Word order phenomena in conversational spoken French
A study on task-oriented dialogue corpora and its consequences on language processing

Jean-Yves Antoine (1), Jérôme Goulian (2), Jeanne Villaneau (3) Marc Le Tallec (1)
(1) Université François Rabelais Tours, LI
(2) LIG, Université Pierre Mendès-France
(3) VALORIA, Université Européenne de Bretagne
Jean-Yves.Antoine@univ-tours.fr

Abstract

This paper presents a corpus study that investigates the question of word order variations (WOV) in spontaneous spoken French and its consequences on the parsing techniques that are used in Natural Language Processing. We have studied four task-oriented spoken dialogue corpora which concern different application tasks (air transport or tourism information, switchboard calls). Two corpora concern phone conversations while the other two correspond to direct interaction. Every word order variation has been manually annotated by 3 experts, following a cross-validation procedure. Our results show that, while conversational spoken French should be highly affected by WOVs, it should also still be considered as a rigid order language: WOVs follow some impressive structural regularity and they result very rarely in discontinuous syntactic structures. As a result, non-projective parsers remain well adapted to conversational spoken French.

1. Introduction

The development of speech technologies would not have been conceivable without the availability of large speech corpora. From the turn of the millennium, numerous electronic corpora have been constituted to fulfil the needs of speech processing. They concern audio file collections (speech corpus), as well as speech transcripts (spoken language corpus). As a result, the speech modality of an increasing number of idioms can now be studied through these resources. It is, therefore, disappointing to observe that the availability of this interesting material did not really lead to better knowledge of spontaneous spoken language.

This situation should be easily explained: speech technologies use in a large extent data-driven techniques (stochastic models, neural networks, support vector machines…) which consider corpora only as training data. Their aim is not to understand the language they are working on, but to use a blind but efficient algorithm to adapt at best a model to some representative data. These data-driven approaches have definitively shown their efficiency: although speech technologies have suffered for years from a damning lack of robustness, they have enabled the emergence of a new industry offering a large panel of applications. This is why knowledge-based systems, which contrariwise are based on an explicit linguistic description, have been taken away from speech and signal processing. However, linguistically motivated approaches still have a future in natural language processing where it has successfully challenged data-driven approaches in a great number of applications (POS tagging, text retrieval, automatic translation, named entities detection…). It is not insignificant to quote that spoken language processing is also concerned by these successes of knowledge-based approaches: for the time being, Markovian models have not revealed a significant superiority on grammar-based systems in spoken language understanding (Pallet, Fiscus *et al*. 1994, Villaneau & Antoine 2009), and most of the research on stochastic dialogue modelling have been abandoned.

The aim of this paper is precisely to show how corpus linguistics can help spoken language processing by providing knowledge-based parsers with some useful linguistic description. More precisely, we will investigate the question of word order variations which is of first importance for spoken language parsers.

2. Word order variations and parsing of conversational spoken French

Word order variations (WOV) have frequently held the attention of linguists, as well as computational linguists. Following the pioneering work of Tesniere (1959), word order variations were, for instance, a central element in the debate between dependency grammars and Chomsky's phrase grammars. More recently, several formal studies (Pollard, Sag, 1994; Rambow, Joshi 1994, Holan et al. 2000) have demonstrated that every parsing formalism (link grammars, TAG, HPSG, LFG…) handles different kinds of word order variations.

In particular, one can distinguish between two kinds of word order variations (Hudson 2000, Bartha et al. 2006): **strong variations** lead to the apparition of a discontinuity in the dependency structure of the utterance, while **weak variations** keep this structure continuous.

(1) on a **un tarif plus intéressant pour Londres** maintenant **qui est nouveau** (AirFrance.II.33)

(Transl.) *we have **a fare more interesting to London** now **which is new***

In the example (1), the extraction of the adverb *now* separates the relative clause from its antecedent, thereby splitting the syntactic structure of the utterance.

Such discontinuous structure can not be parsed by projective formalisms such as dependency grammars (Holan et al. 2000). A precise knowledge on how word order variations occur in a considered language should therefore shed light on the choice of an adequate parsing formalism in a useful way. Considering human-machine dialogue applications, this paper investigates this question on conversational spoken French. It is a common practice to distinguish between free or rigid word order languages (Hale 1983, Covington 2000). Written French is usually considered, in its written modality, as a rigid order language. However, our observations on task oriented interactions tend to show that spontaneous spoken French presents a higher variability. This study aims at quantifying this spoken influence. More precisely, we have carried out a quantitative corpus study to answer the following questions:

- To what extent is conversational spoken French concerned by WOV?
- Do these WOV follow some general structural tendencies?
- Do language registers (Biber 1988) have an influence on WOV?

- What is the average frequency of strong variations, and consequently, are projective formalisms adapted to the parsing of task-oriented conversational spoken French?

3. Corpus study: methodology

3.1. Corpus collection

To reach some inter-register genericity, our study investigates WOV on four task-oriented spoken dialogue corpora (Table 1) which concern three application tasks:

- air transport reservation (Air France Corpus)
- tourism information (Murol corpus and OTG corpus)
- switchboard calls (UBS corpus)

Three corpora (Air France, Murol, UBS) concern phone conversations while the two others correspond to direct human-human interaction.

Corpus	Overall duration	Number of dialogues	Number of speech turns	Number of words	Media	Task
Air France	n.c.	103	5,149	49,700	Phone	Air transport information
Murol	n.c.	9	1,078	13,500	Phone	Tourism information
OTG	2 hours	315	n.c.	25,000	Direct	Tourism information
UBS	1 hour	40	n.c.	10,000	Phone	Switchboard call

Table 1 − Synthetic description of the studied corpora

The Air France corpus was collected in the 1990s by Marie-Annick Morel (U. Sorbonne Nouvelle) and transcribed under the supervision of Pierre Nerzic (IRISA). It contains real conversations between the Air France call centre and different customers who would be either a private individual or a travel agent. The dialogues exclusively concern flight reservations. The audio recordings of this corpus are no longer available, unlike the orthographic transcripts.

The Murol corpus was collected and transcribed by the CLIPS-IMAG (now LIG) laboratory (Bessac, Caelen 1995). It concerns interaction between two subjects who are simulating a real phone conversation between a tourist and the receptionist of a tourism office. According to the followed scenario, the dialogue should concern the resolution of a localisation problem (e.g. *where is the zoological park situated*?) or the definition of a one-day activity schedule (e.g. *are there interesting things to do with children in your town*?). Once again, the audio recordings seem to be lost but the transcripts are still available.

The OTG (*Office du Tourisme de Grenoble*) corpus was collected by the CLIPS-IMAG and transcribed by the VALORIA laboratory (Nicolas et al. 2002). It contains hundreds of real dialogues between tourists and a receptionist of the tourism office of Grenoble. The microphones were hidden during the recording: the tourists were informed of their existence at the end of the conversation. The corpus is distributed freely on the *Parole Publique* website[1].

Finally, the UBS corpus was collected and transcribed by the VALORIA laboratory. It concerns real phone conversations between individuals and the receptionist of the switchboard and reception office of a university. The dialogues concern various topics, from a simple switchboard inquiry (e.g. *may I talk to Mr X please*) to a complex schooling question (e.g. *I have a problem, the elective module I have passed does not appear on the transcripts*). This corpus is also distributed freely on the *Parole Publique* website[1].

3.2. Corpus annotation

The speech transcripts of the four corpora have been annotated to carry out a quantitative analysis on word order variations. The automatic annotation of such phenomena is beyond the current state of the art in natural language processing. As a result, every word order variation has been manually annotated by 3 experts, following a cross-validation procedure. Such an annotation represents an important workload, which explains why our study is restricted to these four corpora (around 100,000 words). However, the results that are presented in this paper have been validated by a statistical test of significance: Student, Wilcoxon-Mann-Whitney or χ2 test (Dudewicz, Mishra 1988). More precisely, every word order variation has been described with four complementary features:

Direction: does the variation correspond to an ante-position or a post-position? The French language follows a standard SVO order. For instance, the example (2a) represents the standard order while (2b) corresponds to an ante-position of the object *Alice* and (2c) to a post-position.

(2a) Je rencontrerai Alice demain
(Transl.) *I will meet Alice tomorrow*

(2b) Alice je la rencontrerai demain
(Transl.) *Alice I will her meet tomorrow*

(2c) Je la rencontrerai demain Alice
(Transl.) *I will her meet tomorrow Alice*

Type: from a structural point of view, four main types of word order variations should be distinguished in spoken French: *inversions* correspond to a simple move of the shifted element (3b), while *marked extractions* come with a pronoun whose aim seems to recall the element at its standard place (3c). While these extractions are lexically marked, *presentative structures* are syntactically marked. Typical illustrations of this kind of variation are cleft (3d) and pseudo-cleft sentences (3e).

(3a) Je rencontrerai Alice demain
(Transl.) *I will meet Alice tomorrow*

(3b) **Demain** je rencontrerai Alice
(Transl.) **Tomorrow** I will meet Alice

(3c) **Alice** je **la** rencontrerai demain
(Transl.) ***Alice*** *I will* ***her*** *meet tomorrow*

(3d) **C'est Alice que** je rencontrerai demain
(Transl.) ***It is Alice that*** *I will meet tomorrow*

(3e) **Celle que** je rencontrerai demain **c'est Alice**
(Transl.) ***The one who*** *I will meet tomorrow* ***it is Alice***

Finally, the last type of WOV is called *binary sentences* since the spoken utterance appears to be completely spit in two or more fragments that do not share any syntactic relationship:

(4) Mon vélo le rouge la roue arrière elle est crevée
(Transl.) *My bike the red one the rear wheel it is flat*

Syntactic function of the shifted element: we have decided to classify the different functions that exist in French into four different categories:

- *subjects*, which cannot be considered as ordinary arguments in French, since the verb always agrees in gender and number with the subject,
- *valence arguments* which are the compulsory complements subcategorised by the verb,
- *modifiers* which usually correspond to adverbial complements,
- *phrase complements* which should be considered as modifiers of the whole speech turn rather than a direct complement of the verb.

The examples below illustrate these different situations (5a: subject, 5b: argument, 5c: modifier, 5d: phrase complement)

(5a) **Jean il** est parti
(Transl.) ***John he** is left*
(5b) **Le gâteau** il **l'**a mangé
(Transl.) ***The cake** he **it** ate*
(5c) **Le lundi** je ne travaille pas
(Transl.) ***On Monday** I do not work*
(5d) L'avion **évidemment** sera plus coûteux
(Transl.) *The plane **obviously** will be more expensive*

Discontinuity: finally, we will also note if the observed word order variation results in a discontinuous syntactic structure or not. One should be aware that if binary sentences always present a discontinuity, the other kinds of WOV should lead to a discontinuity as well. See, for instance, the example (1) where the displacement of the modifier *now* corresponds to a simple inversion.

4. Annotated corpus analysis: results

We have carried out several quantitative analyses (average frequencies and standard deviation, statistical distributions…) on the annotated corpus to draw a precise picture on how WOV occur in conversational spoken French. Every individual analysis has been conducted on the four corpora to reach a certain amount genericity. Some analyses deal with a unique annotated feature alone, while the other ones combine several features in order to study their mutual influences. This section describes the main results provided by this corpus analysis.

4.1. Frequency of occurrence of WOV

Table 2 presents the average frequency of occurrence of word order variations on the four corpora. This frequency has been computed as the percentage of speech turns that have at least one WOV. Standard deviation is counted on the basis of a per-dialogue distribution. Likewise, minimum and maximum correspond to the extreme variations of the frequency on every individual dialogue. For instance, at least one dialogue in the Air France corpus presented no WOV, while in another dialogue, around 3 speech turns over 10 were affected by a WOV.

Corpus	Average frequency	Standard deviation	Minimum	Maximum
Air France	**13.6 %**	10.5 %	0.0 %	30.8 %
Murol	**25.6 %**	10.2 %	10.2 %	37.5 %
OTG	**13.5 %**	11.7 %	0.0 %	50.0 %
UBS	**12.2 %**	7.1 %	0.0 %	22.1%

Table 2 − Frequency of occurrence of WOV (% of affected speech turns) in the four corpora.

Generally speaking, these results show that conversational spoken French should be highly affected by word order variations. On the average, from 12.2% to 25.6% of the speech turns of our corpora are affected by at least one WOV. One should also note that the average frequency noticeably varies from one dialogue to another and from one corpus to another as well. This observation is quite intuitive, since word order variations are usually motivated by a topicalisation whose need is directly related to the evolution of the dialogue. Clearly, there is a positive correlation between the frequency of WOV and the interactivity of the dialogue. For instance, WOV are twice more frequent in the Murol corpus than in the other ones. Now, we have observed that the interactivity was significantly higher in the Murol corpus (overlapping speech turns are, for instance, very frequent). One possible reason for this higher interactivity is certainly that the dialogues were simulated in the Murol corpus. As a result, the interaction was more informal and less civilized than in the real dialogues.

4.2. Direction of WOV

Table 3 gives the distribution of word order variations according to their direction. The percentage of both variables Ante-position (“***Jean il** est parti*“ (Transl.) ***John he** is left*) and Post-position (*“**Il** est parti **Jean**"* (Transl.) ***he** is left **John***) is given for each of the corpora. Since the sum of these two variables is equal to 1, Standard-Deviation (as previously counted on the basis of a per-dialogue distribution) is the same for two of the variables.

Corpus	Ante-position	Post-position	Standard-Deviation[1]
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 %

Table 3 − Distribution of the wov according to their direction.

[1] The standard deviation of the ante-posisition and post-position is of course the same as P (ante-position) = 1 – P (post-position)

The table shows a strong pre-eminence of *Ante-position* on *Post-position*. Results are stable, with percentages of Ante-position between 82.5% and 89.3%.

These results are not surprising since *Ante-position* is a classic way of topicalization in spoken French, as is shown in the studies of linguists who have studied spoken French, such as Claire Blanche-Benveniste and Françoise Gadet (Blanche-Benveniste 1998, Gadet 1989). According to S. Pekarek-Doehler, *Ante-position* is also used to intensify interaction between the speakers of a spoken dialogue: by signalling the link with the spoken expression in the previous turn, it points out the legitimacy of talking to the speaker (Pekarek-Doehler 2001).

4.3. Syntactic function of the extracted element

Table 4 shows the syntactic function of the extracted elements. The four chosen categories are *Subject*, *Valence argument*, *Modifier* and *Phrase Complement:* they are described section 3.2. The results are not as clear or as stable as those related to the direction of WOV: from 25.4% to 42.5% for *Subject* function, from 5.3% to 15.3% for *Valence argument*, from 21.4% to 27.4% for *Modifier* and from 20.3% to 45.8% for *Phrase Complement*. To have a clearer view of these results, Figure 1. gives a representation of them as a diagram.

Corpus		**Subject**	**Valence argument**	**Modifier**	**Phrase complement**
Air France	average	30.7 %	12.0 %	27.4 %	30.0 %
Murol	average	25.4 %	5.3 %	23.5 %	45.8 %
OTG	average	42.5%	11.8%	25.4%	20.3%
Accueil UBS	average	34.4%	15.3%	21.4%	29.0%

Table 4 - Syntactic function of the extracted elements.

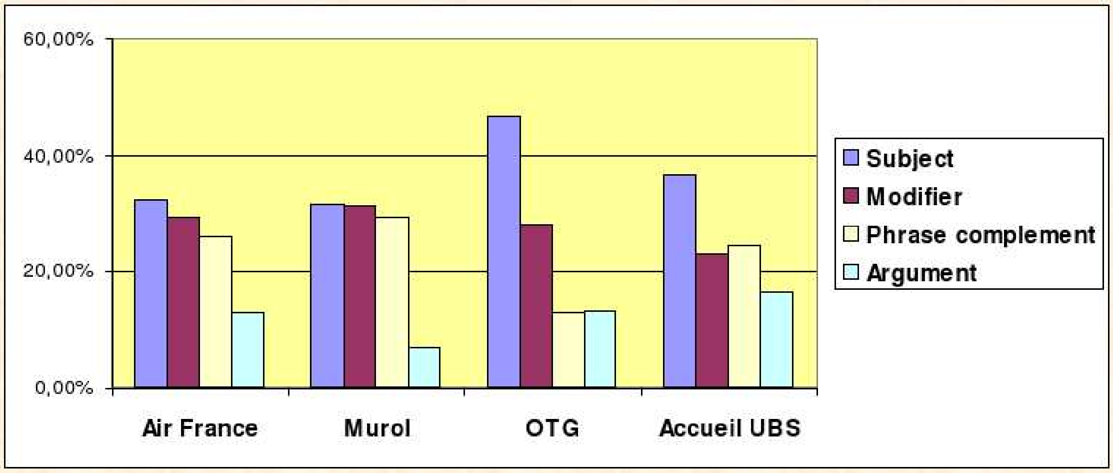

Figure 1 – Syntactic function of the extracted elements as a diagram.

In spite of the differences between corpora, we can observe that the *Subject* function is the most frequent in each of them. The *Modifier* function comes second, except in the Accueil UBS Corpus, while the *Argument* function *is* the less frequent, except in the OTG corpus. While the order *Subject, Modifier*, *Phrase Complement,* *Argument* is not always strictly respected, no corpus really deviates from it. It can therefore be regarded as a significant tendency.

The pre-eminence of the *Subject* function is not surprising. In many cases, the subject of a sentence is related to its topic. So, an extraction of this subject is used to topicalize this subject, as in (6a), or to present it, as in (6b).

(6a) l**e Cargo il** est là là où il y a la la la demoiselle [OTG]
(Transl.) ***The Cargo it*** *is there there where there is the the the young lady.*
(6b) **j'ai ma fille qui** s'est inscrite [Accueil UBS]
(Transl.) ***I have my daughter who*** *registered.*

Since subject extraction is an *Ante-position* in most of the cases (90%) and since the pattern used is of an SS'OV form, the base order, SVO, is strictly respected.
Syntactic functions which come behind the *Subject* function in terms of WOV are *Modifier* and *Phrase Complement*. Both can be found in variable positions of a sentence in French. The ante-position or post-position of a *Modifier* (7a) or of a *Phrase Complement* (7b) can be used to emphasize this element, while respecting the canonical word order SVO.

(7a) ***avant le six juin*** *elle a dû dû valider un de trois voeux* [Accueil UBS]
(Transl.) *before six in June she had to to validate one of three wishes.*
(7b) *et* ***à ce moment-là*** *vous aurez une heure et une date* [Accueil UBS]
(Transl.) *and at this moment you will have an hour and a date.*

Argument is the less frequent of syntactic functions used in Word Order Variations (WOV). This type of WOV can be prejudicial to base order SVO and sophisticated structures may be observed. In most of the cases, argument WOVs are used for topicalization, as in the following examples (8a), (8b) and (8c). In both ante-position, as in the example (8a), and post-position, as in (8b), we have *marked extractions*, with a pronoun found at the right place of the argument. Speech turn (8c) is an example of post-position WOV in an interrogative sentence. In spoken French, the mark of interrogation is intonation only in most of the cases, preserving SVO order, while according to grammatical rules, the “right“ form of interrogation is inversion, with VSO order. According to F. Gadet, the intonation form of interrogation, named *total form* by F. Gadet, can reach 95% of the interrogations in a speech corpus (Gadet 1989): a clue to the preference of spoken French for SVO order. In example (8c), SVO order is respected but the argument is announced by the pronoun (*“**les”)*** for topicalization on the object of the search.

(8a) **la 'Science en fête'** non non on **l'**a pas reçu [OTG]
(Transl.) ***the “Science in feast”*** *no we* ***it*** *not received.*
(8b) vous **l'**avez pas **celui-là** [OTG].
(Transl.) *You have not* ***it that one***
(8c) où vous **les** rangez **vos grands sacs poubelles** [Accueil UBS]
(Transl.) *where you tidy up* ***them your big bin liners****.*

4.4. Syntactic function and WOV type

The results presented in this section are related to WOV type: *Marked Extraction, Presentative Structure, Inversion* and *Binary* according to syntactic function (*Subject, Argument, Modifier* and *Phrase Argument*) of the extracted element. These WOV types are described in section 3.2. One of our objectives is to investigate the validity of our previous assumption according to which, in spoken French, the canonical order SVO is preserved in most of WOV.

Table 5. shows WOV type related to the *Subject* function by opposing both types *Marked Extraction* and *Presentative Structure* to both types *Inversion* and *Binary*.

Corpus	**Extraction + Presentative**	**Inversion + Binary**
Air France	95.4 %	4.6 %
Murol	100.0 %	0.0 %
OTG	97.1 %	2.9 %
Accueil UBS	100.0 %	0.0 %

Table 5 –*Subject* function and WOV type.

In each of the four corpora, a very strong pre-eminence of *Marked* E*xtraction* and *Presentative Structure* types may be observed for the *Subject* function. In both types, *Presentative Structure* (9a) or *Marked Extraction* (9b)*: Ante-position* is a common rule.

(9a) **c'est lui qui** l'avait remplacée [Accueil UBS]
*(*Transl.) ***it is him who*** *had replaced her*
(9b) **la dame elle** veut quelques renseignements [Accueil UBS]
(Transl.) ***the lady she*** *want some information*

In both examples, (9a) and (9b), the main pattern is S'SVO: S' is a presentation in (9a) and a noun phrase in (9b), and pronouns are found in the canonical place S of the element. So, both correspond respectively to a syntactic and a lexical marking, which aims at recalling the standard SVO order.

The following table (Table 6) is related to the WOV type for the less frequent syntactic function: the *Argument* function. Since such WOV can violate SVO order, their examination is especially interesting for our purpose. As in the previous table, both *Marked* E*xtraction* and *Presentative Structure* types are opposed to both *Inversion* and *Binary* types.

Corpus	**Extraction + Presentative**	**Inversion + Binary**
Air France	67.3 %	32.7 %
Murol	77.3 %	22.7 %
OTG	80.3 %	19.7 %
Accueil UBS	60.0 %	40.0 %

Table 6 – Function *Argument* and WOV type.

Pre-eminence of both *Marked Extraction,* as in examples (9c) and (9d), and *Presentative Structure,* as in example (9e), is less clear than for *the Subject* function and the results are more corpora dependent. Nevertheless, with numbers between 60% (Accueil UBS) and 80.3% (OTG), these two types are always more frequent that both *Inversion* (9f) and *Binary* (9g), whose frequency is contained between 19.7% (corpus OTG) and 40% (corpus Murol).

(9c) si **les diplômes** on pouvait venir **les** retirer... [Accueil UBS]
(Transl.) *if **degrees** we could come to remove **them.***
(9d) vous pouvez pas **le** perdre **celui-là** [OTG]
(Transl.) *you can not lose it that one.*
(9e) **c'est ce que** j'ai fait [Accueil UBS]
(Transl.) ***it is what** I made.*
(9f) oui **AES** elle a eu [Accueil UBS]
(Transl.) *yes **AES** she has passed.*
(9g) **c'est pour quand votre location** vous m'avez dit [OTG]
(Transl.) ***it is for when your rent** you said to me.*

In both examples of *Argument* WOV*:* (9c) for *Ante-position* and (9d) for *Post-position*, a pronoun marks the right place of the argument. These pronouns act as French clitics and are therefore regularly localised between subject and verb. In *Presentative Structure* (9e), we have an SVO order in the first part of the sentence "*c'est ce que*", followed with the canonical order OSV related to relative clauses in *"que j'ai fait"*. This order is due to the dual role (pronoun and conjunction) that relative pronouns play in French. It doesn't correspond to the standard SVO order of principal clauses. This is why popular French relative clauses try more and more frequently to avoid this particular structure: for example, a form such as *"c'est cela que je l'ai fait"* (Transl. "*it is it that I made it*") can be found (Gadet 1989). In this example, the use of the pronoun "*l'*" makes it possible to replace OSV order by SOV order. In *Binary* WOV as (9f), SVO order is meaningless, since global sentence structure is broken. Besides, such *Argument* inversions are unusual: such an alteration of the canonical SVO order is rather shocking for a French speaker, and only the pragmatic context (all the interaction is about the "AES" diploma of the student) should motivate this specific WOV.

As a conclusion, *Argument* WOV are less common than *Subject*, *Modifier* or *Phrase Complement* WOV and in most of *Argument* WOV, we have *marked extractions*, with the use of pronouns to preserve canonical word order. In French, nevertheless, *Argument* inversions can break canonical word order and spoken French parsers have to take into account these types of phenomena.

On the opposite, *Modifiers* are not directly concerned by SVO canonical order. One should, therefore, expect the Modifier WOV are not based on the same types. Table 7 precisely shows the WOV types related to the *Modifier* function. It opposes inversions with the other kinds of variations.

Corpus	**Inversion**	**Others**
Air France	96.8 %	3.2 %
Murol	93.5 %	6.5 %
OTG	78.2 %	21.8 %
Accueil UBS	89.3 %	10.7 %

Table 7 – Function *modifier* and WOV type.

In all corpora, we note a predominance of *Inversion* with regard to other WOV types.

(10a) **là** je suis à Lorient [Accueil UBS]
(Transl.) *now I am at Lorient**.***

(10b) **actuellement** j'ai des problèmes d'internet [Accueil UBS]
(Transl.) *Actually I have some problems with internet.*
(10c) vous prendrez votre billet à l'aéroport **directement** [Murol]
(Transl.) *You will take your ticket at the airport directly.*
(10d) **c'est à la TAG** que pouvez [...] vous pouvez la retirer [OTG]
(Transl.) *It is at TAG you can […] you can remove it.*
(10e) **c'est où que** je peux me renseigner [OTG]
*(*Transl.) *It is where that I can inquire.*

For instance, in (10a) and (10b), we have an *Inversion* with ante-position. In (10c), inversion is created by post-position. As *Modifiers* are not directly concerned by SVO canonical order, one should easily understand that Modifiers WOV are frequently achieved by a simple inversion. In the last two examples (10d) and (10e), WOV type is *Presentative structure* (clef sentences) with ante-position. But it would be possible to find also a simple inversion in the example (10d) :

(10d') **à la TAG** vous pouvez la retirer
(Transl.) *At TAG you can remove it.*

The presentative in example (10e) corresponds to an interrogative sentence. This kind of cleft-sentence is frequently used in French for topicalization of question words ("wh-question").

The same kind of observations should be noticed with *Phrase Arguments*, which are once again not directly concerned by SVO order. Table 8 presents the distribution of WOV types related to the *Phrase argument* function. It opposes *Inversions* to other structures. *Phrase argument* variations are always marked by *Inversions* in three of the corpora, while only 6.2% are marked by other types of WOV in the OTG corpus. As a result, we notice a very strong pre-eminence of *Inversions.* Like *modifiers*, *Phrase Argument* WOV do not affect the canonical SVO order. As a result, they generally correspond to inversions, which are not lexically or syntactically marked.

Corpus	**Inversion**	**Others**
Air France	100 %	0.0 %
Murol	100 %	0.0 %
OTG	93.8 %	6.2 %
Accueil UBS	100 %	0.0 %

Table 8 – Function *phrase argument* and WOV type .

In (11a) the inversion is in *Post-position* and (11b) is an *Ante-position.*

(11a) c'était pas marqué **en fait** que j'avais fait un choix... [Accueil UBS]
*(*Transl.) *It was marked in fact that I had chosen*
(11b) et **donc** je postule pour l'IGER [OTG]
(Transl.) *And thus I apply for IGER*

As a synthesis, Figure 2 compares the distribution of the four types of WOV (inversion, extraction, presentative, binary) according to the studied corpus.

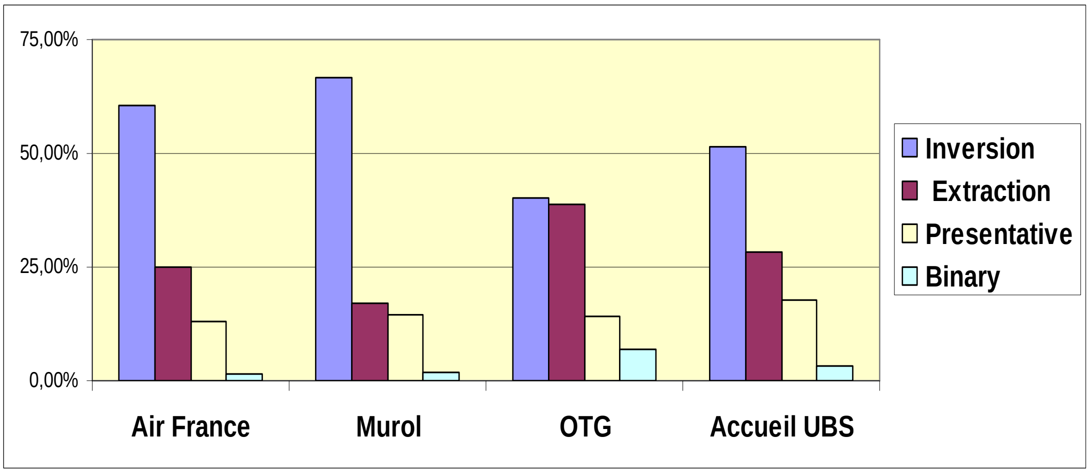

Figure 2 – Distribution of WOV types on every corpus.

This variationist study clearly shows that WOV variations follow strong structural regularities. Indeed, we can observe that inversions are always the most used word order variations. Marked Extractions are second before presentative structures and binary sentences. Most of the time, they are used to preserve canonical word order by using a pronoun at the right place of the extracted element. Even if in the OTG corpus, the difference between *Inversions* and *marked extractions* is not as great, the order *Inversions, marked extractions, Presentative, Binary* sentences is strictly respected in all corpora. This predominance of Inversion is certainly due to the relative importance of Modifier and Phrase complement WOV.

4.5. Projectivity

In this part, we investigate the projectivity of extractions. As shown in Table 9, the frequency of discontinuities due to the extractions is very limited.

Corpus	% of non-projective extractions	% of discontinuous speech turns
Air France	2.3 %	0.4 %
Murol	0.5 %	0.2 %
OTG	2.2 %	0.3 %
Accueil UBS	3.1 %	0.4 %

Table 9 –Distribution of the WOV to the syntactic function.

We can observe a very low amount of non-projective extractions in all corpora, between 0.5% (Murol) and 2.3% (Air France). As a result, detachments leading to non-projective statements represent less than 0.4 % of the statements of our four corpora.

This result is very important from a NLP perspective. It clearly shows that one should use projective formalisms to parse spontaneous spoken French. The resulting degradation of performance will remain very limited, especially when it is compared with the influence of automatic speech recognition errors.

Conclusion

In this paper, we have presented a corpus analysis, the aim of which is to provide natural language processing with a detailed description of word order variations in conversational spoken French. This study has been achieved with the use of four annotated corpora of task-oriented dialogues, which guarantee certain genericity to our results.

This study shows that, while conversational spoken French should be affected by a high rate of word order variations, spontaneous spoken French should still be considered as a rigid order language: most of the observed variations correspond to weak variations and result very rarely in discontinuous syntactic structures. Non-projective parsers therefore remain well adapted to conversational spoken French. Besides this important result for natural language processing, this study shows that WOV follow some impressive regularities:

- *Ante-positions* are preferred to *Post-positions*,
- *Subjects* are significantly more affected than *arguments*, while order varations also concern significantly *modifiers* and *phrase complements*,
- Most *subject* WOV are lexically (pronoun) or syntactically (cleft or pseudo-cleft sentence) marked, while modifier variations usually result in a simple inversion.

Such results are very interesting for the prototyping of spoken language systems. In our opinion, they illustrate the contribution of corpus linguistics to natural language engineering quite well.

Notes

1. Parole Publique website: http://www.info.univ-tours.fr/~antoine/parole_publique

References

Bartha, C., Spiegelhauer, T., Dormeyer, R., Fischer, I. (2006). Word order and discontinuities in dependency grammar. *Acta Cybernetica*, 17(3), 617–632.

Bessac M., Caelen J. (1995) “Analyses pragmatiques, prosodiques et lexicales d'un corpus de dialogue oral homme-machine”. Proc. *JADT'95*, Roma, Italia. 363–370

Biber D. (1988). *Variation across speech and writing*. Cambridge : Cambridge University Press.

Blanche-Benveniste C. (1998). *Le français parlé, études grammaticales*. CNRS,.

Covington M. (1990). “Parsing discontinuous constituents in dependency grammar”. *Computational Linguistics*, 16(4), 234-236

Dudewicz E. J., Mishra S. N. (1988). *Modern mathematical statistics*. Wiley series in probability and mathematical statistics, New-York:John Wiley & Sons, NJ.

Gadet, F. (1989). *Le français ordinaire*. Colin. Paris.

Hale, K. (1983). “Warlpiri and the Grammar of Non-configurational Languages”. *Natural Language and Linguistic Theory*, 5-47.

Holan T., Kubon, Oliva K., Plátek M. (2000). “On complexity of word order”. *Traitement Automatique des Langues, TAL.*, 41(1) 273-300.

Hudson R. (2000). “Discontinuity”. *Traitement Automatique des Langues, TAL*. 41(1), 15-56.

Nicolas P., Letellier-Zarshenas S., Schadle I., Antoine J.-Y., Caelen J. (2002). Towards a large corpus of spoken dialogue in French that will be freely available: the “*Parole Publique*” project and its first realisations. Proc. *Language Resources and Evaluation Conference, LREC’2002*. Las Palmas de Gran Canaria, Spain. 649-655

Pallet, D.S., Fiscus, J.G. and al. (1994). “1994 benchmark tests for the ARPA spoken language program”. Proc. *ARPA workshop on spoken language technology*. Princenton: Morgan Kaufman, NJ. 5–36.

Pekarek Doehler, S. (2001). *Dislocation à gauche et organisation interactionnelle*. in: Marges Linguistiques , vol. 2, 177-194.

Pollard C., Sag I. (1994). Head-driven Phrase Structure Grammar. Chicago: University of Chicago Press, IL.

Rambow O., Joshi A. (1994). “A formal look at dependency grammars and phrase-structure grammars with special considerations of word-order phenomena ”. In Wanner L. (ed.). *Current issues in Meaning-Text Theory*. London: Pinter, UK.

Tesnière L. (1959). *Eléments de syntaxe structurale*. Paris :Klincksiek.

Villaneau, J., Antoine, J.-Y. (2009). “Deeper spoken language understanding for man-machine dialogue on larger application domains: a logical alternative to concept spotting”. Proc. *EACL Workshop on the Semantic Representation of Spoken Language, SRSL’2009, EACL’2009*, Athens, Greece, 50-57.