Chapter 3 *Syntactic annotation of the Rhapsodie corpus: An overview*

Sylvain Kahane, Paola Pietrandrea

Abstract. This chapter presents the principles underlying the syntactic annotation and in particular the reasons for separating this annotation into two levels, micro and macrosyntax, presented in Chapters 4 and 6. The particular focus on paradigmatic lists (coordination, lists, reformulation, disfluencies, etc., see Chapter 5) is justified, as well as the integration of hard-to-describe elements such as disfluencies or discourse markers at various level of the syntactic structure. Our decision to discard the notion of the sentence and replace it by the notions of Illocutionary Units [IUs] and Government Units [GUs] is discussed.

1. Annotating three mechanisms of cohesion and two levels of analysis

The syntactic annotation of the *Rhapsodie* corpus was guided by the theoretical argument that not only competence but also performance phenomena undergo regular constraints that should be made explicit and formalized in an analysis of spoken language. Such a theoretical underpinning led us to develop an empirical, inductive scheme for the annotation of syntax.

In line with the general principles that inspired the annotation schemes of the Rhapsodie corpus, we elaborated an annotation scheme that can be regarded empirical and inductive. Our scheme is "empirical" because we annotated the entirety of the data in the corpus, without neglecting any segment whatsoever (disfluencies, reformulations, corrections, discourse markers). It is "inductive" because the set of relevant syntactic structures occurring in spoken languages was identified through a data-driven incremental strategy of annotation.

Beside a classical morphosyntactic analysis (segmentation into words, lemmatization, tagging in parts of speech), we could recognize and annotate in our corpus three mechanisms of cohesion that seem to define the syntactic structures of spoken French:

1. Government relations (Tesnière 1959, Mel'čuk 1988).

2. List phenomena, that is, phenomena syntactically characterized by multiple realizations of one and the same structural position. These phenomena include coordination, as well as particular cases of reformulation, correction, and disfluency (Blanche-Benveniste 1990, Gerdes & Kahane 2009, Bonvino et al. 2008, Masini & Pietrandrea 2010, Kahane & Pietrandrea 2012, Masini et al. 2012).

3. Macrosyntactic relations, that is, phenomena of syntactic constraints operating between the constituents of the utterance structure due to the illocutionary dependency of some constituents on others (Blanche-Benveniste 1990, Berrendonner 1990, Cresti 2000).

The two former mechanisms belong to what is traditionally called in French literature on syntax "microsyntax", whereas the latter defines a further level of linguistic analysis traditionally called "macrosyntax" (Blanche-Benveniste 1990, Berrendonner 1990, Cresti 2000).

1.1 Microsyntax: Government and listing

1.1.1 Government

Government describes, quite classically, the constraints that a word can operate over a constituent by determining its occurrence, its categorical nature, its position as well as its markers. For example in the utterance (1), the verb *fais* 'do' governs *je* 'I', *des petits boulots* ‘odd jobs’ and *en plus* ‘too’: it constraints the occurrence of the pronominal subject *je* in the nominative case to its left and of the nominal direct object *des petits boulots* to its right.

(1) *je fais des petits boulots en plus* [Rhap-D0006, CFPP2000]

‘I do odd jobs too’

We represent government relations as dependencies between the governor and the head of the governed constituent. This is done recursively and the constituent *des petits boulots* is analyzed with the noun *boulots* ‘jobs’ as its head governing the determiner *des* (indefinite plural article) and the adjective *petits* lit. ‘small’ (see Figure 1).

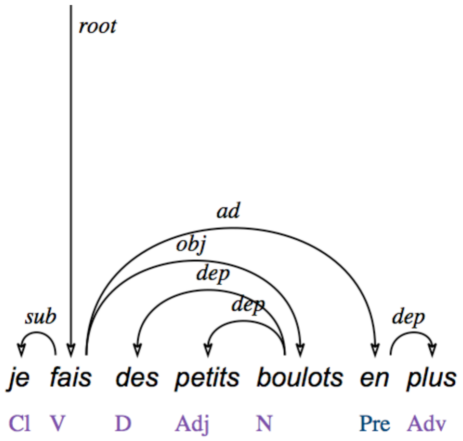

Figure 1. Microsyntactic analysis of example (1)

As shown in several works (Lecerf 1961, Kahane & Mazziotta 2015), dependency-based and constituency-based approaches (that is, approaches describing the dependency relations holding between constituents rather than words) are more or less equivalent. The government relations described in our annotation are, therefore, similar to the hierarchical relations captured by X-bar trees (Jackendoff 1977). It must nevertheless be noted that adopting a dependency rather than a constituency-based approach allowed us to annotate our corpus without resorting to the notion of the sentence, which, as will be shown below, is particularly unsuitable for the analysis of spoken language (Section 2.3). Our annotation task did not consist, indeed, in a top- down identification of the constituents of a pre-segmented sentence, rather we could identify through a bottom-up strategy the government relations linking each word of our corpus to the others. Such an approach allowed for a description of the syntactic relations holding between words even when syntactically connected segments did not belong to "well-formed sentences" but to fragmentary or interrupted structures (see Chapter 16, Section 4.10, for a discussion).

1.1.2. Listing

Listing is a mechanism of cohesion orthogonal to government that guarantees the cohesion between the conjuncts of a coordination construction (2) as well as the cohesion between the words involved in a lexical reformulation (3) or a disfluency construction (4). According to Blanche-Benveniste (1990), the relation between the coordinated (or reformulated) elements of a list cannot be satisfactorily described in terms of government, but is better explained by taking into account the fact that these elements realize one and the same structural position. For example, in (2), the conjuncts *vos photos* 'your pictures', *vos vidéos* 'your videos' and *un agenda* 'an organizer' realize the direct object of the verb *partager* 'share'; the conjuncts *vos parents* 'your parents', *vos cousins* 'your cousins', *vos tontons* 'your uncles' realize the complement of the preposition *avec* 'avec'; in (3) the conjuncts *ziki point com* 'ziki dot com', *ziki* 'ziki' and *un site communautaire dont le but est de vous aider à créer une identité numérique* 'a community site whose goal is to help you to create a numerical identity' realize the object of the verb *citer 'to* quote'; in (4) the three disfluent repetitions can also be analyzed as conjuncts filling the same syntactic position.

(2) *Famibook est une sorte de Facebook cent pour cent français réservé aux familles pour partager* ***vos photos vos vidéos ou encore un un agenda*** *avec* ***vos parents vos cousins ou vos tontons*** [Rhap-M2005, Rhapsodie-Broadcast]

‘Famibook is a kind of hundred percent French Facebook reserved to families to share **your pictures your videos or also an an organizer** with **your parents your cousins or your uncles**’

(3) *on peut citer* ***ziki point com Z I K I un site communautaire dont le but est de vous aider à créer une identité numérique*** [Rhap-M2005, Rhapsodie-Broadcast]

‘we can quote **ziki dot com Z I K I a community site whose goal is to help you to create a numerical identity**’

(4) *et euh* ***il faut il faut*** *apprendre* ***à à*** *s~ je dirais* ***se se*** *faire à ces variations* [Rhap-M2002, Rhapsodie-Broadcast]

‘and um **you need you need** learn **to to** g~ I’d say **get get** used to these variations’

As we will show in Chapter 5, the recognizing of listing as an independent cohesion device allowed us to minimize a great number of difficulties raised by the syntactic annotation of coordinations, reformulations, and disfluencies.

1.2 Macrosyntax

Macrosyntax describes the relations holding between a number of syntactic constructions typical of spoken language and particularly frequent in spoken French - such as paratactic structures, detachments, discourse markers - and the rest of the production. The idea behind macrosyntax is that any utterance can be analyzed in a constituent, the nucleus, which could be uttered alone and a number of optional peripheral constituents, which, though not necessarily governed at the microsyntactic level, are not autonomous and undergo distributional constraints due to their illocutionary dependency on the nucleus.

As an example, in (5) it is possible to recognize a potentially autonomous constituent, *on n’a pas de lycée* ‘we don’t have any high school’, which is thus the nucleus of the macrosyntactic structure, two dependent constituents, *nous* ‘us’ et *dans le quartier* ‘in the neighborhood’ located on its left, and one constituent located on its right, *déjà* ‘first’:

(5) *nous dans le quartier* ***on n’a on n’a pas de lycée*** *déjà* [Rhap-D0004, CFPP2000]

‘us in the neighborhood **we don’t we don’t have any high school** first’

We will argue in Chapter 6 that the syntactic cohesion of sequences such as (5) is ensured by the fact that they realize one and only one illocutionary act, whose nature is determined by the nucleus. In the example (5) the entire sequence constitutes indeed an answer to the question reported in (6) and the illocutionary nature of the sequence is determined by the nucleus *on n’a pas de lycée*. As we can see in (7), this constituent is, indeed, the only constituent that can commutate with (5) as an answer to the question in (6). The other constituents could not be uttered autonomously as an answer in this context. It is, indeed, the presence of the nucleus *on a pas de lycée*, the only autonomous constituent that allows for their occurrence.

(6) *mais euh du coup vous pouvez pas travailler par exemple avec les lycées ?* [Rhap-D0004, CFPP2000]

‘but um given this situation couldn’t you work for example with high schools?’

(7) A : *mais euh du coup vous pouvez pas travailler par exemple avec les lycées ?*

B: ***on n’a on n’a pas de lycée*** ‘we don’t have any high school’

*B: *nous* ‘we’

*B: *dans le quartier* ‘in the area’

*B: *déjà* ‘first’

1.3 Government Units and Illocutionary Units

By acknowledging two mechanisms of syntactic cohesion, microsyntax and macrosyntax, we could recognize two types of syntactic units in the organization of spoken French: (i) Government Units (GUs), whose cohesion is ensured by government and listing, and (ii) Illocutionary Units (IUs), whose cohesion is ensured by macrosyntactic relations.

It should be noted that the distinction between GUs and IUs resemble, at least to some extent, to Biber et al.'s (1999: Chapter 14) distinction between C-units ("clause" units) and T-units ("text" units). Biber et al. claim that a level of proper syntactic organization can be indeed identified beyond C-units, that is, beyond sequences whose cohesion is ensured by government proper. Such a level, composed of what Biber et al. call T-units has been widely explored in the literature with respect to its pragmatic properties. Biber et al.'s characterization is instead focused on the syntactic aspects, namely the distributional properties, of T-units.

Like Biber et al., we recognize a double level of syntactic organization and we focus on the formal rather than the functional aspects of this level of linguistic organization, let us say that we focus on the signifiers rather than the signifieds of the upper level units. Two differences distinguish though our approach from Biber et al.'s. First, we include list phenomena among the syntactic phenomena characterizing the clause level - our GUs. Secondly, and most importantly, the annotation task led us to revise Biber et al.'s idea that upper-level units are composed of clauses: we could observe indeed that the syntactic organization governing GUs can extend over the borders of macrosyntactic components, IUs and speech-turns. Let us illustrate this important point.

When starting our annotation task, we were convinced like Biber et al. (1999), but also Blanche-Benveniste et al. (1990), that spoken discourse is made up of GUs whose internal cohesion is ensured by government relations and which are connected to one another through macrosyntactic relations. We had in mind, in other words, a highly idealized and somehow simplistic model of the interaction between government and macrosyntax according to which macrosyntactic relations should take as an input the output of government relations. Such a syntactic configuration characterizes indeed most IUs of our corpus: an example is in (8), where we can observe four GUs *enfin* 'so', *mon métier* 'my work', *donc* 'then', and *c'est de l'électronique* 'it is in the electronics' connected to one another through macrosyntactic relations.

(8) *enfin mon métier donc c'est l'électronique* [Rhaps D2005, Lacheret]

'so my job then it is in the electronics'

It is also entirely possible, however, that microsyntax extends beyond the borders of macrosyntactic components, as in (9).

(9) *et* ***dans la foulée de ce sommet social à vingt heures ce soir*** *une intervention du chef de l'État à la télévision* [Rhap-M2006, Rhapsodie-Broadcast]

'and following this social meeting at eight P.M. a speech of the Head of State on TV'

In (9), as in (5), it is possible to distinguish a central unit, the nucleus, that can be uttered alone, *une intervention du chef de l'État à la télévision* 'a speech of the Head of State on TV' and two optional units, *dans la foulée de ce sommet social* 'following this social meeting' and *à vingt heures ce soir* **'**at eight P.M.', whose presence is only legitimated by the presence of the nucleus. These three units are connected macrosyntactically since they present an illocutionary asymmetry that determines the illocutionary dependency of the latter from the former; but they are also microsyntactically connected since the head of the nominal constituent *une intervention du chef de l'Etat* governs the heads of the two prepositional constituents *dans la foulée de ce sommet* and *à vingt heures ce soir*.

In (9), as well as in a great number of IUs in our corpus (Chapter 16), one and the same GU crosses the borders of macrosyntactic components. We will also see that the microsyntactic relations holding between the words of a GU can extend beyond IUs and speech turn boundaries (Chapters 6 and 16).

All in all, the annotation task led us to characterize government relations, listing, macrosyntactic relations, and speech-turn organizations (as well as prosodic organization - Chapters 8) as different types of organization that operate simultaneously and independently on one another on the same stretch of discourse.

The data-driven model that we could elaborate during the annotation task is therefore radically modular, that is, composed of independent and parallel modules of linguistic organization. This distinguishes our approach from similar models proposed by Biber et al. (1999), Blanche-Benveniste et al. (1990), and Cresti (2000) and it makes our approach closer to modular approaches to discourse organization, such as Roulet's (2001).

2. The advantages of annotating a complex structure

In this section, we want to highlight three important consequences of our choice to provide an annotation for different mechanisms of syntactic cohesion: (i) the uniqueness of the Rhapsodie treebank, which can be considered one of the most complete syntactic treebanks for spoken language; (ii) the integration of hard-to-describe elements such as disfluencies or discourse markers at some level of the syntactic structure; (iii) the possibility of providing a rich syntactic annotation without resorting to the highly problematic notion of the sentence.

2.1 A rich syntactic treebank

The schema developed for the syntactic annotation of the Rhapsodie treebank completes in a sense previous schemes of syntactic annotation.

Existing treebanks for spoken language provide either a more or less rich annotation of microsyntactic phenomena or an annotation of macrosyntactic phenomena. The British component of the International Corpus of English (Nelson et al. 2002), for example, as well as the treebanks of English, German, and Japanese, created within the VERBMOBIL project (Hinrichs et al. 2000) include a microsyntactic annotation for POSs and constituents. The Ester treebank for spoken French (Cerisara et al. 2010), the CNG (Spoken Dutch Corpus; Schuurman et al. 2004), and the Spoken Italian Corpus AN.ANA.S. MT (Cutugno et al. 2004) include a microsyntactic annotation for POSs and dependencies. The Switchboard corpus (Meteer et al. 1995, Taylor et al. 2003, Calhoun et al. 2010), the CHRISTINE treebank (Sampson 2000) and the Venice Italian treebanks (Delmonte et al. 2007) provide a microsyntactic annotation of both constituency and dependency phenomena. Finally, the C-ORAL-ROM treebank of spoken Italian, French, Spanish and Portuguese (Cresti and Moneglia 2005) provides a complete annotation of macrosyntactic phenomena.

The Rhapsodie treebank is thus the first existing treebank that provides both a complete microsyntactic and macrosyntactic annotation and an extensive annotation of list phenomena. The macrosyntactic annotation has now been partly integrated in some new developments of dependency treebanks, as the Universal Dependencies project (Nivre et al. 2016; see Gerdes & Kahane 2017 for a comparison between Rhapsodie's and UD's annotations of macrosyntax).

2.2 Discourse markers, reformulations and disfluencies as part of the syntactic structure of spoken language

Discourse markers, corrections, reformulations, disfluencies are extremely pervasive in spoken discourse. These elements, traditionally considered external to the "sentence" structure, are analyzed, in Rhapsodie, as syntactically integrated with the rest of the production via macrosyntactic or listing relations.

Discourse markers can be analyzed as a particular type of macrosyntactic component: as shown by example (10), the discourse markers *eh bien* 'well'*, alors* 'then'*, hein* 'you know' are linked to the nucleus *je vais à pieds* 'I walk' by an illocutionary dependency, that is, their presence is licensed by the presence of the nucleus.[1] They can be analyzed, therefore, as well as the detached pronoun *moi* 'me', as elements linked to the nucleus not at the government level but at the macrosyntactic level.

(10) ***eh bien alors*** *moi je vais à pieds* ***hein*** [Rhap-D2001, CFPP2000]

'**well then** me I walk **you know**'

In Chapter 6 we will provide a classification of discourse markers based on their macrosyntactic distributional properties.

As far as reformulations are concerned, we could analyze them as particular types of lists. Let us take for example (11).

(11) *et j'avais* ***une circonscription un un rayon d'action*** *d'à peu près euh cent kilomètres tout autour de cet endroit* [Rhap-D2004, Lacheret]

'and I had **an area a a mission range** of about um one hundred kilometers all around that place'

As we will argue in Chapter 5, the two conjuncts *une circonscription* 'an area' and *un rayon d'action* 'a mission range' realize the direct object position governed by the verb *j'avais* 'I had'. They are therefore analyzable as a case of list, which allows for regarding them as

[1] Discourse markers are highly language-specific and their translation can only be very rough.

integrated to the microsyntactic structure of the sequence. We will see in Chapter 5 that corrections and other dialogical moves display a syntactic cohesion with the rest of the utterance, often through list relations.

The syntactic relation holding between the different components of a (syntactic) disfluency phenomenon can also be described as a case of listing. Let us consider for example the disfluent repetition of the determiner *un* 'a' in (12):.

(12) ***un un*** *rayon d'action*

'**a a** mission range'

Levelt (1983) would analyze this typical case of disfluency as composed of a *reparandum* (the first occurrence of *un*) and a *reparatum* (the second occurrence): it is not hard to show that both the *reparandum* and the *reparatum* can be regarded as realizing one and the same syntactic position: the determiner position for the noun rayon 'range'. They are in other words in a list relation with each other. We systematically annotated the disfluencies made up of a *reparandum* and a *reparatum* that realize the same syntactic position as a case of list. Treating this type of disfluencies as list phenomena allowed us to integrate them within the microsyntactic analysis of our texts.

It goes without saying that such a choice, theoretical in nature, raises some practical questions concerning the automatic parsing of the corpus. Automatic parsers, especially those available for French in 2008 - when the Rhapsodie project started - are conceived for the analysis of written language; they are not suitable therefore for parsing spoken language especially when disfluencies are incorporated to the text to be analyzed. As shown by Gilquin and De Cock (2013), indeed, a quite common solution for this problem consists in adapting spoken texts to automatic parsers, by deleting disfluencies, rather than adapting automatic parsers to spoken texts, with some notable exceptions such the SOUP parser for the analysis of spontaneous speech (Gavaldà 2004). Honnibal and Johnson (2014) have shown that they obtain better results with a disfluency detection joint to parsing. We think that even better results could be obtained if the disfluencies were analyzed as part of the syntactic structure: the "disfluency detection" would become part of the syntactic parsing. Such an approach raised a number of practical questions. We will describe them, as well as the solutions we found for them in Chapter 7.

The choice to include disfluencies in the annotation schemes of spoken corpora, is, we think, crucial for both theoretical and practical reasons: from a theoretical point of view, the encoding of disfluencies may reveal basic aspects of utterance planning and dialogic interaction (Ginzburg et al. 2014), from a practical point of view the annotation of disfluencies provides essential data for the training of parsers of spontaneous speech.

2.3 A syntactic annotation without sentences

As mentioned above, our annotation choices allowed us to provide an annotation of spoken discourse without resorting to the highly problematic notion of the sentence. Let us see why we decided to discard the notion of the sentence for our annotation and how we could replace such a notion with more appropriate and consistent maximal units.

In the grammatical tradition, sentences have been regarded as undisputed units forming the "maximal syntactic units" of language. Nevertheless, several linguists have suggested that the sentence cannot be considered a fully adequate notion, especially when applied to the description of spoken data (Berrendonner 1990, Miller & Weinert 1998, Kleiber 2003, Blanche-Benveniste 2002, Cresti 2005, Pietrandrea 2014 et al., among others). As Berrendonner (1990) puts it:

> “Traditional sentences, since they are nothing but informal and intuitive graphic approximations of linguistic units, are commonly considered as inefficient grammatical tools when it comes to segmenting a spoken text or even to analyzing, in written discourse, relations beyond syntactic government in written data […].” (Berrendonner 1990: 25, our translation)

Miller & Weinert (1998: 30) add:

> “The central problem is that it is far from evident that the language system of spoken English has sentences, for the simple reason that text-sentences are hard to locate in spoken texts.”

As shown by Pietrandrea et al. (2014), indeed, the notion of the sentence is commonly conceived as a linear sequence characterized simultaneously by semantic, syntactic and prosodic unity and contiguity. Such an idea is based on an over-idealized conception of linguistic cohesion, which posits that semantic, syntactic and prosodic units should necessarily be coextensive with each other. As shown by Sabio (2006) and Pietrandrea et al. (2014) among others, data drawn from textual corpora clearly indicate that such a strict coincidence between these three kinds of cohesion mechanisms is indeed possible but by no means necessary, especially in spoken speech. For example, let us examine the following sequences:

(13) *ils étaient tout à fait normaux* [Rhap-D0002, CFPP2000]

‘they were absolutely normal’

(14) *alors là la psychiatrie c’est autre chose* [Rhap-D0006, CFPP2000]

‘then now psychiatry that’s something else’

(15) *ça c'est le problème de Paris ... je pense* [Rhap-D0004, CFPP2000]

'that that's the problem of Paris … I think'

(16) *c'est un Chinois … très riche* [Rhap-D2001, Mertens]

'he is a Chinese man … very rich'

While it is possible to recognize a semantic, a microsyntactic and a prosodic cohesion in (13), the utterance in (14) is characterized by a semantic and a prosodic cohesion, but it does not display any microsyntactic cohesion (there is no government between all the words of the sequence); the utterance in (15) is characterized by semantic cohesion, but it does not display microsyntactic or prosodic cohesion (there is no government relations between the words of the sequence and a pause before *je pense* 'I think' breaks the prosodic unit of the sequence); the utterance in (16) is semantically and syntactically cohesive, but it does not display prosodic cohesion: a long pause intercuts between the words *Chinois* 'Chinese man' and *très riche* 'very rich' (see Chapter 6 for an analysis of this utterance as two successive illocutionary units). In spite of the fact that all these sequences seem to show a certain degree of cohesion which constraints their realization, the notion of the sentence as is commonly conceived could only be applied to the first sequence.

This state of affair makes the notion of the sentence as a weak candidate for the status of reference unit for the syntactic annotation of spoken texts. Still, syntactic annotations need a reference unit. A text needs to be segmented in syntactically relevant units to be further analyzed and annotated internally. How did spoken language syntacticians try to deal with this problem?

The most common solution, which is in our view more a workaround than a solution, has been to reproduce in the syntactic annotation of spoken corpora what is usually done for the syntactic annotation of written corpora. The syntactic annotation of written corpora has traditionally relied on the punctuation of texts to identify its maximal units: the span of text comprised between two full stops is taken as a reference unit. Following a poorly discussed tradition, this unit, which is graphical in nature, is called a "sentence" and it is analyzed and annotated as a truly syntactic unit.

In the same vein, most projects of annotation of spoken corpora, especially the less recent ones, have been based on a pre-segmentation of the transcriptions in "sentences". The annotators of the Christine Corpus, for example (Sampson 2004) worked on texts transcribed and pre-segmented in orthographic sentences by the transcribers of the British National Corpus. The annotators of the Prague Dependency treebank annotated a "reconstructed" text, that is, a spoken text that was transposed - through a normalization of the word transcription, of the word order and through an insertion of punctuation - to the form of a written text (Bejček et al. 2013).

More recently, the creators of treebanks have raised a debate about the nature of the maximal units of annotation without arriving, though, to a precise definition. The annotators of the Corpus AN.ANA.S._MT (Cutugno et al. 2004) did not rely on pre-segmented texts, rather they annotated what they call "clauses" and "sentences" as the maximal units of their constituency-based annotation. Unfortunately the notions of the clause and the sentence are not explicitly defined in the presentation of their treebank (although we can consider that their clauses are similar to our GUs). The annotators of the "treebank du français parlé", who worked on the Ester Corpus (Gravier et al. 2004), decided to pre-segment the transcription of the Ester corpus, which was mainly based on prosodic criteria, by segmenting their transcriptions in sentences, that is, in sequences necessarily (and somehow arbitrarily) included within the boundary of the speech turn and defined on the basis of not better specified "syntactic, prosodic and discursive" criteria (Abeillé & Crabbé 2013).

All in all, in spite of the fact that the need for a discussion about the nature of the maximal units of spoken syntax finally begins to be felt, the attachment to the poorly defined and quite confusing notion of the sentence seems unlikely to die.

The structure of discourse that we could identify through the annotation task is better described adopting a modular approach to discourse structure (see among others Nolke & Adam 2000 and Roulet 2001), which provides several levels of description of discourse cohesion without attempting to reduce this straightforward complexity within the narrow limits of the notion of the sentence.

3. Conclusion

The choice to annotate three mechanisms of cohesion and two levels of syntactic structure allowed us to develop a richly annotated syntactic treebank of spoken French in which, on the one hand, phenomena typical of spoken discourse such as disfluencies, reformulations, corrections, discourse markers are all accounted for and, on the other hand, all the cohesion mechanisms at play in spoken discourse are taken into account to describe the rich syntactic structure characterizing spoken language.

Our microsyntactic annotation combines government relation and listing and is presented in two separated chapters: government is encoded by dependency trees which are presented in Chapter 4, while the list representation leads us to substantial complications of the microsyntactic structure which are explained in Chapter 5. The macrosyntactic structure is explained in Chapter 6, along with the mismatches between micro- and macrosyntax.

We must underline that macrosyntax and lists have been annotated manually while the microsyntactic annotation has been computed automatically and corrected manually. Indeed, at the time we conducted our annotation, no parser for spoken French was available, still some relatively good tools for written French did already exist. We decided by consequence to use a state-of-art parser for written French (de la Clergerie 2005b) and to give to it as inputs segments that were similar to written sentences. This is not just a segmentation of the transcription but a set of multiple sentences in which lists have been unfolded, each sentence containing just one layer. The strategy we adopted, as well as the tools we have used, are described in Chapter 7.