extrait, donnant lieu à une cataphore.

Le tableau 7 montre qu'il est toutefois possible de rencontrer des déplacements de sujets non marqués. L'écrasante majorité de ces exceptions correspond toutefois à des tournures idiomatiques:
(15a) ***l'agence X*** *est à l'appareil*
(15b) ***c'est l'agence X*** *à l'appareil* [Air France]
(15c) ***c'est l'agence X qui est*** *à l'appareil*

L'expression "*c'est X à l'appareil*" que l'on retrouve dans l'énoncé (15b) constitue ainsi un mode d'introduction au téléphone ritualisé dans le milieu professionnel. Il ne viendrait à l'idée à personne d'utiliser une formulation, pourtant canonique, telle que (15a) et il est intéressant de noter que le procédé observé ici correspond à la réalisation d'une ellipse sur un procédé d'extraction marqué syntaxiquement, à savoir le clivage de l'énoncé (15c).

Arguments – Le tableau 7 représente la distribution des extractions d'arguments de valence en utilisant le même regroupement de procédés (marqués vs. non marqués), puisque les arguments sont eux-aussi concernés par l'ordre SVO.

Corpus	Murol	Air France	OTG	UBS
Dislocation + présentatif	77.3 %	67.3 %	80.5 %	63.1 %
Inversion + énoncé binaire	22.7 %	32.7 %	19.5 %	26.9 %

Tableau 7 – Distribution des extractions d'arguments de valence en fonction du type de procédé

On observe là encore un usage privilégié des procédés marqués. Cette prédominance est toutefois moins marquée, et varie sensiblement d'un corpus (63.1% sur UBS) à l'autre (80.5% sur OTG). Les exemples de variations non marqués sont ainsi plus fréquents et bien plus variés. Considérons l'exemple ci-dessous :
(16a) *oui elle a eu AES*
(16b) *oui* ***AES*** *elle a eu* [Accueil UBS]

Nous sommes ici en présence d'une simple inversion de l'objet *AES*. Tout en restant compréhensible, cet énoncé, qui enfreint l'ordre SVO attendu pour une topicalisation OSV, sera jugé incorrect à l'écrit par un lecteur francophone. L'observation assez fréquente de ces inversions simple nous montre toutefois que dans le fil du dialogue, les interlocuteurs acceptent sans problème cette disfluence orale: la cohérence pragmatique de l'interaction semble leur importer plus que la correction syntaxique des énoncés. Si l'ordre SVO a une influence sensible sur l'extraction des arguments de valence en français oral, les locuteurs peuvent donc également se permettre de l'enfreindre sans conséquence. Les analyseurs automatiques chargés de traiter la parole spontanée se devront de modéliser de tels procédés de mise en relief.

Modifieurs – Comme nous l'avons déjà observé, les modifieurs ne sont pas concernés par l'ordre SVO. Si l'on considère que cet ordonnancement canonique guide le choix du procédé

[4] Dans l'énoncé (13b), l'objet est certes positionné avant le verbe, cette situation étant simplement due au positionnement canonique des pronoms clitiques en français.

utilisé pour réaliser une extraction, le recours à un dispositif de marquage n'est pas nécessaire ici. C'est exactement ce que montre le tableau 8.

Corpus	Murol	Air France	OTG	UBS
Inversions	93,5 %	96,8 %	78,9 %	89,3 %
Autres procédés	6,5 %	3,2 %	21,1 %	10,7 %

Tableau 8 – Distribution des extractions de modifieurs en fonction du type de procédé

On observe ici un usage privilégié des inversions sur les autres types structurels d'extraction. Cette prédominance est variable, puisqu'elle va de la quasi-exclusivité (Murol, Air France) à une forte majorité d'utilisation (78,9% sur le corpus OTG). L'utilisation d'extractions marquées n'est ainsi pas rare. Par exemple, l'énoncé (17b) correspond à l'utilisation d'un clivage pour mettre en avant le complément de lieu « *à la TAG* ». Comme le montre l'exemple inventé (17c), l'utilisation d'une simple inversion non marquée aurait été parfaitement acceptable à l'oral. Il n'est toutefois pas évident que ces deux énoncés correspondent exactement à la même intention discursive: il semble ainsi que ce soit le contexte pragmatique, et non des considérations syntaxiques, qui influe sur le choix du procédé à retenir.

(17a) *vous pouvez la retirer à la TAG*
(17b) ***c'est à la TAG*** *que pouvez* [...] *vous pouvez la retirer* [OTG]
(17c) ***à la TAG*** *vous pouvez la retirer*

L'analyse détaillée des observations nous montre par ailleurs que les structures clivées sont privilégiées lorsque la topicalisation apparait dans une question, comme sur l'exemple (18b).

(18a) *où puis-je me renseigner*
(18b) ***c'est où que*** *je peux me renseigner* [OTG]

Enfin, il est intéressant de noter que nous avons observé un nombre significatif d'énoncés binaires correspondant à l'extraction d'un modifieur. Dans l'exemple (19b), le groupe prépositionnel *de Madame X* est un modifieur qui joue le rôle de complément du groupe nominal *le numéro direct*. Son antéposition crée une structure binaire: il n'y a plus de relation syntaxique entre l'élément extrait *Madame X* et le reste de la phrase, comme le montre la figure 1. Seule l'anaphore associative due au déterminant possessif *son* permet de maintenir la cohérence sémantique de l'énoncé.

(19a) *J'ai pas le numéro direct de Madame X*
(19b) ***Madame X*** *j'ai pas son numéro direct* [UBS]

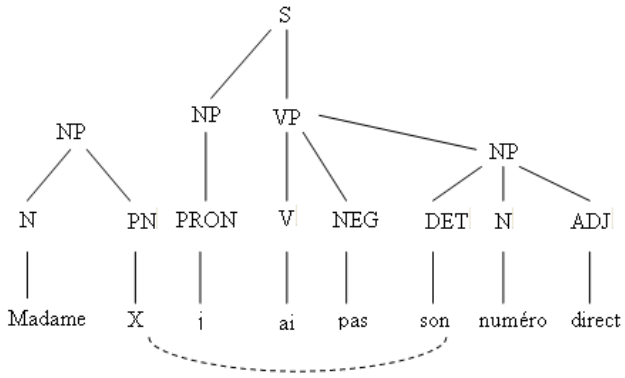

Figure 1 – Structure syntaxique de l'énoncé binaire *Madame X j'ai pas son numéro direct*

Compléments de phrase – Comme les modifieurs, les compléments de phrase ne sont pas concernés par l'ordre canonique SVO. Il n'est donc pas étonnant que les simples inversions soient privilégiées lors de leur extraction, comme le montre le tableau 9. Cette prédominance

est toutefois bien plus marquée que pour les modifieurs. Les inversions constituent la seule forme de variation de complément de phrase observée sur trois des quatre corpus, et elles représentent 99,2% des inversions sur le corpus *OTG*. Les rares exceptions que nous avons observées ici correspondent à des formulations très spécifiques qui ne peuvent être considérées comme représentatives. On peut donc estimer que les simples inversions constituent la seule forme d'extraction de compléments de phrase en français parlé spontané.

Corpus	Murol	Air France	OTG	UBS
Inversion	100 %	100 %	99.2 %	100 %
Autres procédés	0 %	0 %	0.8 %	0 %

Tableau 9 – Distribution des extractions de compléments de phrase fonction du procédé

En guise de synthèse, on peut conclure que de fortes contraintes structurelles influent sur la réalisation des extractions en français parlé spontané. Celles-ci sont dues avant tout à l'influence de l'ordre canonique SVO du français et se caractérisent par deux observations principales :

- d'une part, les extractions d'éléments concernés par l'ordre canonique sont réalisées de manière très privilégiées par des procédés marqués (double marquage ou présentatif)
- d'autre part, l'extraction des éléments positionnellement plus libres que sont les modifieurs ou les compléments de phrase fait appel essentiellement à des inversions. Dans le cas des compléments de phrases, le recours à l'inversion est même exclusif.

Enfin, on remarque que les énoncés binaires, qui cassent la structure syntaxique de l'énoncé, sont rares dans nos corpus (tableau 10). Le corpus OTG en présente une proportion non négligeable (6,9% des extractions). Une étude attentive montre toutefois qu'un quart de ces énoncés binaires sont concentrés sur des prise de parole de la même hôtesse d'accueil. Il semble donc que nous soyons en présence d'une idiosyncrasie personnelle qui a d'autant plus d'influence que ce procédé est globalement rare.

Une autre forme d'altération de la structure syntaxique d'un énoncé est l'absence de projectivité de la structure extraite. Notre dernière étude va précisément consister à étudier l'ensemble des variations fortes de l'ordre linéaire canonique attendu.

Corpus	Murol	Air France	OTG	UBS
Inversion	67,2 %	60,2 %	34.8 %	48,0 %
Dislocation (double-marquage)	17,0 %	28,3 %	44,2 %	30,1 %
Présentatif	15,6%	10,5 %	15,4 %	18,7 %
Enoncé binaire	1,2 %	0,1 %	6.9 %	3,3 %

Tableau 10 – Distribution globale par procédé des variations d'ordre linéaire.

4.5. Projectivité

Cette dernière analyse a consisté à étudier dans quelle mesure les variations d'ordonnancement correspondent à des variations fortes, ce qui signifie qu'elles altèrent la structure syntaxique de l'énoncé au point de la rendre non projective. En traitement automatique des langues, on qualifie de structure non projective tout arbre syntaxique dont les branches se croisent au moins une fois (Holan, 2000). De telles structures se rencontrent fréquemment dans les langues à ordre variables, telles que le russe (figure 2), où l'on observe la présence de constituants discontinus. En disloquant la structure syntaxique des énoncés, les extractions fortes conduisent précisément à des arbres de constituants ou de dépendances non projectifs. Or, il se trouve que certains formalismes d'analyse syntaxiques, telles que par exemple les grammaires de liens (Sleator et Temperley 1991), ne peuvent traiter les structures non projectives. Il est donc essentiel de quantifier la part d'extractions fortes en français parlé spontané.

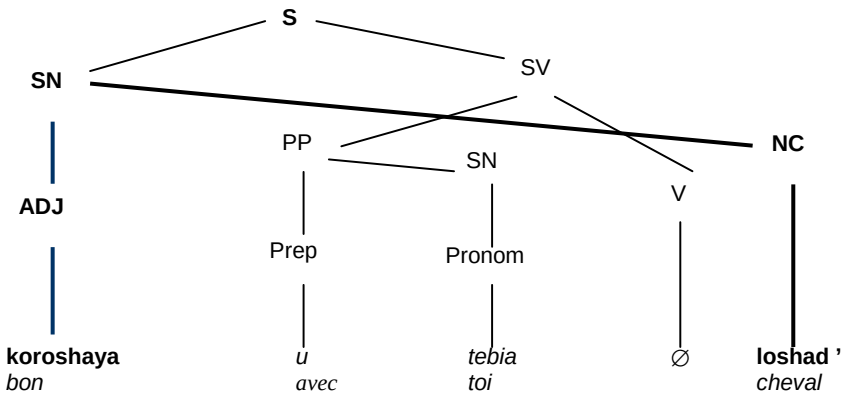

Figure 2 – Exemple de structure non projective en russe.

Le tableau 11 présente précisément la proportion et la fréquence moyenne de ces variations fortes, observées sur nos quatre corpus d’études.

Corpus	**Murol**	**Air France**	**OTG**	**UBS**
% d’extractions non projectives	0.7 %	2.3 %	2.2 %	3.1 %
% de tours de parole avec extraction non projective (variation forte)	0.3 %	0.4 %	0.3 %	0.4 %

Tableau 11 – Distribution des variations fortes d’ordonnancement linéaire.

Au total, les variations fortes conduisant à des énoncés non-projectifs représentent dans tous les cas moins de 0,4 % des énoncés de nos corpus oraux. Au final, on peut tenir les extractions non-projectives pour marginales sur ce genre de français parlé. Ce résultat est important du point de vue de l’ingénierie des langues : il montre en effet que des formalismes d’analyse projectifs restent parfaitement adaptés au traitement de la parole spontanée.

5. Conclusion

En conclusion, ce travail mené sur une banque de corpus diversifiés de par leur cadre applicatif et leur situation d’interaction montre l’intérêt pour l’orientation des recherches en TALN des études linguistiques amont sur des corpus pilotes hétérogènes. Il nous a en effet permis de tirer des conclusions opératives sur le type de formalismes d’analyse adaptés au langage oral, et a également permis de caractériser les formes d’extractions qui devaient être modélisées par les analyseurs de l’oral. Il suggère également la nécessité de mener des études variationnistes sur des collections hétérogènes de corpus. Ces travaux ont en effet permis de mieux caractériser les sources d’influence qui président à l’apparition des variations d’ordonnancement linéaire en français parlé conversationnel : alors que le genre, la tâche et la position du locuteur dans l’interaction ne semblent pas avoir d’influence majeure sur la réalisation des dislocations, le contexte discursif et plus précisément le degré d’interactivité jouent sur la fréquence d’occurrence de ces phénomènes. Globalement, on relève toutefois que ces variations respectent de remarquables régularités structurales qui doivent certainement au système même de la langue française. C’est en ce sens que nous proposons de considérer que le français parlé spontané reste une langue à ordre fixe sous influence de l’ordre canonique SVO. Seule la constitution et l’étude d’une banque de corpus hétérogènes en termes de genre et de situation discursive peut, à notre sens, permettre de tirer ce type de conclusion.