Chapter 6. ***Macrosyntactic annotation***

Paola Pietrandrea, Sylvain Kahane

Abstract. This chapter describes the macrosyntactic annotation of the Rhapsodie corpus, from the linguistic heritage to Rhapsodie's own theoretical approach to macrosyntax. Macrosyntactic phenomena, such as dislocation, discourse markers, inserts, or parenthesis, are quite frequent in spoken French. The major unit of this level is the illocutionary unit, the main component of which is the nucleus, bearing the illocutionary force. Three kinds of peripheral components are considered: adnuclei, openers, and associated nuclei, including discourse markers. The different types of relationships between illocutionary units – contiguity, hierarchy, parallelism, and bifurcation – are discussed, as well as the interplay between macro- and microsyntactic units. We show in particular that microsyntactic relations can go beyond macrosyntactic boundaries and even speech turns.

1. Introduction

In Chapter 3, we explained why we decided to endow the Rhapsodie treebank with two levels of syntactic annotation: microsyntax, that is, syntactic cohesion ensured by government (presented in Chapters 4 and 5) and macrosyntax, that is, syntactic cohesion ensured by illocutionary dependency. In the current chapter we will present the details of the macrosyntactic annotation.

In order to clarify what we mean by macrosyntax, we will first describe the different linguistic traditions from which we inherited the notion of macrosyntax (Section 2). As we will see, these traditions characterize macrosyntax in slightly different terms. This situation led us to explicitly posit our own theoretical approach to macrosyntax before attempting the annotation task. This theoretical definition is described in Section 3. In Section 4, we will define and characterize the macrosyntactic units that were annotated in the Rhapsodie treebank, namely the Illocutionary Units (IUs, 4.1) as well as their components: Ad-Nuclei (4.2), Illocutionary Unit Openers (4.3), and Associated Nuclei (4.4). We will describe the annotation of the interaction between IUs in Section 5

and the annotation of the interaction between macrosyntactic and microsyntactic units in Section 6. In Section 7, the annotation procedure will be presented and, in Section 8, some general conclusions will be drawn.

1.1. Macrosyntactic traditions

The distinction between micro- and macrosyntax was first proposed by Blanche-Benveniste *et al.* (1990), Berrendonner (1990), and Cresti (2000) (but see also Andersen & Nølke (2002) for an overview). These studies put forward macrosyntax as a level of linguistic description capable of accounting for a number of cohesion mechanisms that are particularly frequent in spontaneous spoken language – especially in spoken French – which cannot be simply regarded as microsyntactic government phenomena, such as, for example, the "paratactic" construction in (1).

(1) [*ceux qui sont en location*] [*la moyenne*] [*c'est environ trois ans*] [Rhap-D0004, CFPP2000]

[those who are on a lease] [the average] [it's about three years]

'Those who are on a lease stay three years on average'

While the different macrosyntactic models acknowledge that sequences such as (1) have to be considered as forming a cohesive unit at some level of linguistic description, they diverge slightly as far as the characterization of the nature of this cohesion is concerned. According to the Aix-en-Provence macrosyntactic model (Blanche-Benveniste *et al.* 1990) a sequence such as (1) constitutes a 'macrosyntactic unit', that is, a succession of distinct Government Units (GUs; see Chapter 4) whose cohesion is guaranteed by a loosely defined notion of *togetherness* based on Bolinger (1968), which basically has to do with discursive autonomy cohesion (Blanche-Benveniste 1990: 114). According to the Fribourg model (Berrendonner 1990, 2002), the sequence in (1) constitutes a 'period' made up of three syntactic clauses (units corresponding to our Government Units) marked by a conclusive 'intoneme' at the end of the third clause (Berrendonner 2002). According to the Florence model (Cresti 2000), the sequence in (1) constitutes an 'utterance', that is, a sequence of

prosodic units whose cohesion is ensured by the fact that the entire sequence conveys one and only one illocutionary act, in this case an assertion.

To sum up, these macrosyntactic models characterize some major linguistic units that go beyond government proper. These units are usually described in the literature from a pragmatic perspective that focuses on their illocutionary or rhetorical values. Macrosyntax, instead, focuses on the span and the form of macrosyntactic units, using syntactic and distributional criteria (such as suppressions, insertions, commutations) to identify and delimit them. For all the macrosyntactic models, the main identifying criterion of a macrosyntactic unit is the possibility that this unit has to constitute an autonomous utterance.

The different models, though, present slight differences as far as the description of the components of macrosyntactic units are concerned. The Aix model proposes that macrosyntactic units are composed of several government units bound together by "discursive cohesion"; the Fribourg school proposes, instead, that it is intonational cohesion that binds together several GUs in discourse; the Florence model proposes that illocutionary cohesion may bind together several intonation units (rather than GUs) in discourse.

1.2. Rhapsodie's approach to macrosyntax

In *Rhapsodie,* we chose not to rely on prosodic criteria to define macrosyntactic units. Since our practical objective was to create a corpus that allows us to study the interface between prosody and syntax, we needed to clearly separate these two levels of analysis. Therefore we did not follow the prosodic definition of macrosyntactic units proposed by Berrendonner (2002) who describes the maximal extension of a macrosyntactic unit in terms of the presence of a conclusive intoneme; nor could we strictly follow the Florence school's approach that characterizes macrosyntactic units as sequences of prosodic, rather than syntactic, units.

Rather, we consider that macrosyntax describes the whole set of relations holding between the microsyntactic units that make up one and only one illocutionary act, although, as we will see in Section 6, microsyntax can sometimes go beyond macrosyntactic units.

This definition combines the syntactic model proposed by the Aix model (Blanche-Benveniste et al. 1990), according to which the minimal units that compose a macrosyntactic unit are syntactic in nature, and the pragmatic model developed by the Florence model (Cresti 2000), according to which the maximal extension of a macrosyntactic unit is defined in terms of illocution.

Such a choice led us to call the maximal macrosyntactic units *Illocutionary Units* (henceforth IUs) and to provide, in our work, an account and an annotation for the syntactic rather than the prosodic units that compose an IU.

2. Rhapsodie's macrosyntactic annotation

Rhapsodie's annotation of macrosyntactic structures consisted of four tasks:

(i) segmenting each sample into IUs;

(ii) identifying and annotating the syntactic components of each IU;

(iii) identifying and annotating the linear relations existing between IUs;

(iv) characterizing the microsyntactic nature of the relations holding between the different components of an IU (when it applies).

We will describe tasks (i) and (ii) in this section. Tasks (iii) and (iv) will be described in Sections 3 and 4 respectively.

2. 1 Identifying the Illocutionary Units

As mentioned in Chapter 3, an IU can be defined as all the microsyntactic units that contribute to realizing one and only one assertion, injunction, interrogation, or exclamation (Gardiner 1932; Jespersen 1924).

In order to identify the IUs composing our samples, we grouped together the syntactic constituents that realized one and the same illocutionary act, whether they were microsyntactically linked to one another or not. Operationally, we considered that one or more syntactic constituents (whether microsyntactically linked or not) realize one and the same illocutionary act if they could be embedded under the scope of a saying verb that makes the illocutionary value of the entire sequence explicit. For example, sequence (1) – reproduced here as (2) – is composed of three syntactic constituents, microsyntactically independent of one another, in other words by three distinct GUs, but it constitutes only one IU realizing the illocutionary value of assertion. Such an interpretation is confirmed by the fact that the entire sequence can be embedded under the scope of the saying predicate *je dis* 'I say', which makes explicit the illocutionary value of assertion of sequence (3).[1]

(2) *ceux qui sont en location < la moyenne < c'est environ trois ans* // [Rhap-D0004, CFPP2000]

'those who are on a lease < the average < it's about three years //'

(3) *je dis :* [*ceux qui sont en location la moyenne c'est environ trois ans*] [Rhap-D0004, CFPP2000]

'I say: [those who are on a lease the average it's about three years]'

Similarly, the sequence in (4) is composed of three GUs; these three GUs constitute one IU that realizes one and the same question. Such an interpretation is confirmed by the test in (5).

1 We denote the right border of IUs with the symbol //. The other symbols, used to separate the macrosyntactic constituents of an IU, will be introduced in the following sections.

(4) ^*et au niveau de des des odeurs <+ est-ce que c'est efficace > également* ?//

^and as for the the the smells <+ is it effective > also ?//

'Is it also effective against smells?'

(5) *je te demande :* [*au niveau des odeurs est-ce que c'est efficace également ?*]

'I ask you: [as for smells is it also effective?]'

As mentioned in Chapter 3, each IU contains a *nucleus* (i.e., an autonomous constituent that bears the illocutionary force of the entire IU) plus a number of optional constituents, which are illocutionarily dependent on the nucleus. We have identified three kinds of peripheral constituents: ad-nuclei (i.e., pre-nuclei, in-nuclei and post-nuclei), associated nuclei and IU openers. In the following sections, we will examine the properties of each of these components in detail.

2.2. Nucleus and ad-nuclei

Let us examine the IUs represented in examples (6) and (7).

(6) *nous < dans le quartier <+ on n'a on n'a pas de lycée > déjà //* [Rhap-D0004, CFPP2000]

we < in the neighborhood <+ we don't we don't have any high schools > first //

'first of all we don't have any high schools in the area'

(7) ^*et ça < j'en garde (de toutes ces années) un excellent souvenir //* [Rhap-D0001, CFPP2000]

^and that < I have kept of that (of all those years) an excellent memory //

'and I have kept an excellent memory of all those years'

According to our analysis, the IU represented in (6) comprises a nucleus (*on n'a on n'a pas de lycée*), two pre-nuclei (*nous* and *dans le quartier*) and a post-nucleus (*déjà*) (see Figure 1), whereas

the IU represented in (7) consists of a pre-nucleus (*ça*), a nucleus (*j'en garde un excellent souvenir*) and an in-nucleus (*de toutes ces années*) (see Figure 2).[2]

Figure 1. Macrosyntactic structure of (6)

Figure 2. Macrosyntactic structure of (7)

Let us justify such an analysis, by explaining what nuclei and ad-nuclei (i.e., pre-nuclei, in-nuclei, and post-nuclei) are.

2.2.1. Nucleus

The nucleus of an IU corresponds to the constituent that bears the illocutionary force of the entire sequence, in other words the constituent that could be uttered alone in the same discursive position (and with the same illocutionary function) as the entire IU. To illustrate this, we reproduce here the discussion already presented in Chapter 3. As shown there, the IU in (6) constitutes an answer to the question reported in (8). As we can see in (9), the only constituent of (6) that can commutate with (6) as an answer to the question in (8) is *on n'a pas de lycée* 'we don't have any high schools'. The other constituents could not be uttered autonomously as an answer in this context.

(8) \$L1 *mais euh du coup vous pouvez pas travailler par exemple avec les lycées* ?

\$L2 *nous* < *dans le quartier* <+ ***on n'a on n'a pas de lycée*** > *déjà //* [Rhap-D0004, CFPP2000][3]

2 The right boundary of the pre-nucleus and the left boundary of the post-nucleus are signaled by the symbols < and >, respectively, and the in-nucleus is in parentheses.

3 \$L1 and \$L2 refer to the two speakers involved in this dialog.

‘$L1 but erm given this situation couldn’t you work for example with high schools?

$L2 we < in the neighborhood <+ we don’t we don’t have any high schools > first //’

(9) **a.** $L2 ***on n’a on n’a pas de lycée*** ‘we don’t have any high schools’

b. *$L2 *nous* ‘we’

c. *$L2 *dans le quartier* ‘in the area’

d. *$L2 *déjà* ‘first’

Similarly, the IU reproduced in (7) realizes, in the context of its occurrence, an assertion that “elaborates” a previous assertion, see example (10).[4]

(10) *on était trente-neuf quarante > nous "hein" "euh" //+ trente-neuf quarante // ^mais ceci dit "euh" < l'ambiance était bonne // ^et ça <* ***j'en garde*** *(de toutes ces années)* ***un excellent souvenir //***

‘we were thirty-nine forty people > us //+ thirty-nine forty people // ^but in spite of that "erm" < there was a nice atmosphere // ^and that < **I have kept of that** (of all those years) **an excellent memory** //’

As shown in the test in (11), the IU reproduced in (7), could only be replaced, in the same context, by the sequence *j’en garde un excellent souvenir* that we have analyzed as the nucleus of the IU; the non-nuclear constituents *ça* and *de ces années* cannot replace the entire sequence.

(11) **a.** *on était trente-neuf quarante > nous "hein" "euh" //+ trente-neuf quarante // ^mais ceci dit "euh" < l'ambiance était bonne //* ***j'en garde un excellent souvenir*** *//*

4 We borrow the notion of “elaboration” from, among others, Segmented Discourse Representation Theory (Asher and Lascarides 2003).

‘we were thirty-nine forty people, thirty-nine //+ forty people // but in spite of that, there was a nice atmosphere // I have kept an excellent memory of that //’

b. **on était trente-neuf quarante > nous "hein" "euh" //+ trente-neuf quarante // ^mais ceci dit "euh" < l'ambiance était bonne // ^et ça //*

‘we were thirty-nine forty people, thirty-nine //+ forty people/ / but in spite of that, there was a nice atmosphere // and that //’

c. **on était trente-neuf quarante > nous "hein" "euh" //+ trente-neuf quarante // ^mais ceci dit "euh" < l'ambiance était bonne // de ces années //*

‘we were thirty-nine forty people, thirty-nine //+ forty people// but in spite of that, there was a nice atmosphere // of these years’

Examples (6) and (7) present two verb-headed nuclei, but nuclei can be also realized by non verbal constituents as in (12), (13), and (14) (see also the second IU in (10)).

(12) $L1 *"bon" maintenant <+ ça se passe mieux // ça va //* $L2 ***ouais //***

‘$L1 "well" now <+ it’s better // it’s OK // $L2 yep //’

(13) $L1 *"ben" je vous remercie beaucoup de votre attention //* $L2 *"ben"* ***de rien //***

‘$L1 "well" I thank you very much for your attention // $L2 "well" you’re welcome //’

(14) ***une autre question*** *> peut-être //*

‘another question > maybe //’

It should be clear indeed that the nuclei are characterized by their function of sequences realizing an illocutionary act that contributes to the evolution of the common ground shared by the interlocutors, rather than by their internal syntactic constituency.

Due to their properties, nuclei show some distributional properties that are not equally shared by other macrosyntactic constituents: (i) they are endowed with an illocutionary force that can be made explicit by embedding the nucleus within the scope of a performative predicate (15); (ii) they can constitute an autonomous speech turn (16); (ii) they can enter the scope of an illocutionary adverb, such as *frankly* (17); (iii) they can commutate with other nuclei having the same locutionary content but a different illocutionary force (18); (iv) their locutionary content can be freely modified (19). Let us examine these properties with example (4).

(15) *je te demande : [est-ce que c'est efficace ?]*

'I ask you: [is it effective?]'

(16) *est-ce que c'est efficace ?*

'is it effective?'

(17) ***franchement,*** *est-ce que c'est efficace ?*

'**frankly**, is it effective?'

(18) **a.** *et au niveau des odeurs* ***c'est efficace*** *également*

'and as for smells, it is also effective'

b. *et au niveau des odeurs* ***qu'est ce que c'est efficace*** *également* ***!***

'and as for smells, how effective it is as well!'

c. *et au niveau des odeurs est-ce que c'est efficace* ***toute la journée*** *également ?*

'and as for smells is it also effective all the day?'

We will see in the following sections that non-nuclear constituents do not show these distributional properties.

Every IU has a nucleus. The nucleus, as well as the other macrosyntactic constituents, is generally a GU, as in examples (2) or (7). But sometimes, as in (4), (6), or (10), it is only a part of a GU,

because some other part forms a separate macrosyntactic constituent in the same IU (see 4.2.2 below) or even another IU (Section 7).[5]

2.2.2. Ad-nuclei

The commutation tests presented in (15) through (18) show that IUs are composed of an autonomous illocutionary constituent, the nucleus, plus a number of constituents that are illocutionarily dependent on the nucleus. The very presence of these constituents is indeed licensed by the existence of the nuclear unit. We consider three kinds of such constituents: ad-nuclei, associated nuclei and IU openers. The *ad-nuclei* can be classified, according to their linear position as *pre-nuclei*, *in-nuclei* and *post-nuclei*.

Figure 3. The macrosyntactic structure of (4)

Unlike nuclei, ad-nuclei do not bear the illocutionary force of the IU. This entails that they do not display some of the distributional properties typical of the nucleus: (i) they cannot enter the scope of a performative predicate separately from the nucleus (19); (ii) they cannot constitute an autonomous speech turn (20); (ii) they cannot enter the scope of an illocutionary adverb, such as *frankly* (21); (iii) they cannot commutate with sequences having the same locutionary content but a different illocutionary force (22).[6] This can be clearly illustrated by taking as an example the pre-nucleus and the post-nucleus of the IU in (4), *au niveau des odeurs* and *également*.

5 We use the symbol + to indicate that the boundary between two macrosyntactic units is not a GU boundary. See <+ in (6) and //+ in (10).

6 The linear sequences *au niveau des odeurs* and *également* can indeed constitute an autonomous speech turn in some contexts, but this is not the case for the two sequences employed to realize the pre-nucleus and the post-nucleus of (4). These two sequences are associated to a particular prosodic profile that prevents them from occurring in isolation.

(19) **je te dis : au niveau de des des odeurs ; je te demande : est-ce que c'est efficace ? ; je te dis : également*

'*I tell you: as for the smells; I ask you: is it effective?; I tell you: also'

(20) a. **au niveau des odeurs* 'as for smells'

b. * *également* 'also'

(21) a. **franchement au niveau des odeurs* 'frankly speaking as for smells'

b. **franchement également* 'frankly speaking also'

(22) a **et* ***au niveau de des des odeurs !*** *est-ce que c'est efficace ? également*

b. * *et au niveau de des des odeurs est-ce que c'est efficace ?* ***également !***

Ad-nuclei can, by contrast, be freely modified. This is due to the fact that they have in most cases a descriptive rather than a procedural content.[7] We will see in the following sections that other constituents, such as IU openers or associated nuclei, do not share the same properties.

In most cases an ad-nucleus is not microsyntactically linked to the nucleus; it constitutes, in other words, an autonomous GU, as is exemplified by *moi* and *de toute façon* in (23).

(23) *"ah" "ben"* ***moi*** < ***de toute façon*** < *oui //* [Rhap-D0002, CFPP2000]

'"ah" "well" **me** < **in any case** < yes //'

The pre-nucleus can even be realized by a verb-headed GU, as shown in (24).

7 For the distinction between descriptive and procedural we refer to the distinction proposed within the framework of Relevance Theory (Blakemore 1987) between meanings that contribute to the propositional contents shared by interlocutors and meanings that provide instructions on how to interpret these propositional contents (this distinction is also reminiscent of the distinction proposed by Ducrot (1980) between descriptive and instructional meanings).

(24) ***je suis arrivée "euh" au Kenya*** *< je voulais travailler (d'abord) pour le gouvernement //* [Rhap-D2004, Lacheret]

'I arrived "erm" in Kenya < I wanted to work (first) for the government //'

In other cases, the ad-nucleus is microsyntactically governed by the nucleus.

(25) *^mais* ***à ce moment-là*** *<+ elle est vue par une dame //* [Rhap-M0018, Rhapsodie]

'^but at that moment <+ she was seen by a woman //'

(26) *^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]

'^and following this social affairs summit <+ at eight p.m. <+ a speech by the Head of State on TV //'

We consider that the prepositional phrases in (25) and (26) are not part of the nucleus essentially for topological reasons: these constituents do not follow the microsyntactic word order constraints of French. In French (as in English), a PP that is an obligatory constituent of a verbal nucleus must always be placed after the verb. This is the case in (27) when the focus PP of (25) is obligatory and therefore cannot be fronted.

(27) a. $L1 *quand est-elle vue par une dame ?*

$L2 *elle est vue par une dame* ***à ce moment-là***

'$L1 when was she seen by a woman?

$L2 she was seen by a woman **at that moment'**

b. *$L2 *à ce moment-là elle est vue par une dame*

'*$L2 at that moment she was seen by a woman'

The same topological argument holds for (26): in French (as in English), a PP depending on a noun must always be placed after the noun when it is part of the noun phrase, see (28).

(28) [*une intervention du chef de l'État à la télévision* ***dans la foulée de ce sommet social à vingt heures ce soir***]$_{NP}$

'[a speech by the Head of State on TV following this social affairs summit at eight p.m.]$_{NP}$'

The fact that the PPs of (26) do not follow the microsyntactic constraint of French indicates that they have been positioned following the macrosyntactic organization into nucleus and ad-nuclei and that they are macrosyntactic constituents.

We have so far separately examined the topological, syntactic and distributional properties of ad-nuclei. We will now describe some functional properties that characterize all types of pre-nuclei, in-nuclei and post-nuclei.

Pre-nuclei. The pre-nucleus can be defined as a macrosyntactic constituent, located on the left of the nucleus, that does not respond to the test of nuclearity and that can be eliminated without prejudice for the nuclearity status of the nucleus. The right border of this constituent is notated with the symbol <.

As we will see in Chapter 16, from a semantic point of view, prenuclei provide different types of contribution to the predication included within the nucleus: they can co-refer to an argument (subject, object, indirect object) of the nucleus (as *moi* 'me' in (30), *les chaises* 'the chairs' in (31), *de la crise aux Antilles* 'of the Antilles crises' in (32); they can realize a circumstantial (as *en ce qui me concerne* 'as for me' in (33); a speech-act modifier (34); a connective (35); an appellative (36); a phrase modifier (37); an hanging element (38).

(30) *moi < j'ai eu aucun problème scolaire pour mes enfants //* [Rhap-D0002, CFPP2000]

'me < I don't have any school problem for my children //'

(31) *les chaises < il faut me les donner //* [Rhap-D0009, PFC]

'the chairs < it is necessary to give them to me'

(32) *de la crise aux Antilles < il ne devrait pas en être question au sommet social de l'Élysée cet après-midi //* [Rhap-M2006, Rhapsodie]

of the Antilles crises < there should not be talk of at the social Summit at the Elysée this afternoon

'At the social Summit at the Elysée palace, this afternoon, there should not be talk of the Antilles crisis'

(33) *ensuite <+ vous allez "euh" jusqu' aux lignes de tram //* [Rhap-D0007, Avanzi]

'afterwards <+ you go till the tramway rails'

(34) *en ce qui me concerne < j'aimerais enseigner dans un établissement public //* [Rhap-M1003, Rhapsodie]

'as for me < I would like to teach in a public school'

(35) *donc < la vigilance sera accrue à ce moment-là //* [Rhap-D2008, Rhapsodie]

'by consequence < the vigilance will be increased at that moment //'

(36) *Françoise Giroud < vous occupez un poste d'observation que des gens haut placés vous envient //* [Rhap-D2001, Mertens]

'François Giroud < you hold an observation post that highest-ranking people envy to you'

(37) *{ plus fraternel | plus volontaire } < il aura les couleurs que nous lui donnerons //* [Rhap-M2004, Rhapsodie]

'{ more fraternal | more willful } < he will have the colours that we will give to him //'

(38) *il y a des gens < ils ont des emmerdes parce qu'ils se sont confiés en télé //* [Rhap-D2007, Rhapsodie]

there are people < they are in trouble because they confided on tv //

'some people are in trouble because they confided on tv'

In-nuclei. The in-nucleus is the syntactic constituent, located in the middle of the nucleus, that does not respond to the test of nuclearity and that can be eliminated without prejudice for the nuclearity status of the nucleus. We notate the in-nucleus between parentheses ( ). An example is given in (39).

(39) *vos journaux (Jean-Christophe) qui soulignent également la faiblesse de la mobilisation des électeurs >+ hier //* [Rhap-D2013, Rhapsodie]

‘your newspapers (Jean-Christophe) which also underline the poor voter turnout >+ yesterday //’

In-nuclei are to be distinguished from parentheses (Section 5), that is, IUs inserted parenthetically in other IUs. The main difference between in-nuclei and parentheses is that while the latter have an illocutionary value and respond therefore to the tests of nuclearity, the former are illocutionarily dependent on the nuclei in which they are inserted.

From a syntactic point of view, in-nuclei can be either realized by autonomous GUs as in (39) or by governed syntactic constituents as in (40). We use the symbol + to indicate that the in-nucleus belongs to the same GU as the nucleus it interrupts.

(40) *alors < le dégagement (+ maintenant) de Carrisou "euh" le dégagement argentin //* [Rhap-D2003, Rhapsodie]

‘so < the kick (+ now) by Carrisou "erm" the Argentine kick’

From a functional point of view, in-nuclei may realize coreferents of the subject (41), circumstantials (42), predicates (43), speech-act modifiers (44), connectives (45), and appellatives (46).

(41) *c'est une expérience (ça) que je n'ai jamais oubliée //* [Rhap-D2001, Mertens]

‘it is an experience (this one) that I never forgot //’

(42) *^et les Argentins (sous les ordres de Maradona) "euh" jouaient un peu plus haut //*

[Rhap-D2003, Rhapsodie]

‘^and the Argentineans (under the command of Maradona) "erm" played a bit higher //’

(43) *il faut s'appeler Rachida Dati (+ écrit aussi Libération Champagne) pour oser affirmer qu'il ne s'agit pas d'un vote sanction //* [Rhap-D2013, Rhapsodie]

‘one should be called Rachida Dati (+ writes also Libération Champagne) to dare claim that this is not a protest vote //’

(44) *on a pu se croiser (peut-être) à Censier //* [Rhap-D0001, CFPP2000]

‘we could meet (maybe) at Censier //’

(45) *^parce ^que ça donne un aspect (quand même) plus moderne si tu mets le tissu tout autour //* [Rhap-D0009, PFC]

‘^because it gives a look (indeed) more modern if you put the cloth all around //’

(46) *vous constatez (maître) comme moi //* [Rhap-D2011, Rhapsodie]

you can observe (Master) like me //

Post-nucleus. The post-nucleus is a syntactic constituent, located on the right of the nucleus, that does not respond to the test of nuclearity and that can be eliminated without prejudice for the nuclearity status of the nucleus. The left border of this constituent is notated with the symbol >. An example is given in (47).

(47) *ça a duré dix ans* > ***le silence autour de moi*** *//* [Rhap-D2010, Rhapsodie]

‘it lasted ten years > the silence around me //’

From a syntactic point of view, the post-nucleus, like any type of ad-nucleus can be realized by either an autonomous GU, as in (47), or a syntactically governed constituent, as in (48).

(48) *"oh la" il y a il y a une mauvaise séquence "hein" >+ pour l'équipe de France >+ actuellement //+ qui se fait là vraiment balader "hein" //* [Rhap-D2003, Rhapsodie]

‘"oh dear" there is there is a bad bit of play "isn’t there" >+ for the French team >+ currently //+ which is being walked over "isn’t it" //’

It is even possible for the post-nucleus to govern the nucleus as shown by Sabio (2006) who provides example (41).

(49) *treize euros* >+ ***ça coûte*** *//* (heard at the farmers' market)

‘thirteen euros >+ it costs //’

From a semantic point of view, post-nuclei can fulfil various functions: they can be occupied by coreferents of the subject of the nucleus (50), coreferents of the object of the nucleus (51), coreferents of the indirect object (52), circumstantials (53), speech-act modifiers (54), connectives (55), appellatives (56), but also predicates (57) and co-referents of the attribute of the nucleus (58).

(50) *ça a duré dix ans > le silence autour de moi //* [Rhap-D2010, Rhapsodie]

‘it lasted ten years > the silence around me //’

(51) *^et j'ai vraiment mal vécu > ça //* [Rhap-D1003, Rhapsodie]

‘and I really took it hard > that //’

(52) *qu'est-ce que vous en pensez > de la boule magique //* [Rhap-D2011, Rhapsodie]

‘what do you think of that > of the magic ball’

(53) *j'ai toujours envie de faire "euh" avancer le monde par mes idées > quitte à les mettre en oeuvre moi-même //* [Rhap-D2005, Lacheret]

‘I have always wanted "erm" to boost the world through my ideas > even if I had to realize them by myself’

(54) *^et "euh" cela désigne l'espèce { de | de } force qui mine l'institution littéraire > certainement //* [Rhap-D2009, Mertens]

‘^and "erm" this indicate the sort { of | of } force thatundermines the literary institution > of course //’

(55) *il ne le tire pas > d'ailleurs //* [Rhap-D2003, Rhapsodie]

'he does not shoot > by the way //'

(56) *vous constatez > maître //* [Rhap-D2011, Rhapsodie]

'can you observe > Master //'

(57) *[la France est rose //] >+ constate La Voix du Nord //* [Rhap-D2013, Rhapsodie]

'[France is rose //] >+ observes La Voix du Nord //'

(58) *la France sera { ce que nous voudrons qu'elle soit |} >+ {| une nation { unie | vivante | solidaire | ouverte } qui n'accepte aucune fatalité } //* [Rhap-M2004, Rhapsodie]

'France will be { what we want her to be |} >+ {| a Nation { united | lively | supportive | open } which does not accept anything as fate } //'

2.3 IU openers

In the annotation task, we decided to distinguish pre-nuclei from what we called *IU openers*. IU openers constitute a class of syntactic elements obligatorily located at the very beginning of an IU (and not merely on the left of the nucleus as is true for pre-nuclei). Examples of openers are the conjunctions *et* 'and', *mais* 'but', *parce que* 'because', *puis* 'then'.

Openers are characterized by the following properties:

- they are not microsyntactically dependent on any other word;
- they are obligatorily located at the beginning of an IU;
- they have a linking function that makes explicit the discursive relation holding between the IU they introduce and other IUs in the discourse.

In our annotation, openers are marked by the symbol ^, as illustrated in (59) and (60)

(59) ^***et*** *tu arrives à la fontaine "euh" place Notre Dame* // [Rhap-M0001, Avanzi]

‘^and you arrive at the fountain "erm" in Notre Dame square //’

(60) ^***mais*** *en fait < "euh" Charlot va dire (en fait) que c’est lui qui les "euh" qui l’a pris > puisqu’il l’a dans la main* // [Rhap-M0023, Rhapsodie]

‘^but in fact < "erm" Charlot will say (in fact) that it was he who took it > because he has it in his hand //’

IU openers may include the so-called ‘subordinating conjunctions’ as far as they introduce a new IU (cf. Section 3.1.). An example is given in (61).

(61) *il y a un homme qui fait la charnière "si vous voulez" qui se situe justement au point de désagrégation de la liter- de désagrégation historique de la littérature < c’est Sartre //* ^***parce_que*** *au fond < il a il a tenu et il tient encore encore cette sorte de leadership de "euh" de la culture et de la littérature* // [Rhap-D2009, Mertens]

‘there is a man who makes the link "if you like" who is located right at the disintegration point of lit- the historical disintegration point of literature < it’s Sartre // ^because basically < he was and he still still is a sort of leader of "erm" culture and literature //’

IU openers are indeed a subclass of what is commonly referred to as “connectives” (van Dijk 1977; Rudolph 1987).[8] As members of this class, they present some well known properties: they do not bear an illocutionary force, they do not contribute to the common ground, they cannot realize a

8 Connectives include both IU openers and microsyntactic subordinators not introducing a separate IU; besides all the connective elements realized by ad-nuclei such as (34) and (36) are commonly considered as connectives. We did not include all these elements within the same class because they do not show the same distributional properties.

speech turn in isolation, they cannot be modified by illocutionary adverbs, and they cannot be freely modified.

2.4 Associated nuclei

During the annotation task we identified a class of GUs that show a "weak" illocutionary force, though without having a full-fledged nuclear status. These GUs are usually realized by reduced parenthetical clauses, such as *je pense* 'I think', *je crois* 'I believe', *j'imagine* 'I guess' - see Blanche-Benveniste & Willems (2007) for a thorough approach to French parentheticals within a macrosyntactic perspective. An example is given in (62).

(62) *ça < c'est le problème de Paris "je pense"* // [Rhap-D0004, CFPP2000]

'that < that's the problem of Paris "I think" //'

These GUs share some distributional properties with nuclei. The sequence *je pense* 'I think' in (62), for example, can constitute an autonomous speech turn (63) and at least to some extent, it can commutate with a GU having the same locutionary content and a different illocutionary force, see example (64).

(63) $L1 ça < c'est le problème de Paris //

$L2 *"je pense" //*

'*$L1 that < that's the problem of Paris //

$L2 "I think" //'

(64) *ça < c'est le problème de Paris* ***"tu ne penses pas ?"*** //

'that < that's the problem of Paris "don't you think?" //'

It should be noted, though, that first the commutation with other illocutionary forces is not completely free, as shown by the test in (65).

(65) **a.** *ça < c'est le problème de Paris **"je pense ?"** //

‘*that < that’s the problem of Paris "do I think?" //’

b. ***ça** < c’est le problème de Paris **"je pense !"*** //

‘*that < that’s the problem of Paris "I think!" //’

Second, the performative value of the sequence cannot be made explicit, as shown in (66).

(66) ***ça** < c’est le problème de Paris "*je te dis je pense*" //*

‘*that < that’s the problem of Paris "I tell you I think" //’

Most importantly, this sequence cannot be freely modified as a true nucleus (and a true ad-nucleus) would be, see example (67).

(67) ***ça** < c'est le problème de Paris "je pense depuis longtemps*" //

‘*that < that’s the problem of Paris "I've thought for a long time" //’

The non-nuclear status of these GUs is confirmed by the fact that they cannot replace the entire IU, by fulfilling its discursive function as in (68).

(68) **a.** *^mais moi < j'ai l'impression ici qu'on devient un peu un quartier dortoir // les gens partent tôt le matin rentrent tard le soir //* ***ça < c'est le problème de Paris "je pense"*** *//*

‘^but I < I have the impression here that it’s becoming a bit of a dormitory suburb // people leave early in the morning get back late in the evening // that < that's the problem of Paris "I think" //’

≈

b. *^mais moi < j'ai l'impression ici qu'on devient un peu un quartier dortoir // les gens partent tôt le matin rentrent tard le soir //* ***c'est le problème de Paris*** *//*

≠

c. *^mais moi < j'ai l'impression ici qu'on devient un peu un quartier dortoir // les gens { partent tôt le matin | rentrent tard le soir } //* ***je pense*** *//*

Nor can these GUs be challenged in discourse as illustrated in example (69).

(69) $L1 *^mais moi < j'ai l'impression ici qu'on devient un peu un quartier dortoir // les gens partent tôt le matin rentrent tard le soir // ça < c'est le problème de Paris "je pense"* //

$L2 *non // ce n'est pas le cas* // (= ce n'est pas le cas qu'on devient un peu un quartier dortoir ≠ ce n'est pas le cas que tu le penses)

'$L2 no // it's not the case // (= it's not the case that it's becoming a dormitory suburb ≠ it is not the case that you think so)'[9]

The distributional constraints these units undergo make it possible to characterize their functional properties: these units do not contribute to the common ground shared by the speakers, still, due to their (albeit weak) illocutionary status, they provide instructions on the illocutionary interpretation of the IU they modify.

We decided to consider this type of sequence as a particular type of macrosyntactic class called *associated nuclei,* which we notated between quotation marks "…".[10] As far as their internal syntax is concerned, associated nuclei are not only realized by reduced parentheticals as shown in the examples above but also by a sub-class of discourse markers, that is, autonomizable discourse markers (Brinton 1996, Aijmer & Simon-Vandenbergen 2011, Bolly & Degand 2013). It is well known that discourse markers as a general class present some formal and functional properties also shared by associated nuclei: formally, they are only loosely syntactically linked to their host sentences; functionally, they do not contribute to the modification of the common ground by adding

9 These tests, proposed by Boye & Harder (2007) and, on other grounds, by Cristofaro (2005), allow for a distinction between what is asserted and what is not asserted in discourse, or in other words they allow for a distinction between what is foregrounded and what is not foregrounded in discourse. It is clear from (51) that the GU "je pense" is not foregrounded in discourse as a true nucleus would be.

10 In previous publications (e.g. Kahane & Pietrandrea 2012), we called them *associated illocutionary units*. But it appears now that these units are closer to nuclei rather than to IUs.

new propositional content, they provide instructions on how to interpret the information provided (see Schiffrin 1987; Degand *et al.* 2013 among others). A subclass of discourse markers - which does not include text-connective markers (see Diewald 2013 for the characterization of text-connective markers) - is also characterized by an (albeit weak) illocutionary autonomy that allows its members to occur in isolation in a speech turn. We decided to include these "autonomizable" discourse markers within the class of associated nuclei, as they present all the formal and functional properties that seem to characterize this macrosyntactic class. An example is given in (70), where the discourse marker *hein* is annotated as an associated nucleus. It displays in fact both the property of not contributing to the common ground and the property of being autonomizable as shown by the tests in (71).

(70) *c' est pas bien compliqué à y aller "hein ?"* //

'it is not too complicated to get there "is it?" //'

(71) $L1 *c'est pas bien compliqué à y aller*

$L2 *hein?*

'$L1 it is not too complicated to get there

$L2 is it?'

5. Linear relations between IUs

We have illustrated so far two of the four tasks in which the macrosyntactic annotation of the Rhapsodie corpus was organized: the identification of IUs and the annotation of the internal structure of IUs. In this section we will examine the third task we had to tackle, that is, to account for the relations holding between IUs. We distinguished three major types of linear relations between IUs: contiguous IUs (5.1), IUs embedded in another IU (5.2), and IUs interrupting other IUs (5.3). We also identified a frequent phenomenon of syntactic and semantic parallelism between contiguous IUs that seems to yield major discourse units (5.4).

5.1 Contiguous IUs

The vast majority of IUs are linearly ordered one after the other. In this case, the discursive relation between two IUs is either unmarked (72) or marked by IU openers (73).

(72) *je traverse la passerelle de Solférino // je vais à mon travail là donc à la Grande Chancellerie de la Légion d'Honneur //* [Rhap-D2001, Rhapsodie]

'I walk over the Solférino footbridge // I walk to my work there so to the Grande Chancellerie de la Légion d'Honneur //'

(73) *je fais quelques courses // ^et le soir <+ je rentre je rentre à pied tout le temps //* [Rhap-D2001, Rhapsodie]

'I do some shopping // ^and in the evening <+ I walk I walk back home every time //'

Sometimes, a difficulty can arise in deciding whether a constituent has to be integrated within an IU in progress or whether it constitutes a new illocutionary unit. This is especially true as far as phenomena of clause combining (Haiman & Thompson 1988) are concerned.

Generally speaking we concurred with the general idea that microsyntactic subordination coincides with illocutionary subordination, that is, with a lack of illocutionary force of the subordinate clause (Cristofaro 2005). This entails that subordinate clauses should be annotated as included within the same IU as their matrix clause (since they do not have illocutionary autonomy and they are therefore dependent on the matrix clause), while coordinated clauses should be considered as independent IUs. Nevertheless, we were also aware of the difficulties raised by the traditional distinction between coordination and subordination (Verstraete 2007; Debaisieux 2008). We decided therefore not to rely on the mere presence of a subordinating conjunction to decide on the microsyntactic (and hence illocutionary) subordinate vs. coordinate status of a GU. For instance, while the presence of the conjunction *parce que* 'because' in (74) indicates a true causal relation between the two linked clauses and therefore a subordinating microsyntactic relation, which entails an illocutionary subordination, the presence of the same conjunction *parce que* does not indicate such a semantic relation between the two clauses in (75). Consequently, the two clauses in (75) are

not to be analyzed as subordinated to one another either from a microsyntactic or from an illocutionary point of view. That is why we annotated a frontier of IU between the two clauses.

(74) *^mais je pense que je me réorienterai en psychologie* ***parce que*** *ça m'intéresse plus //* [Rhap-M1001, Rhapsodie]

'^but I think that I will transfer to psychology because it interests me more //'

(75) *ça n'a rien à voir avec "euh" la littérature //* ***^parce_qu'*** *en fait < "euh" j'aime la biologie //* [Rhap-M1001, Rhapsodie]

'it has nothing to do with "erm" literature // ^because in fact < "erm" I like biology //'

Operationally, we distinguished between subordinate and non-subordinate clauses by applying the extraction test proposed by Blanche-Benveniste *et al.* (1984): while a subordinate clause, governed by the verb of the matrix clause, can be extracted through the cleft construction *c'est ... que* 'it is … that' (see (76)), a non-subordinate clause cannot (see (77)).

(76) *mais je pense que* ***c'est*** *parce que ça m'intéresse plus* ***que*** *je me réorienterai en psychologie*

'but I think that it is because I am more interested in it that I will transfer to psychology'

(77) **c'est parce qu'en fait j'aime la biologie que ça n'a rien à voir avec la littérature*

'it is because in fact I like biology that it has nothing to do with literature'

Another case, in which we could not automatically rely on microsyntactic cues to segment discourse into IUs is represented by sequences such as (78), where a single major perceptual break divides a single GU into two sequences. In (78), the break occurs after the word *Chinois*.

(78) $L1 *alors < qui vous regarde* //

$L2 *c'est un Chinois* //+ *très riche* // [Rhap-D2001, Mertens]11

'$L1 so < who is looking at you //

11 Interview with Marguerite Duras.

$L2 it's a Chinese //+ very rich //'

From a microsyntactic point of view the sequence is composed of a single GU, in which the adjectival phrase *très riche* (very rich), is analyzable as an adjunct to the noun *chinois* 'Chinaman'. The perceptual break that separates the nominal constituent from the adjectival one, though, is an index of the fact that the two sequences realize two different nuclei and hence two different IUs. This intuition is corroborated by the application of the tests of nuclearity described in 4.2.1. The sequence *très riche* 'very rich' can indeed commutate with other illocutionary forces, can be uttered as a question, for example (79b); it can be introduced by an IU opener (79c); it can be modified by an illocutionary adverb (i.e. an adverb capable of modifying an illocutionary force) such as *franchement* 'frankly' (79d).

(79) **a.** *$L2 c'est un Chinois //+*
$L1 très riche //

b. *$L2 c'est un Chinois //+* ***très riche ?*** *//*

c. *$L2 c'est un Chinois //+* ***^et*** *très riche //*

d. *$L2 c'est un Chinois //+* ***franchement*** *< très riche //*

The fact that the sequence responds to the distributional and commutative tests led us to consider this microsyntactically cohesive sequence as organized in two IUs.

A final difficult case of segmentation in IUs is sequences of verb-headed constructions. In spite of the fact that from a microsyntactic point of view, the paratactic sequence of two verb-headed constructions, such as (25) reproduced here, in a slightly simplified form, as (80) can only be considered as composed of two distinct GUs, from a macrosyntactic point of view it is entirely possible for one of the GUs to be illocutionarily dependent on the other. This is indeed the case in (68), where the first GU does not show any illocutionary autonomy (it fails all the nuclearity tests)

and it is better analyzed as illocutionarily dependent on the second GU, which has instead all the properties of a nucleus, see (80).

(80) *je suis arrivée "euh" au Kenya < je voulais travailler pour le gouvernement //* [Rhap-D2004, Lacheret]

'I arrived in Kenya "erm" < I wanted to work for the government //'

Such an analysis led us to annotate the two GUs in (80) as a sequence of a pre-nucleus followed by a nucleus.

5.2. Embedded Illocutionary Units

An *embedded IU* is an IU occupying a governed position inside another IU. In other words, an embedded IU is an IU governed by an element of its host IU. From a distributional point of view embedded IUs can be characterized as IUs whose left boundary is located in the middle of another IU and that cannot be removed without prejudice to the consistency of the host IU. Two examples of embedded IUs are provided in (81) and (82).

(81) *Marcel Achard écrivait* ***[elle est très jolie //= elle est même belle //= elle est élégante //]*** [Rhap-D2001, Mertens]

'Marcel Achard wrote **[she is very pretty //= she is even beautiful //= she is elegant //]**'

(82) *vous t~ vous suivez la ligne du tram qui passe vers la &* ***[je crois que c'est une ancienne caserne "je crois" //]*** *//* [Rhap-M0003, Avanzi]

'you t~ you follow the tram line which goes toward the & **[I think it used to be barracks "I think" //]** //'

Figure 4. Macrosyntactic structure of (70)

As our examples show, two different types of embedded IUs can be distinguished: (i) *reported*

speech IUs, which occupy a position governed by a speech verb (see (81)) and (ii) *grafts* (Deulofeu 1999), that is, IUs that occupy a syntactic position where an embedded IU is not expected (see (82)).

While reported speech is a widely known syntactic object, grafts - which are more common in spoken language - have often been overlooked or treated as mere disfluencies in the literature. Let us take a closer look at grafting, which was first identified by Deulofeu (1999). Grafting is defined as the syntactic process consisting in occupying a syntactic position with an item belonging to a category different from the category expected in that position (Deulofeu, 1999). In the sequence (82), for example, one expects the position governed by the preposition *vers* 'towards' to be realized by a noun. This is not the case. The position is realized by an entire IU, co-referent with the unrealized noun phrase *[je crois que c'est une ancienne caserne "je crois"]* [I think it used to be barracks "I think"].

Unlike reported speech IUs, which are systematically governed by a specific class of words – speech verbs – and which systematically realize the syntactic position of object of the speech verb, grafts are neither governed by a specific class of words nor realized in a specific syntactic position. For example, the graft in (82) realizes the dependent of a preposition, while the graft in (83) realizes the subject position.

(83) *vous avez dit que "euh"* ***[disons ma carrière pour simplifier //]*** *témoigne de ma bonne conduite //* [Rhap-D2001, Mertens]

'you said that "erm" **[let's say my career to simplify //]** testifies to my good behavior //'

5.3. Parentheses and bifurcations

A parenthesis is an IU that interrupts another IU. From a distributional point of view, parentheses can be characterized as IUs whose left boundary is located in the middle of another IU and that can be removed without affecting the linear consistency of the host IU. We use curved brackets to

annotate them; the double slash marking the end of an IU distinguishes them from in-nuclei (that do not have illocutionary force):

(84) *"euh" d'autre part* < ***(il ne faut pas se mentir //)*** *les vacances sont nombreuses* // [Rhap-M1003, Rhapsodie]

'"erm" on the other hand < **(one cannot ignore that //)** there are many holidays //'

(85) *aujourd'hui plus que jamais* ***(^et vous le savez mieux que personne //)*** <+ *c'est sur le terrain que se gagne ou se perd le combat* // [Rhap-M2001, C-Prom]

'today more than ever **(^and you know that better than anyone //)** <+ it is on the ground that you win or lose the fight //'

When a parenthesis interrupts an IU, the host is a discontinuous IU. A discontinuous IU must not be confused with the case of discontinuous GUs discussed in Section 6.

Another interesting configuration is the *bifurcation* of an IU or GU. Bifurcations often occur when two speakers overlap: one speaker hesitates and the other speaker proposes a continuation at the moment when the first speaker finally manages to complete her utterance. (The symbols $-…-$ indicate an overlap between the two speech turns inside.) In such a case, the IU/GU started by the first speaker has two continuations. The point of bifurcation is marked #+; the first continuation is just after #+ and the second continuation after ##.

(86) *$L1 ^et ils étaient vraiment très très #+ $- assis //*

$L2 ## très bas -$ // [Rhap-D0009, PFC]

'L1 ^and they were really very very #+ $- seated //

$L2 ## very low -$ //'

Thus (86) contains two IUs: *ils étaient vraiment très très assis* 'they were really very very seated' and *ils étaient vraiment très très très bas* 'they were really very very very low'.

5.4. Discourse units beyond IUs: the case of parallelism

We are perfectly aware of the fact that the scrutiny of the semantic relations holding between IUs could provide interesting information on the interaction between prosody and syntax in the construction of discourse units; however, for the sake of simplicity, we decided not to annotate this kind of phenomena. We made an exception, though, for one particular case of relation that is both functional and formal, namely the quite frequent phenomenon of lexical and syntactic parallelism between IUs, a phenomenon reminiscent of 'dialogical resonance', recently investigated in depth by Dubois & Giora (2014). We distinguished four types of parallelism:

(i) Sequences of identical IUs. This type of parallelism has the semantic function of intensifying or confirming the propositional content of the first IU, as in (87).

(87) *"oh" tout est relatif //= tout est relatif //* [Rhap-D0009, Rhapsodie]

'"oh" everything is relative//= everything is relative //'

(ii) Sequences of IUs realized by identical syntactic heads governing in each IU elements that stand in a synonymous relation to one another. This type of parallelism has a confirmation function, as in (88).

(88) *elle est très jolie //= elle est même belle //= elle est élégante //* [Rhap-D2001, Rhapsodie]

'she is very pretty //= she is even beautiful //= she is elegant //'

(iii) Sequences of IUs realized by identical syntactic heads governing in each IU elements that stand in a co-hyponymy relation to one another. An example is (89): two IUs have been built around the predication *ils savaient pas* 'they didn't know'. In the first IU, the verb *savoir* 'to know' governs the constituent *travailler* 'to work'. In the second IU it governs the co-hyponym *utiliser un ordinateur* 'use a computer'. This type of parallelism is often employed in reformulations.

(89) *ils savaient pas travailler un & //= ils sa~ ils savaient pas utiliser un ordinateur //* [Rhap-D2005, Rhapsodie]

'they could not work a & //= they cou~ they could not use a computer //'

(iv) Contrastive sequences of syntactically parallel IUs. The head of the first IU has a semantic relation (opposition, synonymy, co-hyponymy) with the head of the following IU. There is also a semantic relation between the governed elements. Each parallel IU in (90) for example comprises a pre-nucleus: the pre-nucleus of the first IU is semantically opposed to the pre-nucleus of the second IU (*plus tard* 'after', *plus tôt* 'before'). The two nuclei also express two opposite facts: *je préviens pas* 'I don't warn', *je préviens* 'I warn'.

(90) *plus tard < je préviens pas //= plus tôt < je préviens //* [Rhap-D2007, Rhapsodie]

'after < I don't warn //= before < I warn //'

It should be noted that parallelisms between IUs recall (at least to an extent) the formal and functional properties of list constructions (Chapter 5) in that they both consist of paradigmatic lists of semantically related elements having a restricted number of functions (reformulation, confirmation, intensification, contrast). However, since from a purely syntactic point of view parallelisms cannot be regarded as a phenomenon of multiple realization of the same syntactic position, we preferred to annotate them as a distinct phenomenon (see Bonvino *et al.* 2009 for a unified account of lists and parallelisms in a constructionist perspective).

6. The interaction between macrosyntactic and microsyntactic units

We mentioned in Chapter 3 the fact that, far from being hierarchically ordered, microsyntax and macrosyntax interplay in a complex way in spoken discourse. In most cases the maximal microsyntactic units (i.e., GUs) can constitute the minimal units of macrosyntax (an example is given in (91)), but, as said above in 4.2.2 (especially Note 2), it is entirely possible for macrosyntactic relations to hold between the constituents of a single GU (or, in other words, for a single GU to be organized into more than one macrosyntactic constituent). Example (92) illustrates this point.

(91) ***donc*** < ***alors*** < *ça date de quand à peu près* > ***ce fauteuil-là*** *//*

‘so < then < it dates from when approximately > this armchair //’

(92) *^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]

‘^and following this social affairs summit <+ at eight p.m. <+ a speech by the Head of State on TV //’

We notated the microsyntactic connection between different macrosyntactic units with the symbol +.

A single GU can indeed span over more than one macrosyntactic constituent, but also across more than one IU. An example is (93).

(93) *^et_puis je suis toujours étonnée //+ maintenant < plus* // [Rhap-D2001, Mertens]

‘^and I’m always amazed //+ now < no longer //’

It is possible for a GU to cross more than one speech turn and to form several IUs, especially in the case of list phenomena (Chapter 5), as in (94).

(94) $L1 *ce qui vous a toujours intéressée < c'était chez les gens le mécanisme de la carrière //+*

$L2 *le mécanisme tout simplement pas spécialement de la carrière //+*

$L1 *et la personnalité aussi //* [Rhap-D2001, Mertens]

‘$L1 what always interested you < it was the career mechanism //+

$L2 the mechanism quite simply not especially of the career //+

$L1 and the personality too //’

An IU can be microsyntactically dependent on a non contiguous IU, giving a discontinuous GU. In (95), *dont les mamans ne parlent pas français* ‘whose mothers don’t speak French’ is not an

independent IU but a relative clause that reformulates the relative clause *qui parlent pas français* 'who don't speak French', which had occurred several IUs before.

(95) *il faudrait qu'il y ait qu'on sépare enfin qu'il y ait des cours de français pour les petits enfants* ***qui parlent pas français*** *//+ # c'est pas compliqué quand même // c'est pas très difficile d'apprendre le français à des petits enfants de cet âge-là // ça ça ça se fait assez facilement // ##* ***dont les mamans ne parlent pas français*** *//* [Rhap-D0002, Rhapsodie]

'there should be French lessons for young children **who don't speak French** //+ # it's not complicated after all // it's not very difficult to teach French to young children of this age // you can do it quite easily // ## **whose mothers don't speak French** //'

The discontinuity is marked by the symbol # and ##. As # follows a markup //+, it means that the segment after ## continues the GU of the IU before //+.

A parenthetical IU can be microsyntactically dependent, like the appositive relative clause in (96).

(96) *il y a une petite rue* ***(+ ^mais dont je ne sais pas le nom //)*** *une petite rue en & qui tourne un peu //* [Rhap-M0011, Avanzi]

'there is a small street **(+ ^but I don't know its name //)** a little street in & which winds a little //'

It is even possible for a parenthetical IU to be microsyntactically governed by another IU than its host IU. In this case, we have a discontinuity marked by # and ## as before. In (97), the segment *enfin dur relativement* 'oh relatively difficult', which is a refutation of *assez dur* 'quite difficult', forms a parenthesis hosted by the following IU.

(97) *^et comme j'étais le seul informaticien < "euh" "ben" ça c'est devenu* ***assez dur*** *//+ # ^puisque maint~* ***(## "enfin" dur relativement //)*** *^puisque maintenant < je suis responsable & //* [Rhap-D0005, PFC]

‘^and as I was the only computer scientist < "erm" it became **quite difficult** //+ # ^because **(## "oh" relatively difficult //)** ^because now < I am the person in charge & //’

7. The annotation procedure

The annotation of macrosyntax required a truly corpus-driven procedure: the annotation scheme was defined through a collective discussion of the attempts at annotating two samples performed by a group of 12 expert annotators. The definition of the repertoire of macrosyntactic units and the development of explicit guidelines (Benzitoun *et al.* 2012) took two years and can be considered as the main result of the work of the syntax group. Indeed, the annotation work did not merely build an annotation schema for pre-existing units, rather it consisted in an annotation-driven definition of the nature and the extension of a number of macrosyntactic units, which were previously poorly described (such as microsyntactic relations beyond IUs) or not described at all (such as associated nuclei).

Once the annotation schema had been defined, each sample was annotated by at least three distinct expert annotators working cumulatively: the second annotator corrected the first one and so on, and every problematic case was discussed by the group of experts.

Table 1. Criteria and markup for macrosyntactic components

	Criteria		**Markup**

	illocutionary autonomous	bears illocutionary value	is a GU	linear constraint		
nucleus	+	+	+/–	+		…//[12]
ad-nucleus	–	–	+/–	+/–		…< (…) >…
associated nucleus	–	+	+	–		"…"
IU opener	–	–	–	+		^…

8. Conclusion

In this chapter we have presented the theoretical underpinnings and the practical implementation of our macrosyntactic annotation. Building on the literature on macrosyntax, we have argued that a level of syntactic relations contributes, with microsyntax, to the formal cohesion of discourse: macrosyntax. Discourse is made up of illocutionary units (IUs). Each IU is composed of a nucleus, the segment which encodes an illocutionary force, and a number of other optional segments that do not carry an illocutionary force and that depend on the nucleus. The dependency of these segments on the nucleus is syntactically marked by the distributional constraints these segments undergo in discourse. It should be highlighted that while the terminology suggests that macrosyntax is an extension of microsyntax, this is not exactly the case: microsyntax and macrosyntax are better regarded as two independent mechanisms of cohesion that operate simultaneously in discourse. A

[12] The markup // notates both the IU and its nucleus because an IU always has a nucleus.

government unit (GU) can cross two or more IUs; an IU can be inserted or embedded within a GU.

Our framework for the macrosyntactic annotation is mainly based on the Aix-en-Provence school, although the notion of IU rather comes from the Florence school. Starting from these models, we have developed a new theoretical apparatus, introducing some new concepts such as the associated nucleus and the IU opener. Moreover the criteria for identifying nuclei and ad-nuclei have been formalized and made operational for systematic annotation. The configurational relations between IUs (that is, embedded IUs and inserted IUs) have been clarified and the fact that we need to consider microsyntactic relations beyond IUs has been systematized.