A corpus-based study of bridge verbs in long distance dependencies in French: a specific construction

Lolita Bérard, Henri-José Deulofeu, Sylvain Kahane

A corpus-based study of bridge verbs in long distance dependencies in French: a specific construction

Lolita Bérard°, Henri-José Deulofeu*, Sylvain Kahane†
°Université Sorbonne Nouvelle – Paris 3, *Aix-Marseille Université,
†Université Paris Ouest Nanterre La Défense

This paper proposes a new analysis of long distance dependencies phenomena. The data collected through corpora indicate that the bridge verb and its dependents follow a very specific template limited to a verb with a modal interpretation and clitic pronouns. We propose an analysis based on complex predicates formation, in opposition to current analyses based on clause embedding. Therefore, there is no longer a need for long distance processes: the movement of the relative or interrogative pronoun remains local. All the constraints on this pattern will be formulated in constructional terms.

Keywords: spoken French syntax; long distance dependencies; constructions; extraction; wh-words

1. Description of selected Long distance patterns

Here are three examples which involve apparent long distance dependencies between the items in italics:

(1) (a) *qui* tu penses qu'il faut *contacter* ?
'*who* do you think that we must *contact*?'
(b) *celui* que tu penses qu'il faut *contacter*
'*the one* that you think that we must *contact*'
(c) c'est *moi* qu'il faut qui *parle* maintenant non
'it is *me* that (it) must (that) *speak* now'

CHIMERA. Romance Corpora and Linguistic Studies 1 (2014), 137-155. ISSN 2386-2629

These data, which are respectively instances of a wh-question (1a), a relative clause (1b) and a cleft sentence (1c), have been widely studied in the literature, from the point of view of their extension as well as of the constraints they obey[1].

Up to recent times, the data used in studies on long distance dependencies mainly consisted of intuition based sentences. Recent corpus-based studies (Verhagen 2005) as well as experiment-based ones (Ambridge & Goldberg 2008) widened the empirical basis. In the present study we will be using authentic data from spoken and written French corpora. The data under analysis came from the following corpora for a total amount of 12M words, including 3M of spoken data: Corpus Evolutif de Référence du Français (CERF), Corpus de Référence du Français Parlé (CRFP) (Equipe DELIC 2004), C-Oral-Rom (Deulofeu & Blanche-Benveniste 2006), Corpus de Français Parlé Parisien (CFPP) (Branca-Rosoff *et al.* 2012), Corpus Oral-Nancy, Traitement de Corpus Oraux en Français (TCOF), Phonologie du Français Contemporain (PFC) (Durand *et al.* 2005), Choix de Textes du Français Parlé (CTFP) (Blanche-Benveniste *et al.* 2002), CPROM (Auchlin *et al.* 2012), Corpus IRIT, Corpus Brassens, Office de Tourisme de Grenoble (OTG) (Antoine *et al.* 2002) and Corpus Accueil UBS. This empirical basis presents variations in interaction structure (monologues, dialogues and interviews), variation in media (face to face, phone call, TV/radio show), variations in contents (real life situations, professional experiences, political discussions, public speeches, literature, technical writing, press etc.) and variations concerning speakers: age, education, social and geographic origin.

1.1 Corpus-based approach

Based on the seminal work of Ross (1967), mainstream generative grammar has favoured syntactic solutions relying on the notion of *island* syntactic constraints on embedded constructs. Since the beginning, however, this formal approach has been challenged by functional ones providing evidence that the constraints are, at least in part, of semantic or pragmatic nature. According to Verhagen's (2005) corpus-based study of wh-interrogatives, the Dutch *bridging* verbs belong to lexical classes defined by their pragmatic function. Basically, they include verbs that contribute to establish an intersubjective relationship between speaker and addressee. For instance, in the case of question (1a), the *bridging* verb conveys an anticipation of the modal stance of the addressee. On the other

[1] Long distance dependencies could be observed in other syntactic contexts such as indirect questions or topicalisation. For the moment, the semi-automatic procedure we use to gather the data from raw transcription makes it too time consuming for us to get the relevant examples in such contexts. We leave this topic open for further researches.

hand, Ambridge & Goldberg (2008), following Erteschik-Shir (1997), claim that information structure is mainly involved in determining possible long distance dependencies: according to them, the *gap* that is identified with the *filler* constituent cannot be within a constituent that has particularly low discourse prominence (i.e., that is "backgrounded"). This is because the filler constituent in long-distance dependency constructions plays a prominent role in the information structure: it is anomalous to consider an element as at once backgrounded and discourse-prominent. Backgrounded elements are defined to be constituents that correspond neither to the primary topic nor to part of the potential focus domain. These assumptions explain for instance the fact that the possibility for manner of saying verbs to behave as bridging verbs is context sensitive.

(2) (a) What did she say that he had given her?
(b) ??What did she purr that he had given her?

Utterances like (b) that sound odd out of the blue are much more acceptable in a reprise context.

In the same way, Hofmeister & Sag (2010) add that a construction with island can be countered by performance parameters. An example with a complex filler (and therefore heavy and referential) (3b) is read faster than the corresponding one with simple filler (3a). And where processing difficulty increases, acceptability decreases. In any case, the information is retrieved, with a comparable accuracy.

(3) (a) I saw *who* Emma doubted a report that we had captured in the nationwide FBI manhunt.
(b) I saw *which convict* Emma doubted a report that we had captured in the nationwide FBI manhunt.

The frequency of the bridge verb could also be interpreted as a parameter increasing the degree of acceptability of an example. As we can see in Figure 1, verbs that appear as bridge verb are among the most frequent completive verbs.

Even though we agree that these pragmatics-based constraints are important, we found an even more complex situation when we tried to check them against data from written and spoken French.

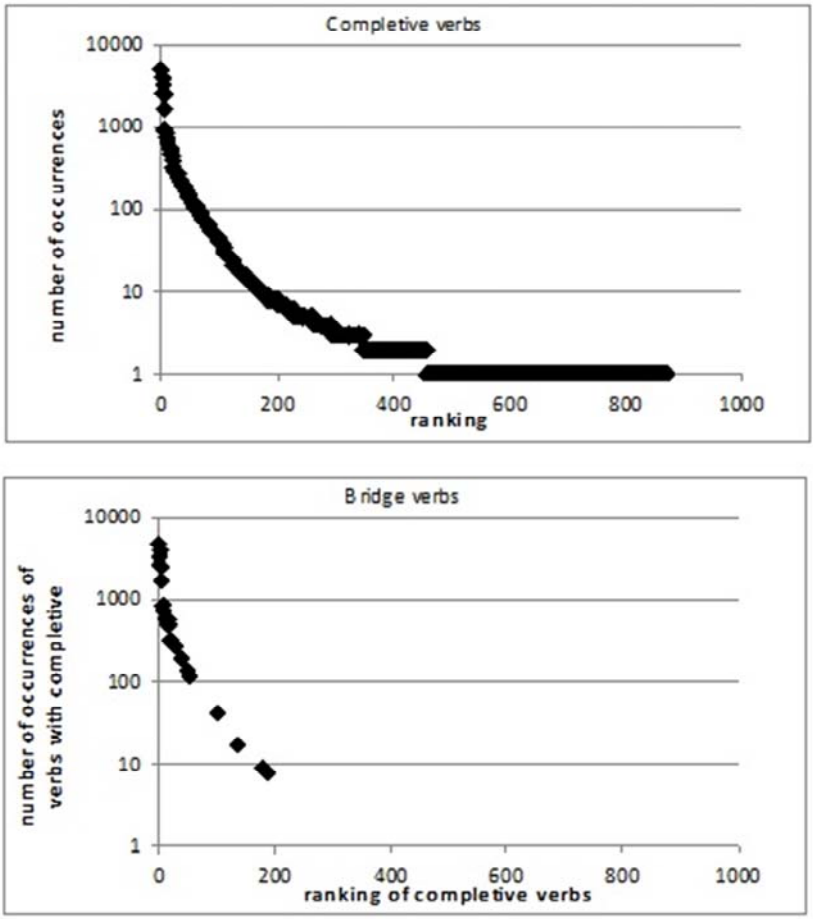

Figure 1. Completive verbs occurring as bridge verb in our corpora

1.2 Corpus-based approach

1.2.1 *The data*

We detected 229 occurrences of constructions involving long distance dependency in our 12M words corpus (Bérard 2012). As the main part of the corpus was neither tagged nor parsed, we could only use queries based on regular expressions. The queries were built using lexical words of the construction as boundaries: the list of the interrogative and relative forms on the one hand (*qui, que, qu', qu'est-ce que, qu'est-ce qu', quoi, dont, où, comment, combien, pourquoi, quand, lequel, laquelle, lesquels, lesquelles, quel, quelle, quels, quelles*

'what, who, whose, where, how, how many, why, when, which') and the conjunction forms on the other hand (*que, qui, qu', si, s'* 'that, if/whether'). Two to five words were allowed in between. We checked on a sample that a span enlarged to 10 words does not trigger more relevant forms. Then, we manually sorted the results of the requests, that is more than 21,000 results.

Through a short experiment we have estimated the benefits of using annotated corpora. First, we compared some spoken corpora with the tagged spoken corpus TCOF (100,000 words). The queries on the tagged corpus used the same lexical boundaries as mentioned above (to avoid the frequent tagging errors with these words), and added a verb in between. They returned 161 results. For 100,000 words, the queries on the non-tagged corpus returned 355 results. A research on a tagged corpus divides the data to be manually sorted approximately by two.

Table 1. Number of results of the query in spoken corpora, tagged or not.

	Non-tagged corpus (CorpAix, CRFP, Nancy, C-Oral-Rom, CTFP, CFPP, PFC)	**Tagged corpus** (TCOF)
Corpus size	2 700 000 words => 100 000 words	100 000 words
Results of the queries	9393 => 355	161

Second, we compared the written corpus CERF with the parsed written corpus French Treebank. The queries on the parsed corpus used parts of-speech (verb, relative pronoun, interrogative adjective and interrogative pronoun, conjunction), forms (conjunction *que* 'that' and *si* 'if') and dependencies (the verb is the governor of a relative pronoun and a conjunction). As we had no available tools to compile this complex query, we wrote a Python program[2]. It returned 69 results while we obtained 468 results for 360000 words with the CERF. A research on a parsed corpus divides the data to be manually sorted approximately by seven.

[2] Of course, there exists tools to query treebanks (Lezius & König 2000; Mírovský 2006; Zeldes *et al.* 2009), but converting our corpus in the specific format of such tools is more complicated than to develop a specific script.

Table 2. Number of results of the query in written corpora, parsed or not.

	Non-tagged non-parsed corpus (CERF)	**Tagged and parsed corpus (French Treebank)**
Corpus size	9 000 000 words => 360 000 words	360 000 words
	11 722 results => 468 results	69 results

The 229 sorted-occurrences are established as follow:

Table 3. Long distance dependencies in written and spoken texts.

	Written (/9M => /3M)	**Spoken (/3M)**
Number of occurrences	104 => 35	125

Table 4. Classification of long distance dependencies by host constructions.

Host construction	**Interrogative clause**	**Relative clause**	**Cleft or pseudo-cleft**
Number of occurrences	108	95	26

According to our corpus, long distance dependencies are more than three times more frequent in spoken French than in written French.

1.2.1 *Analysis of the data*

We focus our analysis on the bridge verb and its dependents. The most frequent syntactic pattern of the bridge is "subject + verb + *que*" ('that') without any other additional term. Here is a typical example of the construction:

(4) Bien sûr le ciné (du moins, celui dont *je présume que* tu parles), c'est facile (CERF)
'For sure the cinema (at least the one of which *I presume that* you are speaking), it is easy '

Between the interrogative/relative pronoun and its governor, we find only the items required by the syntactic well formedness conditions, that is in French the subject (*je* 'I'), the verb (*presume* 'presume') and the complementizer (*que* 'that'). This pattern represents 96.1% of the occurrences. The syntactic structure of the bridge verb is highly constrained. The objects are rare (seven occurrences) and they only take the form of clitic pronouns, as in (5). In the whole set of examples no adjuncts are found.

(5) Ils interprètent leur sentiment comme de la colère de l'euphorie ou comme rien du tout, en fonction de ce qu'on *leur* a dit qu'ils ressentiraient (CERF)
'They consider their feelings as anger euphoria or as nothing at all according to what they have said *to them* that they would feel'

The other modifications are negation: three occurrences (6). Besides, only four occurrences of recursive structures were found, limited to two verbs (7):

(6) Ne fais pas aux autres ce que tu *ne voudrais pas* qu'on te fasse (CERF)
'Do not do to others what you *do not want* they do to you.'

(7) En Grande Bretagne où il *croit savoir* que les contrôles sont moins nombreux […] (CERF)
'In Great Britain where he *believes to_know* that checkings are less numerous […].'

Some adverbs are present (six occurrences), but they are not syntactically integrated (8): they cannot be questioned nor clefted. Semantically they intensify the meaning of the bridge verb or express an attitude of the speaker, without real descriptive value (*exactement* 'exactly', *bien* 'well', *réellement* 'really', *peut-être* 'maybe').

(8) À moins qu'il y ait des raisons que je n'ai pas perçues et qu'il faudrait *réellement* qu'on m'explique. (CERF)
'Unless there are reasons that I have not perceived and that it is *really* necessary that one explains to me.'

The two remaining adverbs (*là* 'here', *ensuite* 'then') are not used as space and time adjuncts but as discourse particles with an argumentative function.

(9) Quand vous allez chez le coiffeur, vous espérez tous que ce sera un professionnel qui vous coupera les cheveux... pourquoi voudriez vous *ensuite* que ce soit des citoyens, la plupart du temps désinvestis politiquement, qui décident de l'avenir d'une nation ? ! ; -) (CERF, Forum)
'When you go to the hairdresser, you all hope it will be a professional who will cut hair... why should you *in addition* expect that citizens, mostly politically divested, decide the future of a nation? ! ; -)'

The syntactic arguments are not only rare, they also have particularities. As we said, objects are only clitic pronouns (5). Subjects are quasi exclusively clitic pronouns too (95.6%), as illustrated in (10).

(10) Et pourquoi *tu* penses qu'ils ont pas voulu (PFC)
'And why do *you* think that they didn't want (that)'

The content of the bridge is not at all descriptive: it does not refer to the world but to the interaction frame and its protagonists. The subjects are either devoid of content (impersonal) or refer to the source of an attitude. The objects are pronouns conveying interpersonal relations (*it seems to me, tell somebody something*) (5) and, as said before, adverbs are intensifiers or attitude markers. We did not find any nominal objects or adverbs of time, space or manner of doing (11a, opposite to 11b). The verb itself is limited to specific domains. Indeed, even by extending our research to the Web, we did not find for instance verbs describing the relationship between two facts as *faire* ('do') in (12), which sounds very unusual in a relative clause (11) as well as in interrogatives (12)[3]:

(11) (a) ??ce que j'ai *gentiment* demandé *hier à ma secrétaire* qu'on m'adresse
'what I *kindly* asked *my secretary yesterday* to be sent to me'
(b) ce que j'ai demandé qu'on m'adresse
'what I asked people to send me'

(12) ??vers où le tremblement de terre *a fait que* les bateaux ont dû partir
'where the earthquake *did that* boats had to leave for'

In our corpus, 34 types of verb are present but the most frequent are "basic" modals: *vouloir* 'want', *falloir* 'have to', *penser* 'think', *dire* 'say', and *savoir* 'know'. In interrogative clauses the first four represent 94.4% of occurrences; in relative clauses, the first three and the last one appear in more than a half of occurrences.

[3] The only example of this pattern was found in the French Treebank (Il en est de même dans l'hôtellerie et la restauration, pour 63% des 44000 emplois supplémentaires créés en dix ans et dans le commerce où *le succès des grandes surfaces a fait que* tous les emplois créés, en net, sont à temps partiel. 'It is the same in hotel management and food service for 63% of 44,000 additional jobs created in ten years and in trade where *the success of supermarkets does that* all jobs created, in fact, are part-time.'). The syntactic context can be analyzed as an instance of *supplementary relative clauses* (Hudleston *et al.* 1999: 1064). In such a context the wh-word could be analyzed as a fronted and not *extracted* adjunct, so that long distance dependency is no more at issue.

As a whole, the attested verbs can be taken as conveying a modal attitude. That is why we found a large lexical variation of bridge verbs belonging to this semantic domain by extending our research to the Web and testing 161 different verbs. In this extended corpus even factive verbs are attested as bridge verbs.

(13) m'enfin pour l'image, il s'est pas foulé, c'est clairement dérivé de la Mégane coupé telle qu'elle était prévue par Renault il y'a quelques années avec ces deux ouvertures sous les optiques. (Ce que *je regrette qu'*ils n'aient pas fait au final sur Mégane et Laguna)
'It is clearly derived from the Mégane Coupé as it was prescribed by Renault a few years ago with these two openings in the optical. (Which *I regret that* they have not done finally for Mégane and Laguna)'

(14) Là encore il y a des bois c'est ainsi que *je me figure qu'*était la Provence il y a mille ans.
'Again there are woods. This is how *I imagine that* was the Provence a thousand years ago.'

Finally, it is worth pointing out that the string "bridge verb + *que* + main verb" undergoes very few discourse and performance phenomena. Beside the rare cases of disjunct adverbs not syntactically integrated, like *réellement* 'really' in (8), we noticed only one occurrence of a discourse particle (*ben* 'well') and three cases of disfluencies. To sum up, the string involved in long distance dependency shows a strong syntactic and prosodic cohesion.

2. A linguistic hypothesis based on the corpus-driven description

The facts pointed out in the last section can receive a linguistic interpretation within a constructional framework. The severe limitations on the surface string "wh-word Subject1 V1 que Subject2 V2" undermine the traditional analysis of the string as a free unbounded embedding of S structures. This will result in a considerable overgeneration of strings compared to the attested ones. To limit this structural overgeneration in order to get the observed facts, many heterogeneous constraints would be needed: syntactic, lexical and performance constraints. In particular, the restrictions on the projection possibilities of this supposed S (for instance: no standard adjunct allowed for the bridge verb) would require an explanation probably based on processing factors not easy to check. This allows us to consider another possibility: instead of constraining a free underlying structure we could start from the existing forms and try to directly

model them. Following this approach, the limited string "wh-word Subject1 V1 que Subject2 V2" can be decomposed into two specific constructions. That is, two conventional associations of form and meaning.

The benefit will be to dispense with overgeneration at a structural level and to obtain the observed extensions, such as disjunct adverbs, by allowing insertion of material in the basic string at the level of discourse production, operation which is independently needed to account for the presence of a disjunct adverb in a PP in (15):

(15) Beaucoup de travail *pour franchement peu de résultats*.
'Many efforts *for frankly few results*.'

We call the first construction *verbal chunk*. Its form consists of the bridge verb + its subject + QUE. The meaning associated with this whole form can be phrased as "non descriptive meaning" to capture the fact that the chunk must convey a modal attitude (epistemic, deontic, evidential, appraisal). This meaning excludes from the chunk, as observed before, verbs the meaning of which is not compatible with this non-descriptive meaning (*faire que* 'do that', *entraîner que* 'entail that', *signifier que* 'mean that'). Subjects with semantic content of bridge verbs (excluding impersonal) are not descriptions of an agent or experiencer. They best signal the source of an attitude towards a state of facts. Also excluded are verbs of "manner of saying" which precisely describe the manner in which an act of saying is accomplished. This makes any possible evidential interpretation of the phrase difficult to obtain. The bridge chunk can further combine with a finite verb to form a *verbal chain*.

The resulting construction can be viewed as a *complex verbal licensor* consisting of a set of concatened (not embedded) verbs with little intervening material. The verbal string with complementizer is a member of a family of verbal strings constructions, the most widespread of them combining modal verbs with one or more infinitives, as in: *à qui doit être capable de répondre Pierre* 'to whom Pierre must be able to answer', where the licensor *répondre* 'answer' is syntactically concatenated with two verbs (*doit* 'must' and *être capable* 'be able') acting as modifiers at a semantic level.

Positing *verbal chunk* and *verbal chain* as new analytic concepts can be independently motivated. This pattern can be straightforwardly applied to the description of the behaviour of *weak verbs* as found in utterances such as: *je pense qu'il va s'améliorer* 'I think that he will improve' *versus* the *strong* interpretation found in *vous pensez à juste titre que la fin du monde est proche* 'you rightly think that the end of the world is close'. An analysis as a syntactic chunk di-

rectly captures the constraints on the *weak* form investigated in Blanche-Benveniste & Willems (2007)[4].

Furthermore, under the name of *verbal nucleus*, Kahane (1997) shows that a unit such as *verbal chain* can be used elegantly to formulate the rule placing the compound negation *ne…personne* 'not…anybody' in the following pair:

(16) Pierre est très jaloux de sa femme. Il *ne* supporte qu'elle parle à *personne*.
'Pierre is very jealous of his wife. He does*n't* stand that she speaks to *anyone*.'

(16') #Pierre est très jaloux de sa femme. Il supporte qu'elle *ne* parle à *personne*.
'#Pierre is very jealous of his wife. He tolerates that she does*n't* speak to *anyone*.'

The possible verb clusters in gapping contexts seem also to obey Bridge Chunk constraints:

(17) Marie *veut qu'on appelle* la police et Pierre _ les pompiers.
'Mary *wants that one calls* the police and Peter _ the firemen'

All these independent evidence in favour of our analysis should obviously be checked on a corpus of authentic data.

3. Modelling the phenomena

We have seen some motivations for considering the *verbal chain* in the description of extraction and some other phenomena (gapping coordination, negation…). We propose to go beyond the descriptive generalisations and to consider the verbal chain as a syntactic unit that occupies a syntactic position in the syntactic structure. For this, we will be working on the theoretical framework of dependency-based grammar, where the syntactic structure is basically a dependency tree, that is, a tree-like graph of relations between words (Tesnière 1959; Mel'čuk 1988). A syntactic position is characterized by the set of dependencies it can realize with other positions. We are going to extend such a representation

[4] A verb may have a week form if it can be constructed with a completive, as an interpolated clause verb and alone in an answer. Semantically, it mitigates the assertion of the sentence. Some examples of this verbs are *je pense* 'I think', *j'imagine* 'I imagine' and *j'espère* 'I hope'.

by considering that another unit than a word can occupy a syntactic position in the dependency structure. We call such a unit a *nucleus* following Tesnière (1959: 44) and more precisely a *verbal nucleus* when the nucleus behaves like a simple verb. The verbal nucleus can occupy a syntactic position exactly like a simple verb. The consequences on the description of long distance dependencies are immediate: only a phrase governed by the main verb of a clause can be extracted, where a verb can be a simple verb as well as a verbal nucleus (Kahane 1997, 2001, Kahane & Mel'čuk 1999).

To take some examples:

(18) (a) tu veux que j'en fasse quelque chose
'You want me to do something with it'
(b) Que veux-tu que j'en fasse ?
'What do you want me to do with it?'

(19) (a) Tu connais la personne qui a fait ça
'You know the person who did it'
(b) *Que connais-tu la personne qui a fait ?
'*What do you know the person who did?'

(20) (a) Elle est passée pendant que je faisais ça
'She came when I did it'
(b) *Qu'est-elle passée pendant que je faisais ?
'*What did she came when I did?'

In (18), *veux* → *que* → *fasse* 'want → to → do' is a verbal chain: *fasse* depends on (is subordinated to, is governed by) *veux* and this dependency is marked by the conjunction *que*. Such a verbal chain is a verbal nucleus. This verbal nucleus is the head of the clause: therefore, the position occupied by *quelque chose* 'something' can be extracted and for instance interrogated as in (18b).

Examples (19) and (20) illustrate the so-called modifier island (Ross 1967): the relative clause *qui a fait ça* 'who did it' or the circumstancial phrase *pendant que je faisais ça* 'when I did it' are modifiers and a position cannot be extracted out of a modifier. In our terms, these positions cannot be extracted because they are not dependent of a main verbal nucleus. In other terms, the chain of dependencies from the head of the clause to the extracted position is not a verbal chain: the first one, *connais* → personne → *qui* → *a* → *fait* 'know → *person* → who → did', contains a noun and the second one, *est* → *passé* → pendant que → *faisais* 'came → *when* → did', contains a conjunction. And both contain a modifier dependency. Such chains are not verbal nuclei. Our modelling can be compared

with the functional uncertainty of LFG (Kaplan & Zaenen 1989), where the constraints on extraction are verified on the chain of functional relations. But we go further by considering this chain as a syntactic unit strictly speaking.

Considering verbal nuclei as syntactic units in the same way as simple verbs gives us a simple description of extraction. There are no longer long distance dependencies, extraction is only local: only a direct dependent of the main verbal nucleus can be extracted. The complexity resides in the description of verbal nuclei: which strings of words can form a verbal nucleus? Before answering this question (which has been already addressed in previous sections), it is worth mentioning some advantages to considering verbal nuclei as proper syntactic unit.

First, as said before, the verbal nucleus appears in other phenomena, like negation (16) or gapping coordination (17). In consequence, describing the possible nuclei is also a contribution to the description of these phenomena.

Second, this modelling has a certain explicative value. It is well known that relative pronouns (*le type à* qui *tu veux que je cause* 'the guy to *whom* you want I speak') or interrogative pronouns in indirect question (*je comprends pas à* qui *tu veux que je cause* 'I don't understand to *whom* you want I speak') play a double role: they both mark the subordination of the clause they introduce and they fill a syntactic position inside the clause (*qui* fills the indirect object slot of *cause* 'speak' in the previous examples). The subordinate role explains the fronted position of the extracted group. This point has been extensively argued by Tesnière (1959: 561) (see also Kahane 2002) and is also constitutive of the X-bar analysis with COMP (Chomsky 1977). There is therefore a double dependency between the extracted group and the rest of the clause: the extracted group both governs the clause (as complementizer) and is governed by the clause (as filler). As the two roles of the extracted group are realized by one word, we can argue that the two syntactic relations between the extracted group and the rest of the clause must be realized by one dependency. This is why the main verb of the clause must also be the governor of the extracted group: they must occupy one position in the syntactic structure and be part of a same unit occupying this position. This unit is our verbal nucleus.

Third, if introducing complex nodes like nuclei make the structure more complicated: the structure is no longer a tree, but a *bubble tree* (Kahane 1997) the relation between the dependency structure and the linear order is simplified: the structure becomes projective. Let us compare the traditional dependency structure of (18b) (Figure 2) with its representation when the verbal nucleus is considered (Figure 3).

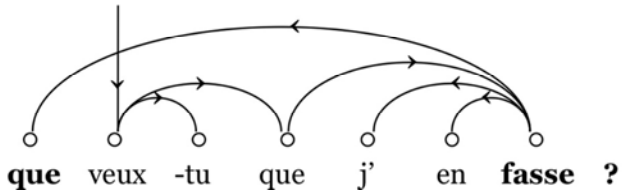

Figure 2. Dependency tree of (18b)

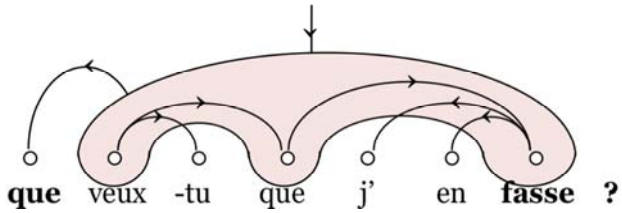

Figure 3. Bubble tree of (18b)

The structure in Figure 2 is non-projective: the dependency between *fasse* 'do' and the fronted *que* 'what' covers elements like *veux* 'want' that dominates *fasse* 'do'. In other words, the projection of *fasse* (*que … j'en fasse*) is a discontinuous phrase and the dependency between *fasse* and *que* cross the root dependency on *veux*. As soon as the dependency between *fasse* and *que* is attributed to the verbal nucleus, non-projective configurations disappear (Figure 3): *que* 'what' is placed directly on the left of the verbal nucleus and no dependency cross another one. This has an immediate consequence on word order: order rule remains local. In other words, the extracted group is placed in relation to its immediate governor and not in relation of one of the ancestor of its governor as in the traditional analysis (where *que* is placed beyond the governor *veux* of its governor *fasse*).

We can now come back to the definition of the verbal nucleus. If we take in consideration the descriptive generalisations found in section 2, we see that they are naturally modelled through the toll of verbal nucleus. The verbal nucleus models in a dependency-based framework the descriptive unit of verbal chain. Contrary to a phrase structure-based model in which any instance of the pattern subject + verb must be considered as a sentence, so that "chunks" revealed by our data show up as arbitrarily reduced sentences, this model can express a direct dependency from verb to verb. The only additional stipulation that we must make is that the verbs entering a verbal nucleus meet the constraints observed by the verbal chunks. That is, as for syntactic structure, limiting the possible de-

pendencies to the subject relation and, as for semantic or lexical properties, excluding the verbs conveying a descriptive meaning.

Note that the definition of verbal nuclei is language-specific. For instance, English contrasts with French by allowing preposition stranding:

(21) (a) the girl (that) I spoke to
(b) *la fille que je parlais à

In terms of nucleus, it means that the governed preposition (*to* and *à* in our examples) can be part of the verbal nucleus in English but not in French. This is probably explainable by the fact that English has phrasal verbs, that is, particles inside the verbal nucleus, and by the fact that particles and prepositions are similar in form and position[5].

A last point must be discussed: Is the verbal nucleus just a chain of verbs or does it aggregate other elements? In other words, is the verbal nucleus of example (18) the verbal chain *veux → que → fasse* 'want to do' or the "aggregate" *veux-tu que j'en fasse* 'you want me to do with it' including clitics?

The extraction is, all things being equal, much easier if the verbal chain forms a quasi-continuous unit. This suggests that the constraints do not apply only on the verbal chain proper (that is the elements belonging to the chain of dependency between the main verb and the extracted position), but concern the whole projection of this verbal chain. The more compact this projection is, the easier the extraction is. The best situation is when the projection of the verbal chain contains only clitics and other grammatical elements.

If we divide the corpus in genres, roughly opposing spontaneous speech and elaborated writing, we see that the projection of the verbal chain has a different pattern according to the genres. In spoken French, it is very rare that the verbal chain in an extraction is interrupted by non-clitic words (9/125 occurrences) and such examples mainly belong to instances of professional speech:

(22) c'étaient des gens qui avaient les moyens d'élever euh sept ou huit enfants et qui avaient des appartements dans lequel *il faut dire que ces enfants s'entassaient* un peu (CFPP)
'these were people who had the means to raise uh seven or eight children and who had apartments in which *it must be said that these children were a little crowded*'

[5] Conversely, pied-piping (the fact that the preposition is extracted with the wh-word it governs and placed in front of the clause) can be explained by the formation of *nominal nuclei* (this point will not be developed here, see Kahane 1997).

(23) tout ça s'intègre dans un contexte où *nous avons vu que euh aussi bien les frais d'inscription que les frais de sécurité sociale que le prix de ticket de restaurant universitaire augmentent bien plus vite que le montant des bourses qui sont alloués* euh aux (Nancy)
'it fits in a context where *we saw the registration fees, social security charges as well as university restaurant coupon prices increase much faster than the amount of scholarships that are allocated* uh to.'

On the other hand, in written genres, lexical subjects of the second verb are much more frequent (34.6%).

We can posit that spontaneous speech reveals a basic pattern with a subject clitic and the finite verb, which are straightforwardly modelled into a verbal nucleus unit. The written styles further elaborate this basic pattern with more lexical material in the subject slot in a way that could result in excluding the full-fledged subject from the verbal nucleus

4. Conclusion

The 229 apparent long distance dependencies detected in our French corpus of 12M words have a specific template: a bridge Chunk "clitic subject + verb + *que*", associated with a specific modal meaning. The bridge and the main verb compose a verbal chain, that is one sentence with a complex predicate instead of an embedding of sentences. This complex predicate – that is the two verbs and *que* (and maybe the two clitic subjects) – can be modelled as one node in a bubble tree. The apparent extracted group is now locally dependent to an immediate governor.

Our corpus-based study on a limited grammatical structure belonging to *core syntax* allows us to draw useful conclusions concerning both the problem of the data and the relation between data and analysis.

As for the data we can point out a balance of a positive contribution and issues to be deepened. Our 12M words corpus of authentic data provided us with a sufficient basis for capturing the most salient features and properties of a spontaneous use of a syntactic pattern. In particular, we could start from the forms actually produced by the speakers and writers, avoiding the overgeneration that generally implies the construction of utterances by the linguist on the basis of a well-studied phenomenon. Having in mind available utterances like *à qui tu as dit qu'il fallait s'adresser* 'to whom did you say that it must be spoken' and of the quite plausible *tu as dit hier avec beaucoup de conviction qu'il fallait s'adresser à Pierre* 'you said yesterday with much conviction that it must be

spoken to Pierre', the linguist will be prompted to consider by extension as acceptable: *à qui tu as dit hier avec beaucoup de conviction qu'il fallait s'adresser* 'to whom did you say yesterday with much conviction that it must be spoken'. Nevertheless the last pattern is totally absent of the authentic data. We can conclude that the artificial manipulation of the data can prevent us from capturing an important descriptive generalisation: the limited possibilities of syntactic projection of the pattern in the context of long distance dependencies.

On the other hand we noticed that we used cases of negative evidence in our argumentation as crucial examples for establishing the lexical limits of the pattern. Obviously the corpus is of no use in such cases. To guarantee the relevance of this kind of negative evidence, it should be necessary to add to the corpus some experimental device to get the proper acceptability judgments. This will be considered in a further step.

From a theoretical point of view, we propose to deal with the descriptive generalisations by positing new types of syntactic unit: Modal chunk and Verbal Nucleus as a type of complex governor. This amounts to generalizing to predicates the idea that a syntactic slot can be filled by a multiword unit. It is interesting that this linguistic hypothesis can easily be modelled in the formalism of an extended version of dependency grammar.

We hope to have shown that a link can be established between the various steps involved in linguistic analysis: observing authentic data, establishing descriptive generalisations and finally phrasing them in a theoretical framework.

References

Ambridge, B. & Goldberg, A. 2008. The island status of clausal complements: Evidence in favor of an information structure explanation. *Cognitive Linguistics* 19(3): 349–81.

Antoine, J.Y., Letellier-Zarshenas, S., Nicolas, P., Schadle, I. & Caelen, J. 2002. Corpus OTG et ECOLE MASSY: vers la constitution d'une collection de corpus francophones de dialogue oral diffusés librement. In *Actes de la 9ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'2002)*. Nancy: Association pour le Traitement Automatique des Langues, 319-324. Available online at http://www.atala.org/taln_archives/TALN/TALN-2002/taln-2002-poster-001 (accessed October 15, 2014).

Auchlin, A., Avanzi, M., Goldman, J.P. & Simon, A.C. 2012. *C-PROM corpus libre de parole multigenre*. Available online at http://sites.google.com/site/corpusprom/ (accessed October 15, 2014).

Bérard, L. 2012. Dépendances à distance en français contemporain. Etude sur corpus. PhD diss., Université de Lorraine.

Blanche-Benveniste, C., Rouget, C. & Sabio, F. (eds) 2002. *Choix de Textes du Français Parlé : trente-six extraits*. Paris: Champion.

Blanche-Benveniste, C. & Willems, D. 2007. Un nouveau regard sur les verbes faibles. *Bulletin de la Société Linguistique de Paris* 102(1): 217-54.

Branca-Rosoff, S., Fleury, S., Lefeuvre, F. & Pires, M. 2012. Discours sur la ville. Corpus de Français Parlé Parisien des années 2000 (CFPP2000). Available online at http://cfpp2000.univ-paris3.fr/CFPP2000.pdf (accessed October 15, 2014).

Chomsky, N. 1977. On wh-movement. In P.W. Culicover, Th. Wasow & A. Akmajian (eds), *Formal Syntax*. New York: Academic, 71-132.

Corpus Accueil UBS. Available online at http://www.info.univ-tours.fr/~antoine/parole_publique/Accueil_UBS/index.html (accessed October 15, 2014).

Corpus Brassens. Available online at http://www.info.univ-tours.fr/~antoine/parole_publique/Brassens/index.html (accessed October 15, 2014).

Corpus Evolutif de Référence du Français. Université de Provence (10 million words).

Corpus IRIT. Available online at http://www.irit.fr/ACTIVITES/LILaC/Pers/Prevot/Corpus.html (accessed October 15, 2014).

Corpus Oral-Nancy. Université de Lorraine (360,000 words).

Corpus TCOF (Traitement des Corpus Oraux en Français). Available online at http://www.cnrtl.fr/corpus/tcof/ (accessed October 15, 2014).

Equipe DELIC 2004. Présentation du Corpus de Référence du Français Parlé. *Recherches sur le français parlé* 18: 11-42.

Deulofeu, H.J. & Blanche-Benveniste, C. 2006. C-Oral-Rom: French Corpus. In Y. Kawaguchi, S. Zaima & T. Takagaki (ed.), *Spoken Language Corpus and Linguistic Informatics*. Amsterdam: Benjamins, 181-198. Corpus distributed by European Language Distribution Agency (ELDA).

Durand, J., Laks, B. & Lyche, C. 2005. Un corpus numérisé pour la phonologie du français. In G. Williams (ed.), *La linguistique de corpus*. Rennes: Presses Universitaires de Rennes, 205-217.

Erteschik-Shir, N. 1997. *The dynamics of Focus Structure*. Cambridge: Cambridge University Press.

Kahane, S. 1997. Bubble trees and syntactic representations. In T. Becker & H.U. Krieger (eds), *Proc. 5th Meeting of the Mathematics of Language* (MOL5), Saarbrücken: DFKI, 70-76.

Kahane, S. 2001. A fully lexicalized grammar for French based on Meaning-Text theory. In A. Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing*. Berlin/Heidelberg: Springer-Verlag, 18-31.

Kahane, S. 2002. A propos de la position syntaxique des mots qu-. In P. Le Goffic (ed.), *Interrogation, indéfinition, subordination*. *Verbum* XXIV(4): 399-435.

Kahane, S. & Mel'čuk, I. 1999. La synthèse sémantique ou la correspondance entre graphes sémantiques et arbres syntaxiques – Le cas des phrases à extraction en français contemporain. *TAL* 40(2): 25-85.

Kaplan, R.M. & Zaenen, A. 1989. Long-distance dependencies, constituent structure, and functional uncertainty. In M. Baltin & A. Kroch (eds.), *Alternative Conceptions of Phrase Structure*. Chicago: University of Chicago Press, 17-42. Reprinted in *Formal Issues in Lexical-Functional Grammar* 47: 137-65.

Lezius, W. & König, E. 2000. Towards a search engine for syntactically annotated corpora. In W. Zühlke & E.G. Schukat-Talamazzini (eds), *KONVENS 2000 / Sprachkommunikation, Vorträge der gemeinsamen Veranstaltung 5. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS), 6. ITG-Fachtagung "Sprachkommunikation"*. Berlin: VDE-Verlag, 113-116.

Mel'čuk, I. 1988. *Dependency Syntax: Theory and Practice*. Albany (NY): SUNY Press.

Mírovský, J. 2006. Netgraph: a tool for searching in Prague Dependency Treebank 2.0. In J. Hajič & J. Nivre (eds), *Proceedings of the TLT 2006*. Prague: Institute of Formal and Applied Linguistics, 211–222.

Ross, J. 1967. Constraints on variables in syntax. PhD diss., Massachusetts Institute of Technology.

Tesnière, L. 1959. *Éléments de syntaxe structurale*. Paris: Klincksieck.

Verhagen, A. 2005. *Constructions of Intersubjectivity*. Oxford: Oxford University Press.

Zeldes, A., Lüdeling, A., Ritz, J. & Chiarcos, C. 2009. ANNIS: A search tool for multi-layer annotated corpora. In M. Mahlberg, V. González-Díaz & C. Smith, *Proceedings of the Corpus Linguistics Conference – CL2009*. Published by UCREL, Article #337(a), 1-23. Available online at http://ucrel.lancs.ac.uk/publications/cl2009/ (accessed October 15, 2014).