The notion of sentence and other discourse units in corpus annotation

Paola Pietrandrea, Sylvain Kahane, Anne Lacheret-Dujour, Frédéric Sabio

The notion of sentence and other discourse units in corpus annotation

Paola Pietrandrea - LLL, Université François Rabelais de Tours & CNRS, France
Sylvain Kahane - Modyco, Université de Paris Ouest Nanterre & CNRS, France
Anne Lacheret - Modyco, Université de Paris Ouest Nanterre & CNRS, France
Fréderic Sabio - LPL, Université Aix-Marseille & CNRS, France

Abstract

The notion of sentence - as it is defined in syntactic, semantic, graphic and prosodic terms - is not a suitable maximal unit for the prosodic and syntactic annotation of spoken corpora. Still, this notion is taken as a reference in many syntactic and prosodic annotation systems. We present here the modular approach we adopted for the annotation of the *Rhapsodie* corpus of spoken French, which led us to distinguish three types of elementary units operating in discourse (government units, illocutionary units, and intonational periods) and to annotate them separately. We describe the types of interactions identified among these various levels of cohesion. On this basis we propose a reappraisal of the traditional notion of sentence and we define two additional types of discourse units that we consider as the minimal and the maximal span for the notion of sentence.

1 Introduction

This article focusses on the question of units of analysis raised by the annotation of spoken corpora in a corpus-driven perspective. Our theoretical considerations are grounded in the experience acquired in developing *Rhapsodie*, a 33,000 word Treebank (57 short samples of spoken French, 5 minutes long on average, amounting to 3 hours of speech) created with the aim of modeling the interface between prosody, syntax and discourse in spoken French. *Rhapsodie* is endowed with a rich prosodic and syntactic annotation, which required at the outset a definition of the maximal units of analysis to be annotated. The complexity of this task convinced us that the notion of sentence - as it is defined in syntactic, semantic, graphic and prosodic terms - is not a suitable maximal unit for either syntactic or prosodic annotation. Rather we observed that in order to identify the maximal structures of syntax and prosody, it is necessary to take into account three mechanisms of cohesion that appear to operate simultaneously and independently from one another in spoken discourse: syntactic cohesion, illocutionary cohesion, and prosodic cohesion. These three mechanisms organize discourse in a number of independent maximal units: microsyntactic maximal units, which we call government units, macrosyntactic maximal units, which we call illocutionary units, and prosodic maximal units, which we call intonational periods.

While these maximal units are independent from one another, they can interact in a finite number of ways. The identification of the possible interactions between maximal units allowed us to define a repertoire of the structures licensed by spoken French.

Within the context of this repertoire, we propose a place for what is commonly called a sentence. As we will see, what is commonly understood by "sentence" is but a particular case of interaction between maximal units: namely, the coincidence of all three maximal units on one and the same span of discourse.

1.1 Organization of the article

Our paper is organized as follows: we will show that although the notion of sentence is quite controversial in general linguistics (Section 2), this notion is nonetheless taken as a reference in many systems of syntactic and prosodic annotation of corpora (Section 3). We will present the modular approach we adopted in the annotation of our corpus (Section 4), which led us to distinguish among government units (4.1), illocutionary units (4.2), and intonational periods (4.3),

and to annotate them separately (4.4). We will describe the types of interaction identified among these various levels of cohesion (Section 5). We will draw some general theoretical conclusions about the concurrency of different cohesion mechanisms in the definition of maximal units of spoken language and we will propose on this basis a reappraisal of the traditional notion of sentence (Section 6).

2. The notion of sentence in grammatical tradition

In 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 as 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, 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 […]."[1] (Berrendonner 1990: 25, our translation)

While linguists working on standard written language can ignore the difficulties raised by the definition of sentence, by relying on the clues provided by punctuation marks, the situation is totally different for linguists working on spoken corpora, since they do not work with sets of isolated sentences that can be analyzed internally. Rather, they deal with whole texts that need to be segmented into syntactically relevant units in order to be further analyzed internally. No matter what term is used to designate those units (sentence, utterance or other terms), their precise nature cannot be taken for granted. As Miller and Weinert (1998: 30) pointed out:

> "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 is well known, sentence-units have been given a variety of definitions, involving such heterogeneous dimensions as syntax, pragmatics, psychological reality, semantics, punctuation and prosody. In particular, the following three criteria are very often taken into account in the definition of sentences:

(i) Locutionary criterion: Sentences are frequently presented as being under the locutionary responsibility of a given speaker, who builds them in order to represent a given State of Affairs.

(ii) Graphic/prosodic criterion: The extension of sentences can be identified by relying on punctuation markers in written texts, or on major prosodic breaks in speech.

(iii) Syntactic criterion: Sentences are regularly regarded as maximal syntactic units: externally, sentences are structurally autonomous; thus they are linked to the surrounding context merely by discursive – not grammatical – relations; internally, the elements located inside the sentence-unit are related with one another by morpho-syntactic rules and can be described with regard to their grammatical function.

[1] "La 'phrase' traditionnelle, parce qu'elle n'est que l'approximation graphique, intuitive et informelle d'une unité de langue [...] constitue, de l'aveu commun, un instrument grammatical à peu près inefficace lorsqu'il s'agit de segmenter un discours oral, ou même d'analyser à l'écrit certaines configurations syntaxiques non rectionnelles"

In our view, this approach is based on an over-idealized conception of linguistic cohesion, which posits that speech segments, prosodic or graphic groupings and syntactic units should necessarily be coextensive with each other.

As shown by Sabio (2006) among others, data drawn from textual corpora clearly indicate that such a strict coincidence between these three kinds of units is indeed possible but by no means necessary. That is why we chose to assume that locutionary representations, syntactic elaboration and prosodic or graphic structuring are not necessarily coextensive and we decided to abandon the excessively "unifying" conception of sentence-units as they are traditionally defined.

3. The notion of sentence in corpus linguistics: a benchmark for the annotation task

As mentioned above, in spite of the inconsistencies of its definition, the notion of sentence is often taken as a reference for both the syntactic and the prosodic annotation of corpora.

Syntactic annotation often consists in a "bracketing" of the corpus, i.e., the single word tokens are tagged in parts of speech and the phrase structure tree is analysed in major categories such as NP, VP, etc.. Such a methodology, which is clearly sentence-based, is widely applied in the syntactic annotation of written corpora. However as observed by Nivre (2008):

> "It remains an open question to what extent the annotation schemes developed for written language are adequate for the annotation of spoken language, where interactively defined notions such as turns or dialogue acts may be more central than the syntactic notion of sentence inherited from traditional syntactic theory."

In this sense, the annotation of dependencies, i.e., the annotation of the relations holding between the words of a text, seems a more promising instrument for both the manual and the automatic analysis of sequences that cannot be represented as sentences *strictu sensu* (Bourigault 2007). In principle, that should mean that the annotation task could be performed with a bottom-up approach in order to identify the relation existing between single words, without necessarily relying on a pre-segmentation of the units to be analyzed. Still, the most important dependency-based corpora (like the Prague Dependency Treebank of Czech (Hajič 1998, Böhmová *et al.* 2003)) take as units of analysis spans of texts delimited by strong punctuation. This means that graphic sentences are, even in this framework, taken as the reference for the annotation. The authors of these corpora do not seem to question the underlying assumption that dependency relations cannot cross the boundaries of a sentence. It should also be said that most treebanks are semi-automatically annotated and that the parsers used for this task often require a pre-segmentation of the text into sentences (see for example Villemonte de la Clergerie 2005).

Concerning prosodic annotation, one of the most popular systems for intonational transcription, TOBI, is put forward by its creators (Beckman & Elman, 1997) as "a system for transcribing the intonation patterns and other aspects of the prosody of English utterances", with no further discussion of what an English utterance is. TOBI is indeed widely used for the analysis of single utterances produced within a controlled lab environment, but in spite of its success, to our knowledge, only a few limited corpora of spontaneous conversational speech have been prosodically annotated with TOBI: The Boston University Radio Speech Corpus (see Hasegawa-Johnson *et al.* 2005), and 75 Switchboard conversations in the NXT edition (Ostendorf *et al.* 2001).We cannot enter into a discussion here as to why TOBI is rarely used for the annotation of spontaneous conversations, but it can be assumed that an utterance-based system requiresmany readjustments in order to be fully exploitable on real data.

4. A modular, bottom-up, inductive approach to the annotation task

In order to overcome the difficulties raised by the weakness of the definition of sentence, we preferred not to resort to this notion in the annotation of our corpus, and chose instead a modular, bottom-up, inductive approach to the annotation task.

Our approach can be defined as "modular" because we assum that languages are organized in a number of autonomous mechanisms of linguistic cohesion operating simultaneously and independently from one another in discourse (see Roulet *et al.* (2001) and Nølke & Adam (1999), among others, for a thorough introduction to modular theories). In particular, we assume that prosodic structures do not always coincide with syntactic structures (see Mithun this volume for a discussion): such an assumption led us to separately annotate and analyze the mechanisms of prosodic and syntactic cohesion we identified in our corpus.

Building on Blanche-Benveniste *et al.* (1990), Berrendonner (1990), Cresti (2000), Andersen and Nølke (2002), we also assume that two different orders of syntactic organization can be distinguished in spoken language: microsyntax and macrosyntax. Microsyntax describes the kind of syntactic relations determined by government (usually represented in terms of dependency and phrase structure trees), whereaas macrosyntax describes other types of syntactic relations which, as we will show later, are not guaranteed by government (see 4.2). We therefore provided separate and independent annotations for all the microsyntactic relations and for all the macrosyntactic relations.

Our approach can be defined as bottom-up because rather than pre-segmenting our corpus into sentences and annotating them, we preferred to examine the dependency relations holding between the words of our texts in order to reconstruct the extension of these relations. In a similar vein, we examined the relations between the prosodic prominences present in our corpus in order to identify the extension of prosodic structures.

This approach allowed us to define inductively, i.e., through a data-driven incremental strategy of annotation, the repertoire of the relevant units of our corpus.

In the following sections we will examine one by one the extension and the definition of the various levels of analysis and annotation that we took into account in our annotation task.

4.1 Microsyntactic units: the notion of Government Unit (GU)

As mentioned above, in order to define and annotate the extension of microsyntactic cohesion mechanisms in our corpus we decided to adopt a dependency-based approach. The basic idea of dependency syntax is to connect linguistic units together (generally by dependencies between words) rather than to decompose a unit into immediate constituents. This makes dependency-based annotations particularly apt for an annotation task that seeks to avoid relying on a notion of sentence defined at the outset.

4.1.1 Berrendonner's clauses and Government Units (GUs)

In order to define the extension of the dependency units to be annotated, we revisited the notion of *clause* developed by Berrendonner (2002: 27) and we proposed the notion of *government unit* (henceforth GU).[2] The notion of GU is crucially based on the notion of government which can be defined as follows: an element X governs an element Y if X imposes constraints on Y regarding its

[2] In previous publications (Deulofeu *et al.* 2010, Benzitoun *et al.* 2010), GUs were called *dependency units*. We prefer to consider *dependency* as a formal notion used to implement various structures (here we use it for the microsyntactic structure but it could also be used for macrosyntax) rather than as a linguistic notion (even if *dependency* often stands for *microsyntactic dependency* in the literature).

linear position, its category, its morphological features, and its restructuration possibilities (commutation with a pronoun, diathesis, clefting).

Let us now specify what a GU is and what it is not.

Berrendonner defines a clause as "the projection of a syntactic dependency tree whose head does not depend on any other word in the sequence." Such a definition accounts for both verbal (1) and non-verbal government units (2).

(1) *ils étaient tout à fait normaux* (Rhap-D0002, CFPP2000)
they were absolutely normal

(2) *petite obstruction de Gabi Heinze* (Rhap-D2003, Rhapsodie,)
little obstruction by Gabi Heinze

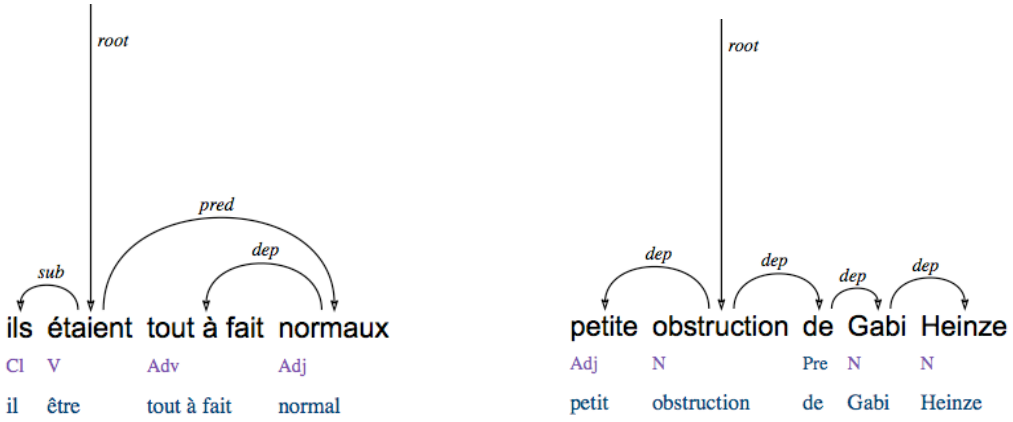


Figure 1. Microsyntactic structures of (1) and (2)[3]

Concerning the extension of the clause, Berrendonner (2011) points out that the right boundary of the clause may coincide with: (i) the absence of syntactic government; (ii) a change in the illocutionary act; (iii) a major prosodic boundary, i.e., what Berrendonner calls a conclusive intoneme; or (iv) a turn change.[4]

Our notion of GU extends Berrendonner's notion of clause in two directions:

(i) We do not think that prosodic, semantic, illocutionary or interactional phenomena should determine the extension of the clause: coherently with our modular approach we claim that only the absence of syntactic government allows for identification of the right boundary of a GU and that the breaks occurring at other structural levels should be accounted for at other levels of analysis.

(ii) We extend the domain of microsyntax beyond the notion of government by including so-called "pile phenomena" (Blanche-Benveniste *et al.* 1979, Gerdes & Kahane 2009).

[3] Our choices for the dependency annotation are described in the Rhapsodie microsyntactic annotation guide (Kahane 2013), which is mainly based on traditional work in dependency syntax (Mel'čuk 1988) except for our particular treatment of coordination and other pile phenomena presented in 4.1.3 below. The annotation was done using the resource developed by Kim Gerdes, Arborator (Gerdes 2013).

[4] Speech turns in dialogs are introduced by speakers' pseudos ($L1, $L2). The symbols // and < are respectively the end of the pre-nucleus and of an illocutionary unit (see the definition in 4.2). The symbol + indicates that a macrosyntactic boundary does not correspond to a microsyntactic boundary. See the appendix for a complete list of the symbols used in our syntactic annotation.

Let us examine these two extensions of the notion of clause in detail.

4.1.1 The boundaries of GUs

We claim that only the absence of syntactic dependency enables identification of the boundaries of a GU: a GU can in principle (as well as in practice) extend over several illocutionary acts, several speech turns, or several intonational periods. We take into account and annotate the presence of pragmatic breaks, prosodic breaks, illocutionary boundaries or turn shifts at other levels of description.

Let us take sequence (3). A major prosodic break – i.e, the end of an intonational period (IPe) as it is defined in section 4.3 - occurs after the word *Chinois* (Figure 2).

(3) $L1 *alors < qui vous regarde* //
$L2 *c'est un Chinois* //+ *très riche* // (Rhap-D2001, Mertens)

$L1 then < who is looking at you //
$L2 he is a Chinese man //+ very rich //

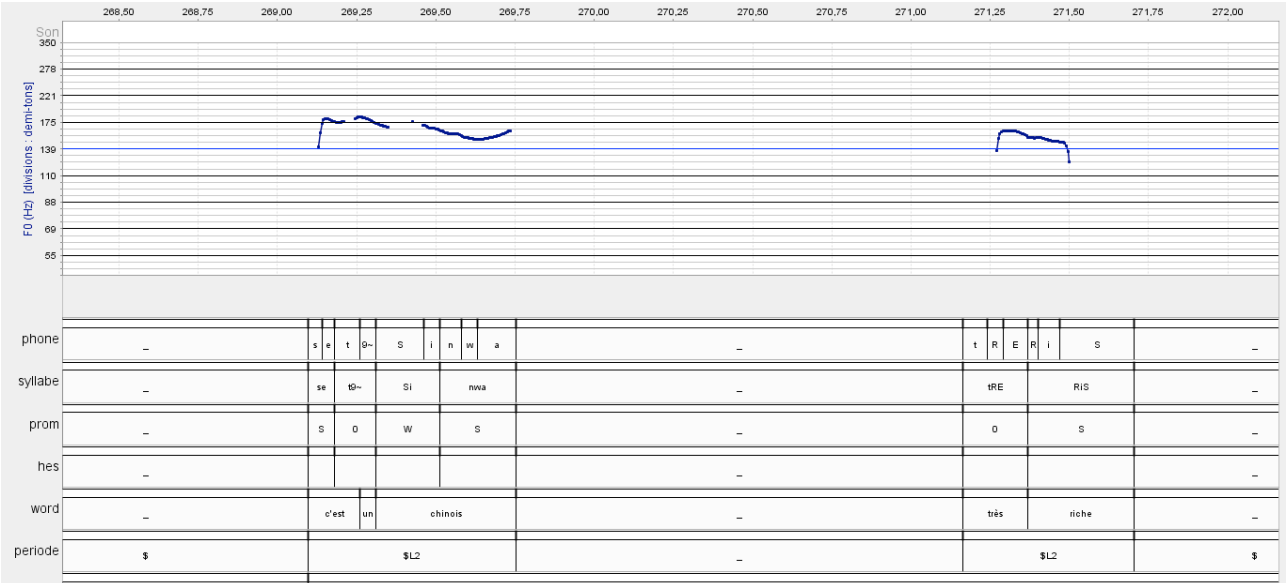


Figure 2. Prosodic annotation of (3)[5]

The presence of a major prosodic break would lead Berrendonner to analyze the sequence as organized in two distinct clauses (Figure 3, on the left) and the second clause (*very rich*) as elliptical (Groupe de Fribourg 2012: 58). By contrast, following the Aix-en-Provence framework (Blanche-Benveniste *et al.* 1990), we analyzed the entire sequence as the projection of one and only one dependency tree (Figure 3, on the right) (and as we will see below, we annotate the prosodic break in the prosodic and macrosyntactic structures).

[5] Prosodic annotation was done with Analor, as described in Avanzi *et al.* (2008). On the abscissa, temporal values are given in milliseconds; on the ordinate, the values of F0 in a logarithmic scale can be seen. Annotation tiers are, from top to bottom: phones, syllables (both in SAMPA), prominences, disfluencies, words and IPes (see *infra*, 4.4).

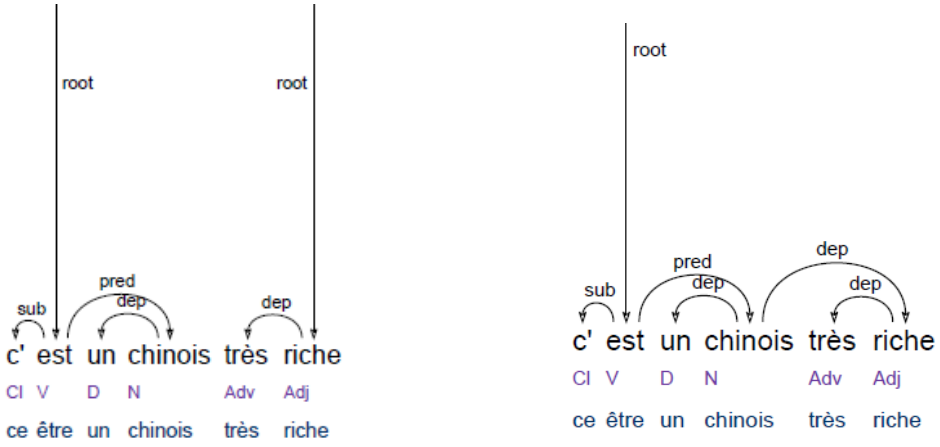


Figure 3. Berrendonner's and Rhapsodie's analyses of (3)

It is also important to highlight that our model, and consequently our annotation schema also licenses discontinuities: it is entirely possible for a GU to continue after having been interrupted by another GU. See for example (4), where, as is shown in Figure 3, the GU *vos journaux qui soulignent également la faiblesse de la mobilisation des électeurs hier* is interrupted by the appellative *Jean Christophe* that constitutes an independent GU.

(4) *vos journaux (Jean Christophe) qui soulignent également la faiblesse de la mobilisation des électeurs hier* // (Rhap-D2013, Rhapsodie, news flash)
your newspapers (Jean Christophe) which also emphasize the poor voter turnout yesterday //

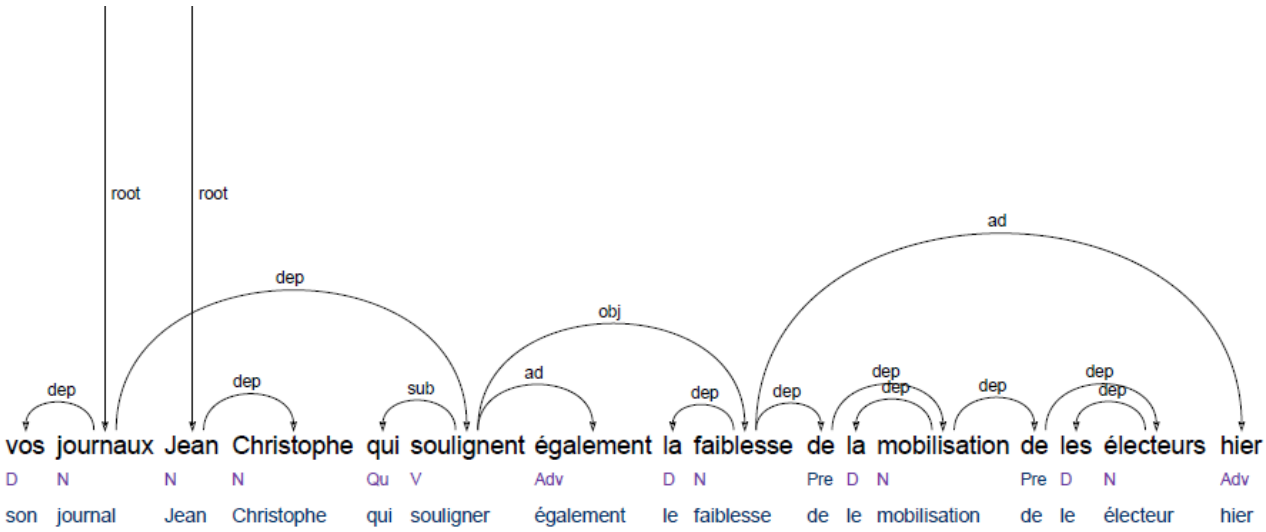


Figure 4. Microsyntactic structure of (4)

On the other hand, we follow Berrendonner in strictly identifying a boundary of GU at each break in microsyntactic dependency. In a sequence such as (5), which is intuitively cohesive, we recognize four distinct GUs (Figure 5), because we observe three breaks in the microsyntactic dependency:

(5) alors < là < la psychiatrie < c'est autre chose // (Rhap-D0006, CFPP2000)
then < now < psychiatry < that's something else //

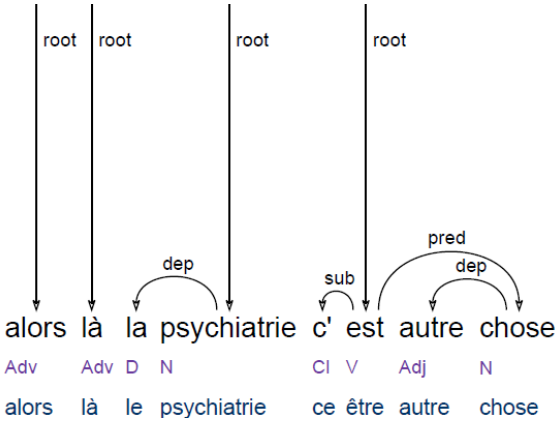


Figure 5. Microsyntactic structure of (5)

We account for the cohesion of sequences such as (5) at another level of analysis and annotation, i.e., the macrosyntactic level (see 4.2).

4.1.2 Extension of the notion of GU: The notion of pile

As mentioned above, our notion of GU includes pile phenomena in the repertoire of microsyntactic phenomena and therefore within the boundaries of a GU. By pile, we designate the fact that, within a given sequence, two or more elements – the conjuncts – occupy the same structural slot, i.e., they have the same syntactic function and the same governor. A pile may correspond to canonical coordinations, as in example (6):[6]

(6) *c'est aussi là l'intérêt fondamental { de l'Europe | ^et de nos partenaires } //*
(Rhap-M2001, C-PROM)
that's also the fundamental interest { of Europe | ^and of our partners } //

The two segments *de l'Europe* and *de nos partenaires* occupy the same structural position in the sequence, a position governed by the noun *intérêt.* The conjuncts are in a paradigmatic relation, represented in Figure 6 by the dependency link labeled *para_coord*. This link is overarched by the junction links between the pile marker *et* 'and' and the conjuncts. Each conjunct depends on *intérêt*, the first one through a true dependency and the second one by an inherited dependency link labeled *dep_inherited* (see Gerdes & Kahane (2009) and Kahane (2012) for details and justification).

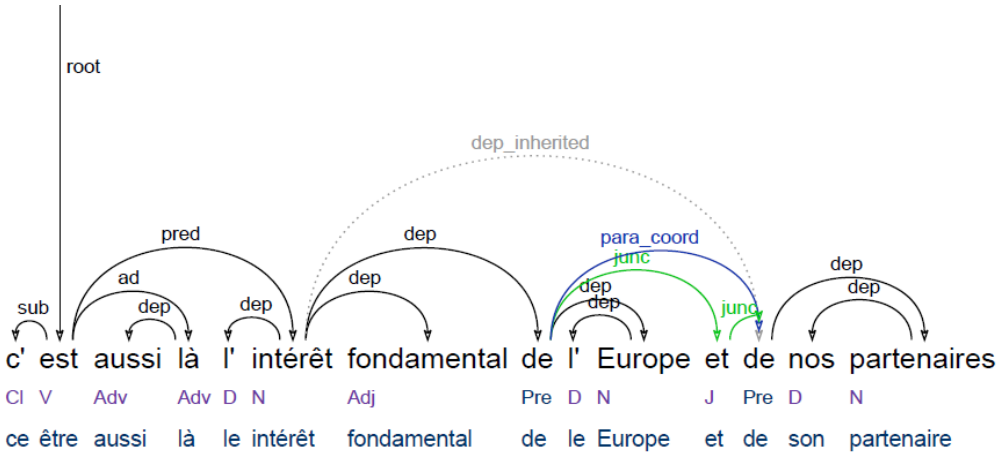


Figure 6. Microsyntactic structure of (6)

[6] Piles are annotated using parentheses {…|…} ; the symbol | indicates the limit between two layers of the pile. Junctors, that is coordinate conjunctions, are marked by ^.

As argued by Blanche-Benveniste (1990), Gerdes & Kahane (2009), Bonvino, Masini & Pietrandrea (2009), Kahane & Pietrandrea (2012a), other phenomena such as intensive repetitions (7), disfluencies (8), reformulations (9), corrections and confirmations (10), etc. can be regarded as pile phenomena due to the fact that the elements piled up occupy the same syntactic slot in the sequence:[7]

(7) *et Rozysky dit* [*on pouvait pas s'empêcher à la fin de* { *Mort* | *^et transfiguration* } *de faire* **{ *résonner* | *résonner* | *^et résonner* | *^et encore* }** *ces accords qui nous enchantaient* //] // (Rhap-D2012, Rhapsodie)
and Rozysky says [one could not avoid at the end of { Death | ^and transfiguration } letting **{ resonate | resonate | ^and resonate | ^and again }** these chords that enchanted us //] //

(8) *ça* < **{ *j'en ai* | *j'en ai* }** *pas beaucoup* > *quand même* // (Rhap-D2002, Rhapsodie)
that < **{ I don't| I don't }** have much anyway //
'I don't have much of that anyway'

(9) *^et si vous faites de la musique* < "*eh bien*" *vous avez l'expérience* **{ *de la poïésis* | { *de la* | *de la* } *production musicale* }** // (Rhap-M2002, Rhapsodie)
'^and if you practice music < "well" you have the experience { of poïesis | { of | of } musical production } //'

(10) *c'est la crise générale* { { ***des*** | ***des*** } ***Français*** |} //+ {(***"enfin" des Français*** //) | (*pas simplement des Français* "*hein*" //) | { { *des* | *de* } *l'humanité* | *^et de la lecture* } } //[8] (Rhap-D0004)
it is the general crisis {{of | of } French people |} //+ {("well" French People //) | (not only French people "ok" //) | {{ of | of } humanity and of readership} } //

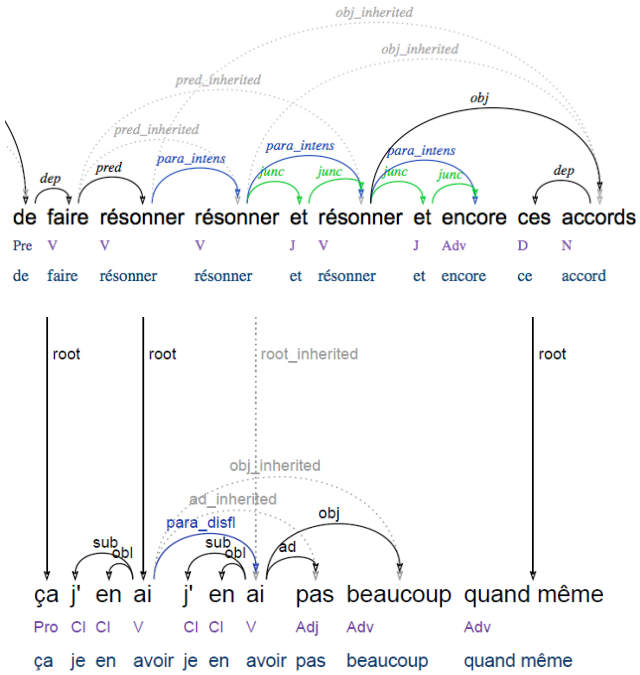


Figure 7. Microsyntactic structures of (7) and (8)

[7] See Bonvino, Masini & Pietrandrea (2009) and Kahane & Pietrandrea (2012a) for a complete typology of pile phenomena.
[8] The quotes "…" mark the discourse markers that function as associated nuclei (see 4.2.3).

We included pile phenomena in the description of the microsyntactic structure because we assumed that the paradigmatic relation between two conjuncts is a particular type of microsyntactic dependency. The inclusion of pile phenomena in the repertoire of microsyntactic phenomena substantially extends the boundaries of microsyntactic units as compared to more traditional analyses. Interestingly, pile phenomena tend to occur in dialogical constructions: speakers often use this cohesion mechanism to pile up with the discourse of their interlocutors. Since we do not consider a turn change as an interruption of a GU, we often came up in the annotation of our corpus with long GUs, made up of the layers of dialogical piles.[9] In (11), a GU spans over four speech turns. It is worth highlighting that by considering the different speech turns as part of one single co-constructed GU we were able to avoid resorting to the notion of ellipsis to account for the cohesion of the entire sequence: each turn is simply the continuation of the microsyntactic structure of the previous one by a pile structure (see Figure 8).

(11) $L1 ^*et il donne* { *à Gaga* |} //+
$L2 {| *à* { *Gago* |} } > *effectivement* //+
$L1 {| *Gago* |} "*pardon*" //+
$L2 {| *Gago* } { *qui est contré* | *qui est contré* } //
(Rhap-D2003, Rhapsodie)

$L1 ^and he gives { to Gaga |} //+
$L2 {| to { Gago |} } > actually //+
$L1 {| Gago |} "sorry" //+
$L2 {| Gago } { who is blocked | who is blocked } //

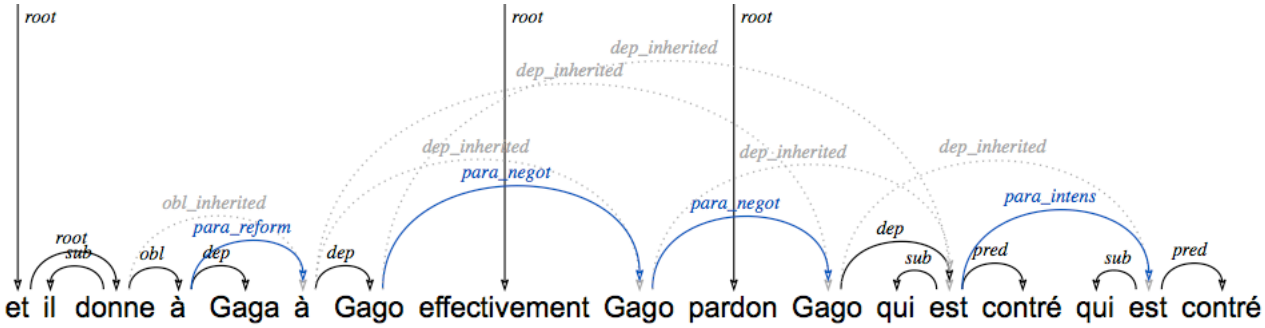


Figure 8. The microsyntactic structure of (11)

4.2 Beyond microsyntax: The notion of IU

The notion of GU is not sufficient for the annotation of spoken corpora. Let us examine for example the following two utterances:

(12) *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 //

(13) *ça a duré dix ans > le silence autour de moi* // (Rhap-D2001, Mertens)
it lasted two years > the silence around me //

The successive sequences contained in each utterance are not microsyntactically related: according to the definition of dependency formulated above it is not possible to identify any (microsyntactic) dependency relation between any of the successive segments; still it is intuitively clear that they do

[9] Such co-constructed microsyntactic units have been characterized as a "collective speaker" phenomenon (*locuteur collectif*) in the Aix-en-Provence framework; cf. Loufrani, 1984

have a cohesive status in certain respects. The question arises what determines the cohesion of these sequences. In order to answer this question, we borrowed some categories of analysis from macrosyntax.

As mentioned above, macrosyntactic models claim that discourse is organized in maximal units whose cohesion is guaranteed by relations that go beyond a strict microsyntactic dependency. All macrosyntactic models would acknowledge for example that the sequences (12) and (13) have to be considered as a unified unit. The question arises what justifies this intuition of cohesion.

The different macrosyntactic models do not provide a unique answer to this question. According to the Aix-en-Provence school, sequences (12) and (13) constitute macrosyntactic units, i.e., a succession of distinct GUs whose cohesion is guaranteed by the rather vague notion of *togetherness* based on Bolinger (1968) and Blanche-Benveniste (1990: 114). According to the Fribourg School, sequences (12) and (13) form a single "macrosyntactic period", that is a sequence of communicative actions marked by a single conclusive intoneme (Berrendonner 2002). According to the Florence School, the sequences in (12) and (13) constitute utterances, i.e., sequences of prosodic units whose cohesion is guaranteed by the fact that the entire sequence conveys one and only one illocutionary act, in this case an assertion (Cresti 2000, this volume).

To remain coherent with our modular approach we could not follow the prosodic definition of macrosyntactic units proposed by Berrendonner. We did not assume therefore that the macrosyntactic cohesion of a sequence is guaranteed by the existence of a conclusive intoneme. Rather, we built on Cresti's proposition that the cohesion of sequences such as (12) and (13) is determined by the fact that they encode a single illocution and we propose therefore that a maximal unit of macrosyntax coincides with the maximal extension of an illocutionary act, i.e., all the GUs that contribute to forming one and only one assertion, injunction, interrogation, etc. We called the maximal units of macrosyntax, *illocutionary units* (henceforth IU). As we will see in the following sections, we proposed a number of criteria that allowed our annotators to precisely identify the extension of IUs and of their components (see below).

4.2.1 The nucleus and the other illocutionary components of an IU (ICs)

In this section we illustrate the criteria used for the identification of the *illocutionary components* of an IU (henceforth IC): the nucleus and the pre-nuclear and post-nuclear components.

Let us take utterance (12), here reproduced as (14): this IU is formed by three GUs: (i) *ceux qui sont en location* "those who are on a lease"; (ii) *la moyenne* "the average"; (iii) *c'est environ trois ans* "it is about three years''.

(14) *ceux qui sont en location < la moyenne < c'est environ trois ans* //
those who are on a lease < the average < it is about three years //

Building on Blanche-Benveniste (1990), Berrendonner (1990), and Cresti (2000) we defined the nucleus as the only unit of an utterance endowed with communicative autonomy. The nucleus is the only unit that can be uttered alone. Such a definition led us to consider *the possibility of being autonomized* as the first test for the identification of nuclei. In (14) for example, the GU *c'est environ trois ans* 'it is about three years'' can be interpreted even when uttered without the two pre-nuclei (15), whereas the two pre-nuclei could not be interpreted without the presence of the nucleus (16) :

(15) *c'est environ trois ans* //
It's about three years

(16) **ceux qui sont en location < la moyenne < & //*[10]
those who are on a lease < the average < & //

According to Cresti, the communicative autonomy of the nucleus is due to the fact that the nucleus is the only unit in an utterance endowed with an illocutionary force: it can be, in other words, interpreted as an assertion, as a question, as an injunction, or as an exclamation (see Cresti this volume, for further details). Such a definition of the nucleus in terms of illocution led us to develop a second test which distinguishes nuclear from non-nuclear units on the basis of *the possibility, within the same context, of making the implicit performative explicit.* In (14), for example, it is possible to make explicit the performative of the GU *c'est environ trois ans* 'it is about three years' (17), but not the performative of the preceding two GUs (18), (19):

(17) *je te dis c'est environ trois ans* //
I tell you it is about three years //

(18) *??je te dis ceux qui sont en location* //
I tell you those who are on a lease //

(19) *??je te dis la moyenne* //
I tell you the average //

The fact that the nucleus is endowed with an illocutionary force makes it possible to qualify such a force through an utterance adverbial (i.e., an adverb qualifying the illocutionary force of a sequence, such as *frankly, briefly speaking, roughly speaking* – Nølke 1990). Such a property led us to develop a third criterion for the identification of nuclei consisting in testing *the possibility for a unit, of entering the scope of an utterance adverb, without changing context*. In (14) the GU *c'est environ trois ans* 'it is about three years' can enter the scope of an utterance adverb (20), whereas the preceding two GUs cannot (21), (22):

(20) *franchement/ pour faire court c'est environ trois ans* //
frankly/ briefly speaking it is about three years //

(21) *?? franchement/ pour faire court ceux qui sont en location* //
frankly/ briefly speaking those who are on a lease //

(22) *?? franchement/ pour faire court la moyenne* //
frankly/ briefly speaking pour faire court the average //

As it is endowed with an illocutionary force, the nucleus can commute with other GUs having the same locutionary content, but a different illocutionary force. Such a property, already noted by Blanche-Benveniste *et al.* (1990) constitutes the basis for a fourth test we developed, which distinguishes between the nucleus and other illocutionary components on the basis of their *commutability with other illocutionary forces.* As shown by the tests (23) through (25), the GU *c'est environ trois ans* 'it is about three years' can commute with other GUs having the same locutionary content, but a different illocutionary force, (23) whereas the preceding two GUs cannot (24) and (25)

(23) *ceux qui sont en location < la moyenne < c'est environ trois ans !* //
those who are on a lease < the average < it is about three years! //

(24) ?? *ceux qui sont en location ! < la moyenne < c'est environ trois ans* //
those who are on a lease ! < the average < it is about three years //

[10] Such a sequence would be perceived as incomplete and hence uninterpretable; the symbol & indicates the fact that the sequence is incomplete.

(25) ?? *ceux qui sont en location < la moyenne ! < c'est environ trois ans* //
those who are on a lease < the average ! < it is about three years //

Once the nucleus of an IU has been identified, it is quite easy to characterize the neighboring GUs as non autonomous from an illocutionary point of view, and to classify them according to their linear position, as pre-nuclei, post-nuclei, and in-nuclei.

An example of a complex IU, made-up of two pre-nuclei, a nucleus, an in-nucleus, and a post-nucleus is (26):[11]

(26) *^et là < ce que je vous propose <+ c'est d'écouter (bien sûr) le spécialiste nous expliquer comment ça marche > notre boule magique* // (Rhap-D2011, Rhapsodie)
^and now < what I propose to you <+ is to listen (obviously) to the expert who will explain to us how it works > our magic ball //

It is worth highlighting that, unlike Cresti (2000, this volume, but see also Moneglia 2011) we do not rely exclusively on perceptual criteria to identify macrosyntactic units.

Obviously, it is often necessary to listen to the sequence in order to identify the right segmentation, but, in our view, perceptual criteria are neither necessary nor sufficient for the identification of the macrosyntactic structure.

We do not rely exclusively on perceptual criteria because, for example, given a sequence such as (14), no matter how this sequence is uttered - whether in three prosodic units or in a single prosodic unit - we analyze it as composed of three distinct GUs and we claim that these three GUs are linked at the macrosyntactic level because of the illocutionary dependency of the first two units on the third one. This analysis is guided on the one hand by syntactic cues (there are two microsyntactic breaks in the sequence, so the sequence has to be analyzed in three GUs) and on the other hand by the nuclearity tests that acknowledge the third GU as the nucleus of the sequence.

Perceptual criteria are not sufficient, because it may happen that even major prosodic breaks serve other functions than marking the macrosyntactic structure (for example, at the pragmatic level, they may mark phenomena related to information packaging, focus marking, specific rhythmic scansion linked to rhetoric style and, more generally, expressive processes – Lacheret (2003), Lacheret *et al*. (2011)).

All in all, we claim that perceptual criteria may sometimes guide the segmentation, but only the application of nuclearity tests capable of verifying the congruity of the syntactic-semantic interface of the units identified allows for a correct characterization of the macrosyntactic structure of a sequence.

It should also be said that we do not believe that a prosodic theory defined at the outset can guide the (macro)syntactic segmentation. Indeed, our project was based on the necessity of keeping prosodic and (macro)syntactic annotations clearly separate in order to identify empirically, in a further step, the correlations between syntactically defined units on the one hand and prosodically defined units on the other hand (section 5.3).

[11] The initial element *et* 'and' is classified as an IU introducer. It is less mobile than an ad-nucleus and must occupy the first slot of the IU (and consequently excludes any other introducer). We mark introducers with ^, usingthe same symbol as for junctors.

4.2.2 Extension of the notion of IU: The notion of associated nucleus

Let us consider the sequence in (27):

(27) *ça < c'est le problème de Paris "je pense" //* (Rhap-D0004, CFPP2000)
that < that's the problem of Paris "I think" //

It is intuitively clear that this sequence is cohesive to some extent; but let us examine its composition in detail. The sequence is made up of three GUs: *ça, c'est le problème de Paris* and *je pense*. The GU *c'est le problème de Paris*, like the GU *c'est environ trois ans* examined in (14), has all the properties of a nucleus: it is autonomizable, its performative value can be made explicit, it can enter the scope of an utterance adverb, and its illocutionary force can commute with other illocutionary forces.

The GU *ça* does not satisfy any of the tests of nuclearity: it cannot be autonomized, it is not possible to make its performative value explicit, it cannot enter the scope of an utterance adverb, and it cannot commute with other sequences bearing different illocutionary values.

Let us now consider the third GU made up of the sequence, *je pense*.

This third GU has some properties of a true nucleus. It can indeed be autonomized (28), and, at least to some extent, its illocutionary force can commute with other illocutionary forces (29):

(28) $L1 *ça < c'est le problème de Paris* //
$L2 "*je pense*" //

$L1 that < that's the problem of Paris //
$L2 "I think" //

(29) *ça < c'est le problème de Paris "tu ne penses pas ?"* //
that < that's the problem of Paris "don't you think?" //

Still, this GU does not meet all the tests of nuclearity. It cannot freely commute with other illocutionary forces (30) and its implicit performative cannot be made explicit (31):

(30) **ça < c'est le problème de Paris "je pense ?"* //
that < that's the problem of Paris "do I think?" //

(31) **ça < c'est le problème de Paris "je te dis je pense"* //
that < that's the problem of Paris "I tell you I think" //

Finally, it shows a property which moves it away from both nuclei and ad-nuclei, and brings it closer to interjections: it cannot easily be modified (32):

(32) **ça < c'est le problème de Paris "je pense depuis longtemps"* //
that < that's the problem of Paris " I have thought for a long time " //

The question arose during our discussions how to analyze and annotate sequences such as *je pense*. On the one hand they seem to be endowed with an illocutionary marker that makes it possible to manipulate their illocutionary force, which would argue in favor of an annotation as true nuclei, on the other hand they undergo a number of constraints that do not allow for classification as fully autonomous nuclei.

We preferred to consider this type of sequence as a particular type of macrosyntactic object and we called them *associated nuclei*.[12] An associated nucleus has some properties of a true nucleus (it has

[12] In previous publications (e.g. Kahane & Pietrandrea 2012b), we called them *associated illocutionary unit*. But it appears now that it is a relation between nuclei rather than between IUs. See the discussion below.

an illocutionary force) but it is less autonomous than a true nucleus. It is anchored to another nucleus (here *c'est le problème de Paris*), but it is neither microsyntactically nor illocutionarily dependent on its anchor.

We observed indeed that the lack of autonomy of these sequences can be considered as a side effect of their semantic dependency, rather than of their illocutionary dependency on the nucleus of the IU. A sequence such as *je pense* is realized by an unsaturated predicate: the predicate *penser* 'to think' is a bivalent predicate, obligatorily selecting a subject and an object. Within the limits of the GU only one of its arguments, the subject, is saturated. A number of analyses agree in considering the anchor as the semantic object of this type of predicate (called *parentheticals* in the literature – Ross (1973), Schelfhout *et al*. (2004), Dehé & Kavalova (2006)). Such a relation explains the syntactic constraints that associated nuclei undergo.

First of all, an associated nucleus can only have a nucleus as anchor. In (33)a, *je pense* can be anchored on *elle est venue* (33)b or on *l'autre jour* (33)c, but in the second case, *l'autre jour* is necessary a nucleus and the segment contains two assertions ('she came' and 'I think it happened the other day').

(33) **a.** elle est venue, l'autre jour, je pense
she came, the other day, I think

b. elle est venue >+ l'autre jour "je pense" //
she came >+ the other day "I think"

c. elle est venue //+ l'autre jour "je pense" //
she came //+ the other day "I think" //

Moreover an associated nucleus can only anchor one nucleus. In (34)a, the associated nucleus *je p~ je pense* can only predicate on the last nucleus (with which it is adjacent), and can be paraphrased by (34)b and not by (34)c.

(34) **a.** *"euh" dans la confusion <+ donc < une "euh" une passante a dénoncé la jeune fille au livreur qui a couru apres la jeune fille "euh" // les policiers sont arrivés en raison du du du vacarme "je p~ je pense"* // (Rhap-M0024, Rhapsodie)
in the confusion <+ then < a "uh" a passer-by denounced the girl to the deliveryman who ran after the girl "uh" // the policemen arrived because of the the the din "I th~ I think" //

b. A passer-by denounced the girl to the deliveryman who ran after the girl. I think that the policemen arrived because of the din.

c. # I think that a passer-by denounced the girl to the deliveryman who ran after the girl and that the policemen arrived because of the din.

We call the sequences formed by an associated nucleus and its anchor *associated sequences* (AS). We extend the notion of IU by considering that ASs are in the same IU as their anchor.

Let us note incidentally that we included in the repertoire of associated nuclei a number of discourse markers that are analyzable as unsaturated predicates taking their anchors as arguments, such as *bon, hein*:

(35) *et c'est vrai que "bon" habitant dans le centre de Paris "euh" < les écoles sont de très bon niveau "hein" "je veux dire"* // (Rhap-D0002, CFPP2000)
and it's true that "well" living in the center of Paris "uh" < the schools have a pretty high standard "eh" "I mean" //

This classification allows for a formal interpretation of the parenthood often identified in the literature between discourse markers and parenthetical units (Brown & Levinson, 1978, Östman 1981, Holmes 1986, Schiffrin 1987, Bazzanella 1995, Aijmer 2002, Kärkkäinen 2003). We claim that discourse markers and parenthetical units not only have the same function, but also that they establish with their anchors the same macrosyntactic and predicate-argument relations (Kahane & Pietrandrea 2012b).

4.3 The role of prosody: intonational periods, intonational packages, rhythmic units

Many fundamental questions remain unsolved for the prosodic annotation of continuous speech in French, in which the prosodic system has a number of typologically peculiar features. Particularly, scholars highlight the syncretism between accentuation and intonation, or, to put it more accurately, the influence of intonation over accentuation (Rossi, 1979). In view of these findings, the approach used for the prosodic annotation of Rhapsodie was based on two principles

(1) As for syntactic annotation, the prosodic annotation is processed autonomously: neither microsyntactic nor macrosyntactic criteria are used to differentiate prosodic labels (ex. lexical tone vs. boundary tone). From this point of view, we do not make any hypothesis a priori regarding the functional role of an annotated segment. Potentially it can have two basic functions: a) phrasing and grouping, b) hierarchisation of groups. More specifically, in French, we cannot say a priori whether phrasing and grouping have only a syntactic function of demarcation; because of the lack of lexical stress in this language, phrasing and grouping may also have a focalization function (see Chafe 1998: the "spotlight of consciousness principle"). We return to this point later when comparing syntactic and prosodic annotations.

Consequently, the first step of the annotation is to select perceptual cues that are relevant for the segmentation of the prosodic flow.[13] The basic tenet underlying our prosodic annotation is the phonetic hypothesis formulated by the Dutch-IPO school ('t Hart *et al.* 1990) stating that, out of all the information characterizing the acoustic domain, only some perceptual cues selected by the listener are relevant for linguistic communication. On this basis we decided to manually annotate three perceptual phenomena characterizing real productions: prominences, non silent pauses and disfluencies.[14] We annotated perceptual syllabic salience in speech by using a 3-level scale distinguishing between strong (S), weak (W), and zero (0) prominences. The annotation of the prominences identified in sequence (36) is represented in the third tier of Figure 9:

(36) *alors < je sais bien* (*Marguerite Duras*) *que* { *v~* | "*c'est vrai*" *vous* } *avez obtenu d'autres succès* // (Rhap-D2001, Mertens)
so < I know well (Marguerite Duras) that { y~ | "it's true" you } have had other successes //

[13] The acoustic properties of prosodic segments (ex: pitch range, change of pitch range, complex pitch movements, steepness of pitch movements, local vs global melodic variations, variations of tempo, etc;) are not taken into account for this first step of annotation, but all the information for a complete analysis is available.
[14] The prosodic notion of disfluency is different from the syntactic notion introduced in 4.1.3 above.

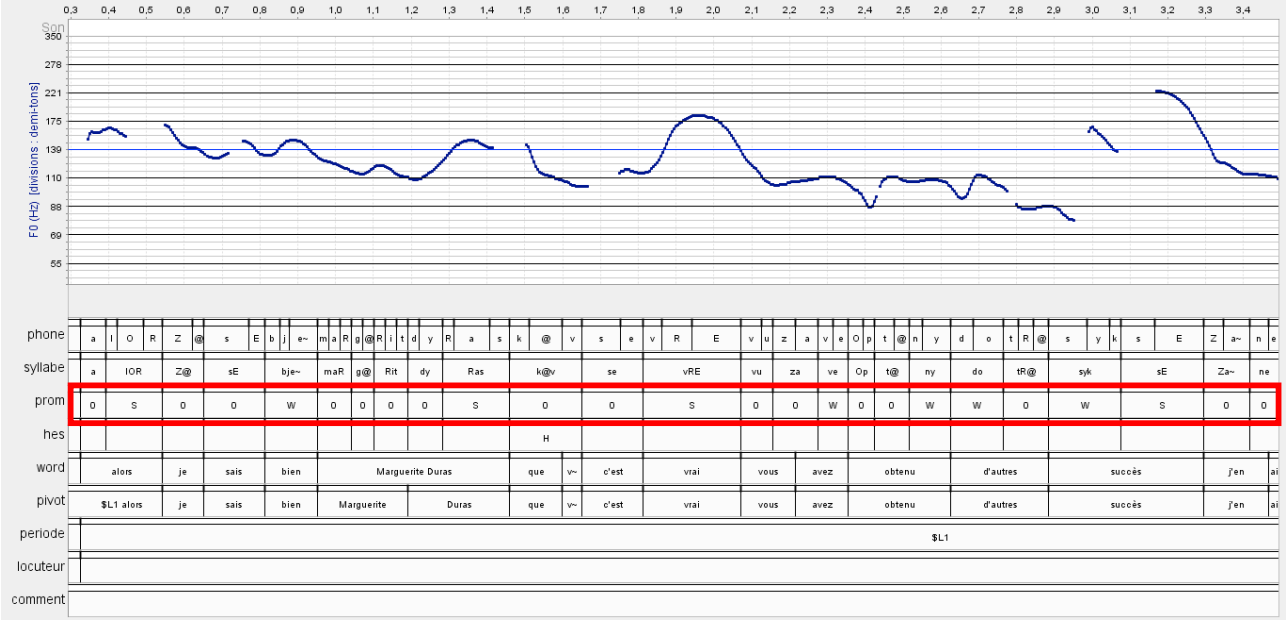


Figure 9. Prosodic annotation of (36): zoom on prominences.[15]

For disfluencies, we considered false starts, repetition of non-lexical elements, "*uh*", and unexpected syllabic lengthening. An example of the annotation of the disfluent sequence (37) is in the fourth tier of Figure 10.

(37) $L2 "*eh ben*" *XXX ^soit on* $- *travaille* //
$L1 *par exemple* //-$[16]
$L2 ^*soit on travaille* // (Rhap-D0001, CFPP2000)
$L2 "well" XXX ^either one $- works //
$L1 for example //-$
$L2 ^either one works //

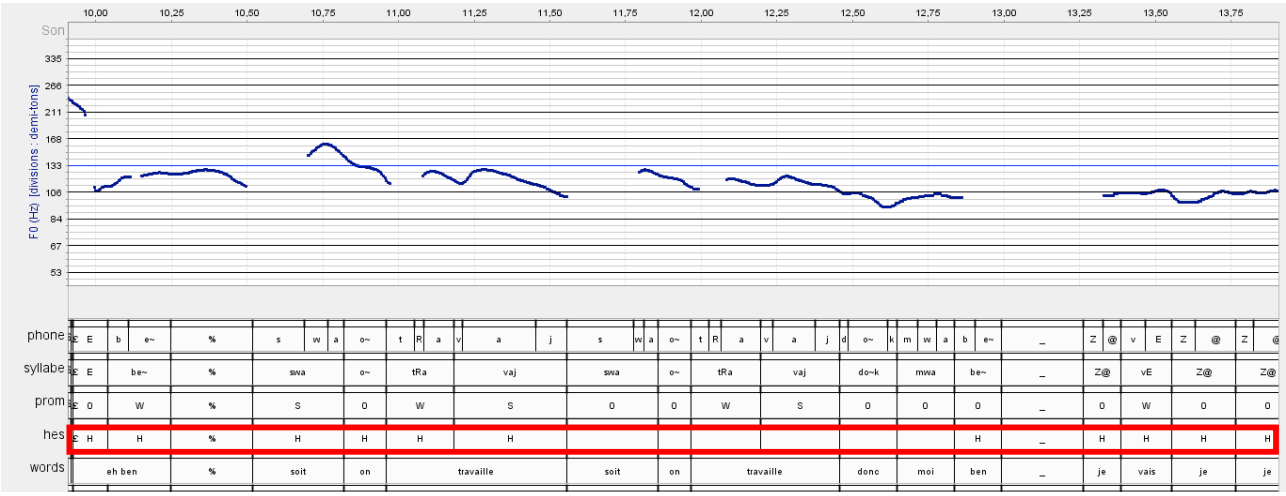


Figure 10. Prosodic annotation of (37): zoom on disfluencies

Building on the hypothesis that prominences are hierarchically organized and that the distribution of prominences, disfluencies and non silent pauses defines different degrees of prosodic cohesion

[15] All our prosodic figures are Analor screenshots (Avanzi et al. 2008).
[16] $- … -$ indicates an area where the two speech turns overlap.

within the intonational period, we took the annotation of these elements as input to automatically generate a prosodic structure, organized around rhythmical and intonation components. Our algorithm identifies four levels of prosodic cohesion. Strong prominences on word-final syllables mark what we called *intonation packages* (henceforth IPa); weak and strong prominences on word-final syllables mark *rhythmic* groups (henceforth RGs); prominences (on whatever syllables) mark *metrical feet*. These three constituents are embedded in macroprosodic units called *intonational periods* (hencefort IPe), which are acoustically defined as sequences delimited by a silent pause of at least 300 ms, a major contour (an F0 pitch movement reaching a certain amplitude), a "pitch reset" (a difference in height between the last F0 extreme preceding the pause and the first F0 value following the pause), and the absence of disfluencies in the immediate proximity of the pause. Figure 11 represents the prosodic annotation of (38):

(38) *je sais également ce que cela signifie pour vos familles* (Rhap-M2001, C-PROM)
I also know what that means for your families

This sequence is constituted by one IPe, organized in two IPas *je sais également* and *ce que cela signifie pour vos familles*. The first IPa embeds the RGs *je sais* and *également*, the second the three RGs *ce que cela*, *signifie* and *pour vos familles*:

([*je sais*]$_{RG}$ [*également*]$_{RG}$)$_{IPa}$ ([*ce que cela*]$_{RG}$ [*signifie*]$_{RG}$ [*pour vos familles*]$_{RG}$)$_{IPa}$ ||$_{IPe}$

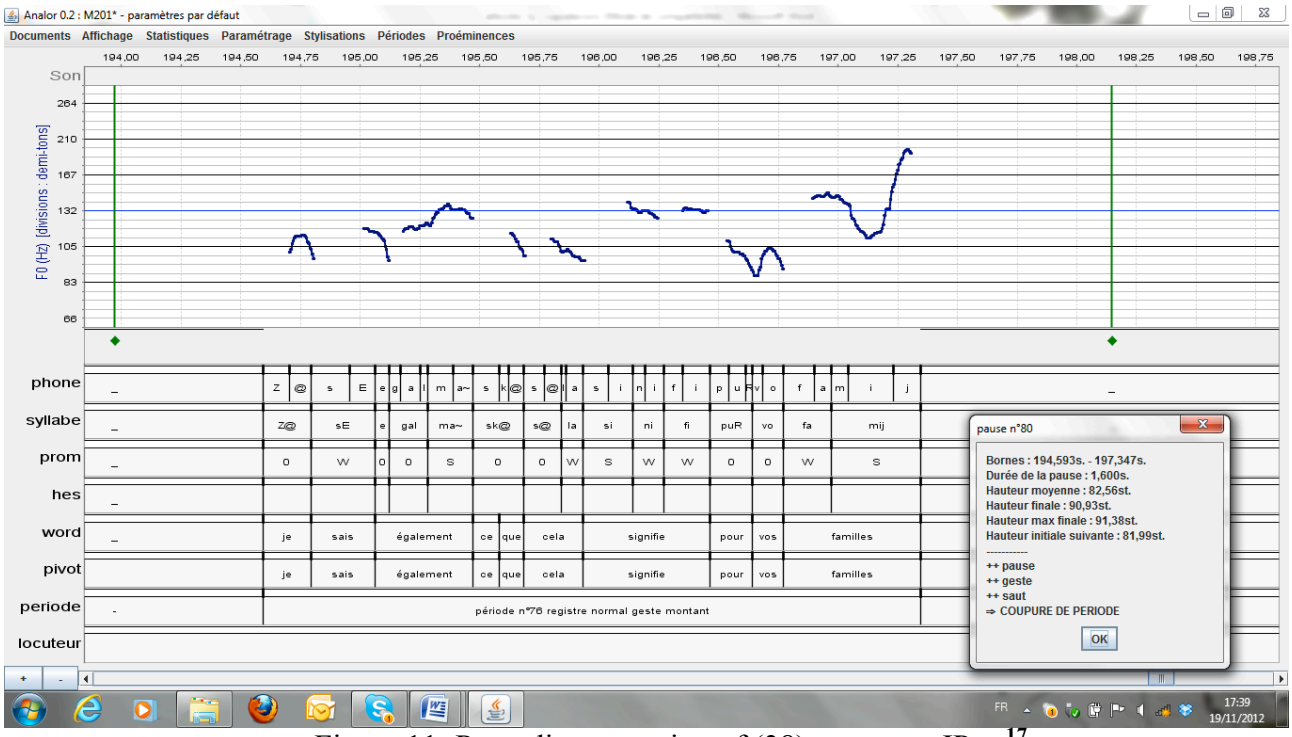


Figure 11. Prosodic annotation of (38): zoom on IPes.[17]

It should be noted that our prosodic annotation schema is substantially different from most of the syllable-based schemata which are most of the time derived through a top-down procedure from a

[17] On tier 7 (from top to bottom), the IPe register (medium) and the final contour direction (rise) are indicated. On the right screen, the values of the different parameters and thresholds activated are given:

- Occurrence of a pause (1.6s)
- Detection of an F0 pitch movement reaching a certain amplitude, defined as the difference in height between the last F0 extremum (91.38 semi-tones) and the mean F0 over the entire portion of the signal preceding the pause (82.56 semi-tones).
- Detection of a "pitch reset", defined as the difference in height between the last F0 extremum preceding the pause and the first F0 value following the pause (81.99).

Each parameter is well above the threshold ('++'); e.g.. for the pause: threshold = 300 ms vs actual duration : 1.6s

pre-existing phonological framework built on abstract syntactic properties of words and groups (see Lacheret & Beaugendre 1999 for a presentation).

The approach we adopted allowed us to provide a complete prosodic annotation without any reference to the notion of sentence or utterance. We were therefore able to completely annotate our corpus and to identify a number of genuinely prosodic primitives to be used as input in the study of the prosodic-syntactic interface.

5. Interactions between units

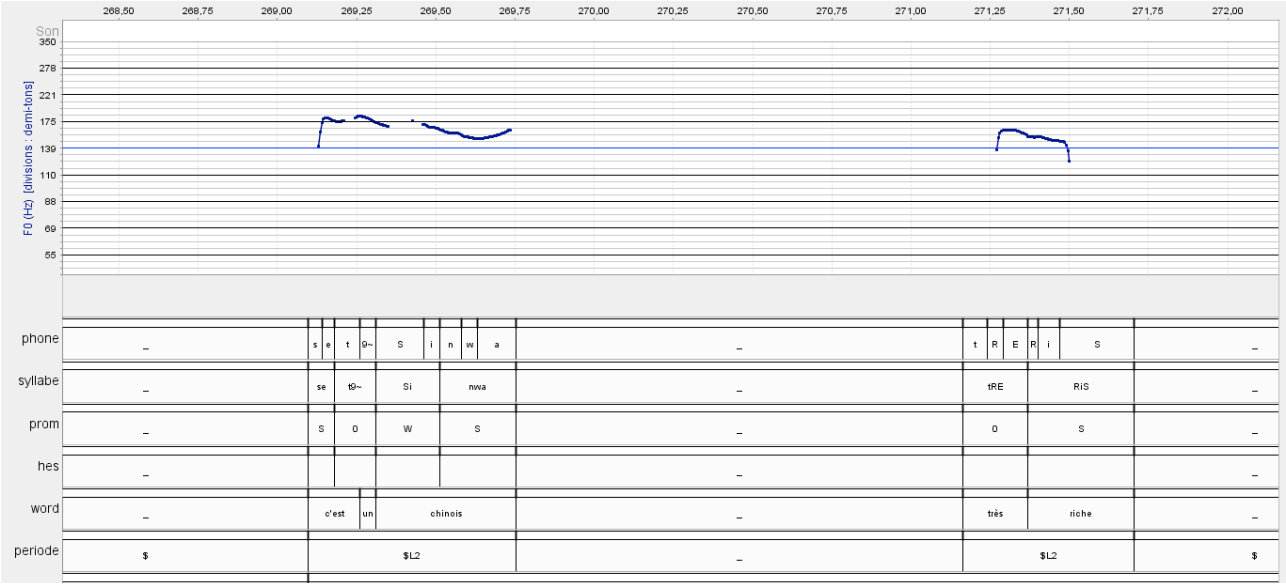
Our modular inductive annotation allowed us to identify different types of syntactic and prosodic units in discourse: GUs, nuclei and IUs, IPas and IPes. As mentioned above, these units operate in principle independently from one another and simultaneously in discourse. In this section, we will examine the types of interactions among units that we identified in our corpus.

5.1 The interaction between IUs and GUs.

Let us begin by observing the interaction between GUs and IUs. We noted above that an IU can be made up of one or more GUs (examples (12) and (13)).

This does not mean that GUs are always included within the limits of an IU: dependency and pile relations can go beyond the limits of an IU (Benzitoun *et al.* 2010; Deulofeu *et al.* 2010). We saw in (3), reproduced here as (39), an example of a GU spanning over an IU:

(39) $L1 *alors < qui vous regarde* //
$L2 *c'est un Chinois* //+ *très riche* // (Rhap-D2001, Mertens)
$L1 then < who is looking at you //
$L2 he is a Chinese man //+ very rich //



The major prosodic break between the two segments *c'est un chinois* and *très riche* (see Fig. 2) organizes the sequence into two IPes and to each of these two IPes corresponds an IU: both the sequence *c'est un chinois* and the sequence *très riche* respond to the tests of nuclearity (they are autonomizable, their illocutionary force can commute with other illocutionary forces). These two sequences are thus two autonomous nuclei. Similarly, the layers of the pile in (11), reproduced here

as (40), are distributed over four dialogical IUs determining another case of a GU spanning over four IUs:

(40) $L1 ^*et il donne* { *à Gaga* |} //+
$L2 {| *à* { *Gago* |} } > *effectivement* //+
$L1 {| *Gago* |} "*pardon*" //+
$L2 {| *Gago* } { *qui est contré* | *qui est contré* } // (Rhap-D2003, Rhapsodie)
$L1 ^and he gives { to Gaga |} //+
$L2 {| to { Gago |} } > actually //+
$L1 {| Gago |} "sorry" //+
$L2 {| Gago } { who is blocked | who is blocked } //

5.2 The interaction between IUs

Let us now observe the interaction between IUs. IUs are not always linearly organized ranged One IU may interrupt another IU, forming a parenthesis, which we annotated as IUs between parentheses: (…//). An example is (41):

(41) "*euh*" *d'autre part* < (*il ne faut pas se mentir* //) *les vacances sont nombreuses* //
(Rhap-M1003, Rhapsodie)
"uh" on the other hand < (let's face it //) there are many holidays //

An IU can also be governed by a word belonging to another IU. Such an embedded IU is marked with square brackets: […//]. The most typical case is reported speech, where a verb of saying governs one or more embedded IUs:

(42) *Marcel Achard écrivait* [*elle est très jolie* // *elle est même belle* // *elle est élégante* //] //
(Rhap-D2001, Mertens)
Marchel Achard wrote [she is very pretty // she is even beautiful // she is elegant //] //

A more general case of embedded IU is a graft (Deulofeu 1999). A *graft* is an IU produced in a governed position, where a noun phrase would be expected. In example (43) for instance, the IU *je crois que c'est une ancienne caserne* 'I think they are old barracks' is governed by the preposition *vers* 'toward' (Figure 12). In other words, an entire IU has been grafted in the place of the noun phrase expected after the preposition *vers*:

(43) { *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 tramline that goes towards { the & | [I think they're old barracks "I think" //] //

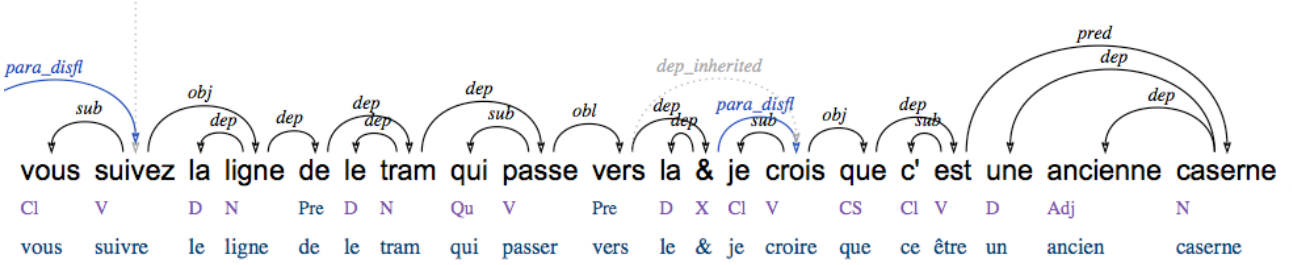


Figure 12. Microsyntactic structure of (35)

5.3 Interaction between prosodic and syntactic units

We saw in section 4.1.2 (example (3)) that the presence of a major prosodic break organizes one GU into two IPes and two IUs. It should be highlighted, though, that the correspondence between prosodic and macrosyntactic units is not always univocal. In a previous study conducted on a small sample of spoken French, not belonging to the Rhapsodie corpus (Lacheret *et al.* 2011), we showed that the correlation between syntactic and prosodic units is strong but not absolute: 65% of IPa boundaries correspond to IU boundaries and 87% of IU boundaries correspond to IPa boundaries.

We showed in the same study that there is a correspondence between the boundaries of IPes and the boundaries of IUs in that usually several IUs are grouped together in one IPe. Even in this respect, however, the correspondence is not perfect. In many cases, the organization in IPes can be determined by performance needs. Speakers may want for example to scan their discourse and to focus on part of it and this may result in a sequence organized in several IPes, regardless of the encoding of the illocutionary information. As an example, let us consider sequence (44) taken from a political speech by President Sarkozy. In this excerpt we see one IU realized by a sole GU (Figure 13) and segmented at the prosodic level into four IPes (Figures 11 and 14).

(44) [*je sais également ce que cela signifie pour vos familles*]
[*que je veux saluer particulièrement*]
[*dont j'imagine qu'elles sont souvent confrontées à l'absence*]
[*et parfois l'angoisse*] (Rhap-M2001, C-PROM)
[I also know what it means for your families]
[whom I want to particularly greet]
[whom I imagine are often confronted with absence]
[and sometimes anguish]

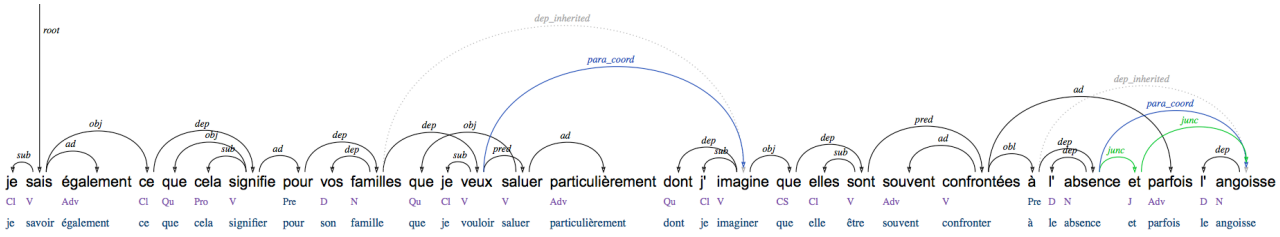
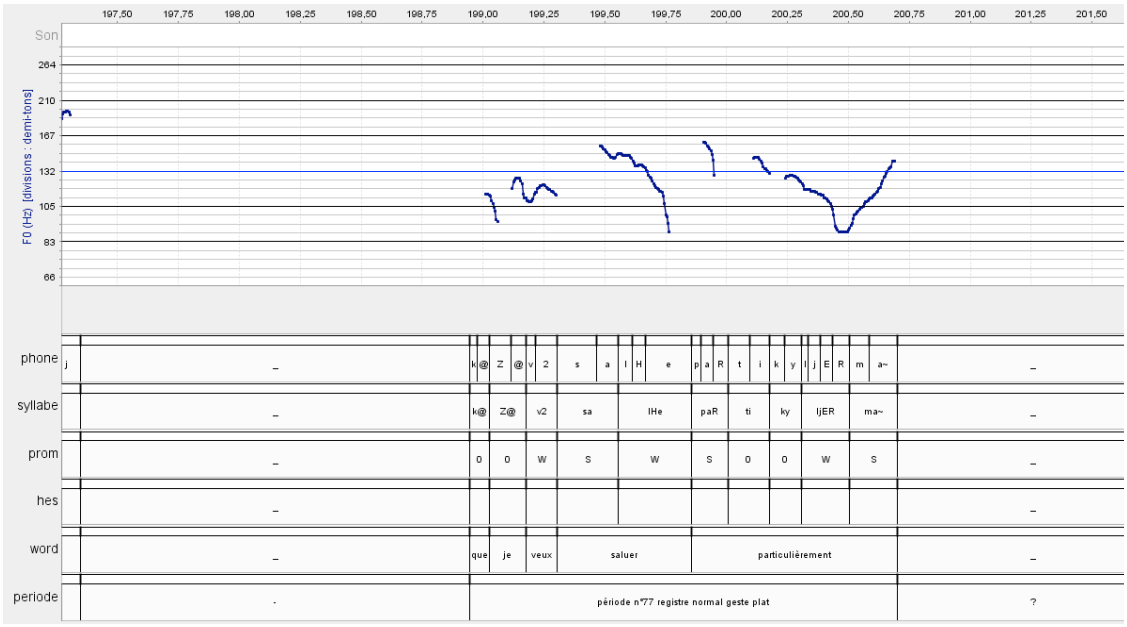
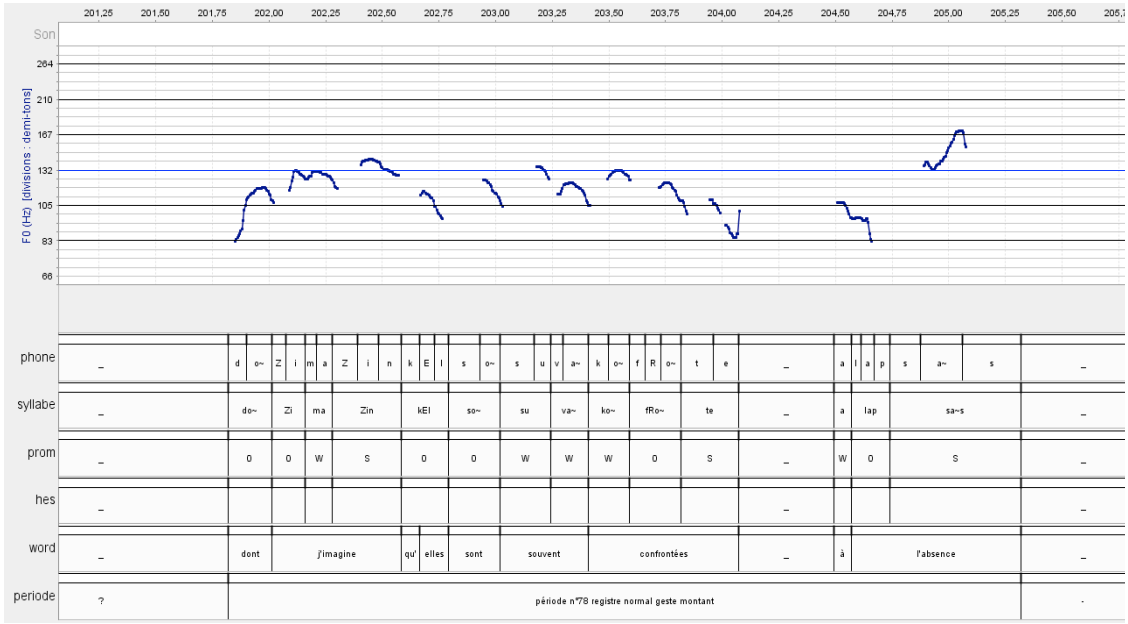


Figure 13. Microsyntactic structure of (44)


Son
F0 (Hz) [divisions : demi-tons]
264
210
167
132
105
83
66
197,50 197,75 198,00 198,25 198,50 198,75 199,00 199,25 199,50 199,75 200,00 200,25 200,50 200,75 201,00 201,25 201,50
phone
syllabe
prom
hes
word
periode
j _ k@ Z @ v 2 s a l H e p a R t i k y l j E R m a~ _
_ k@ Z@ v2 sa lHe paR ti ky ljER ma~ _
_ 0 0 W S W S 0 0 W S _
_ _
_ que je veux saluer particulièrement _
. période n°77 registre normal geste plat ?


Son
F0 (Hz) [divisions : demi-tons]
264
210
167
132
105
83
66
201,25 201,50 201,75 202,00 202,25 202,50 202,75 203,00 203,25 203,50 203,75 204,00 204,25 204,50 204,75 205,00 205,25 205,50
phone
syllabe
prom
hes
word
periode
_ d o~ Z i m a Z i n k E l s o~ s u v a~ k o~ f R o~ t e _ a l a p s a~ s _
_ do~ Zi ma Zin kEl so~ su va~ ko~ fRo~ te _ a lap sa~s _
_ 0 0 W S 0 0 W W W 0 S _ W 0 S _
_ _
_ dont j'imagine qu' elles sont souvent confrontées _ à l'absence _
? période n°78 registre normal geste montant .

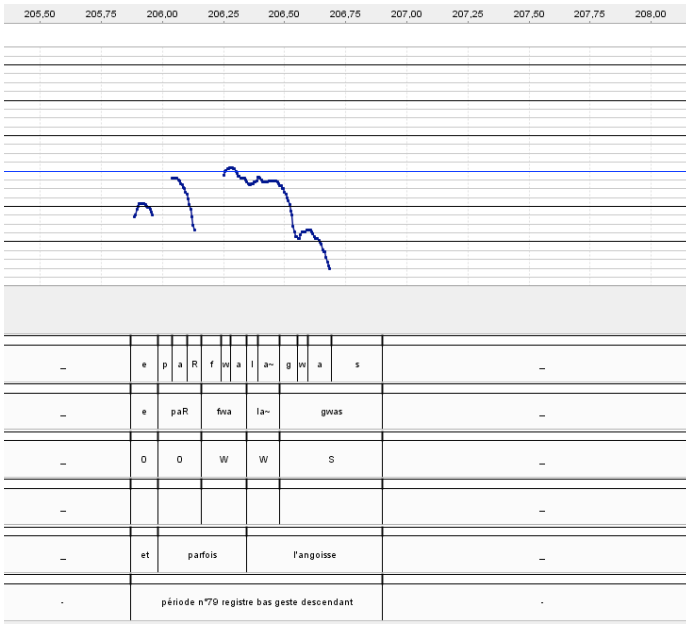


Figure 14. Prosodic structure of the last three IPE of (44)

It is important to highlight that example (44) shows that the prosodic organization may be independent not only of the macrosyntactic, but also and above all of the microsyntactic organization, since a single GU is realized in several IPes.

All in all, our analysis of the interaction between prosodic and syntactic structures questions most of the hypotheses put forward by top-down approaches to the prosody-syntax interface. According to these hypotheses, the mismatch between prosodic and microsyntactic structure when regarded as a consequence of a specific structuring imposed by pragmatic constraints, is explained by the fact that prosodic units necessarily include syntactic components (Nespor & Vogel 1986, Delais-Roussarie 2005, Selkirk 2005). Actually, our data show that this is not always the case and that pragmatic constraints (information processing, topic and focus marking, expressive constructions, etc.) also lead to fragmentation.

6. Conclusions

In our annotation task we identified three separate mechanisms of discursive cohesion: microsyntax, macrosyntax and prosody. These cohesive mechanisms yield three types of maximal units operating in discourse: GUs, IUs, and IPes. Since microsyntax, macrosyntax and prosody operate (for the most part) independently from one another, GUs, IUs and IPes are not necessarily co-extensive.

The study of the interaction between GUs, IUs and IPes led to a redefinition of the notion of "discourse unit", such as proposed by Degand & Simon (2009). By taking into account the interaction of microsyntactic dependency and prosody, Degand and Simon put forward a typology of discourse units. We have extended this notion by also taking into account macrosyntax and piling phenomena.

Namely, we propose the notion of *extended discourse unit* (EDU). By EDU we mean a sequence characterized by the fact that its components are linked to one another by at least one of the three mechanisms of syntactic or prosodic cohesion identified above.

Let us take as an example (45):

(45) ["*eh ben*" "*euh*" *tu prends* { *le boulevard* "*euh*" *là qui part de Nef Chavant* | *là le boulevard qui passe à côté d'Habitat* } // $||_{\text{IPe}}$]$_{\text{EDU}}$ [*tu continues* // $||_{\text{IPe}}$]$_{\text{EDU}}$

["well" "uh" you take {the boulevard "uh" that starts in Nef Chavant | the boulevard that passes close to Habitat } // ||IPe]EDU [you go on // ||IPe]EDU

The sequence between the words *eh ben* and the word *Habitat* is connected through macrosyntactic relations connecting the associated nuclei "eh ben", "euh" to the nucleus of the IU, *tu prends le boulevard là qui part de Nef Chavant* and microsyntactic relations between the piling of the two objects *le boulevard là qui part de Nef Chavant | là le boulevard qui passe à côté d'Habitat* ; besides, the sequence realizes one and only one IPe. After the word *Habitat* there is a break in all the cohesion mechanisms (prosodic, microsyntactic, and macrosyntactic): in other words we encounter, after the word *Habitat,* a boundary which is at the same time a GU, an IU and an IPe boundary. We can say, thus, that the sequence in (45) is organized into two EDUs *et ben euh tu prends le boulevard euh là qui part de Nef Chavant là le boulevard qui passe à côté d' Habitat* and *tu continues.*

In (46), the sequence between the words *je* and *angoisse* constitutes one EDU. Unlike the first EDU of (45), the cohesion of this EDU is not guaranteed by prosody (the sequence is indeed organized into four distinct IPes) but by microsyntactic and macrosyntactic relations: the dependency and piling links between the words of the sequence make it a single GU realizing one and only one IU. After the word *angoisse* we have a boundary of IPe, GU and IU, i.e., an EDU boundary: the sequence in (46) is therefore analyzed as two distinct EDUs, as shown by the annotation.

(46) [je sais également ce que cela signifie pour vos familles ||IPe { que je veux saluer particulièrement ||IPe dont j' imagine qu' elles sont souvent confrontées à { l' absence ||IPe ^et parfois l' angoisse } } // ||IPe]EDU [je sais aussi "hélas" le { lourd tribut payé par certains de vos compagnons d' armes | tribut qui peut aller jusqu' au sacrifice ultime } // ||IPe]EDU

[I also know what it means for your families ||IPe { whom I want to particularly greet | ||IPe whom I imagine are often confronted with { absence |||IPe ^ and sometimes anguish} } // ||IPe]EDU [I also know "unfortunately" {the heavy tribute paid by some of your comrades in arms | a tribute that involved paying the ultimate sacrifice } // ||IPe]EDU

Quite interestingly, having posited the notion of EDU, this enables us to reappraise and to provide a new definition for the traditional notion of sentence. Indeed we found in our corpus a number of EDUs characterized by the fact that they were realized by one and only one IU, realizing one and only one GU, headed by a verb, and included in one and only one IPe :

(46) *le lycée Voltaire est un bon lycée* // (Rhap-D2002, Rhapsodie)
Voltaire high school is a good school //

(47) *il y en a des moins bons* // (Rhap-D2002, Rhapsodie)
there are some that are less good //

(48) *ils ne parlent jamais français* // (Rhap-D2002, Rhapsodie)
they never speak French //

It is easy to see that each of these sequences corresponds to what is commonly called “a sentence”. In other words, we might say that what is commonly called a sentence can be regarded as but a particular case of an extended discourse unit whose cohesion is guaranteed at the same time by prosody, microsyntax and macrosyntax and which is microsyntactically governed by a verb. This type of EDU is rare, but not absent from our corpus.

It is important to highlight, though, that in spite of the fact that “sentences” do exist in discourse, they do not deserve a special epistemological status: they are only one type of EDUs among others

and by no means should such a particular case of EDU be considered as the rule from which all other types of EDUs deviate, nor as a viable unit for the annotation of spoken corpora.

Appendix

//	End of an illocutionary unit (IU)
<	End of a pre-nucleus
>	Beginning of a post-nucleus
( )	Beginning and end of an in-nucleus
(//)	Beginning and end of a parenthetical IU
+	Indicates the continuation of a governed unit. This symbol is always combined with a macrosyntactic tag: //+ or <+ or >+ or (+
" "	Beginning and end of an associated nucleus
^	IU opener and pile marker
[//]	Beginning and end of an embedded IU
#	Indicates a discontinuity in a governed unit
&	Indicates an unfilled governed position.

The symbols used in the macrosyntactic annotation

References

't Hart J., Collier R., Cohen A. (1990*). Perceptual Study of Intonation: An Experimental-Phonetic Approach to Speech Melody*. Cambridge: Cambridge University Press.

Aijmer K. (2002), *English Discourse Particles, Evidence from a Corpus*, Amsterdam/ Philadelphia, Benjamins.

Andersen H. L., Nølke H. (eds.) (2002). *Macro-syntaxe et macro-sémantique, Actes du colloque international d'Århus, 17-19 mai 2001*. Bern: Peter Lang.

Avanzi M., Lacheret A., Victorri B. (2008). Analor, a Tool for Semi-automatic Annotation of French Prosodic Structure, *Speech Prosody 2008*, Campinas, Brazil, 119-122.

Bazzanella C. (1995). I segnali discorsivi. In L. Renzi, G. Salvi, A. Cardinaletti (eds.), *Grande grammatica italiana di consultazione*, vol. III, Bologna, Il Mulino, 225-257.

Beckman M. E., Elman G. A. (1997). Guidelines for ToBi Labelling, version 3, The Ohio State University Research Foundation.

Benzitoun C., Dister A., Gerdes K., Kahane S., Pietrandrea P., Sabio F. (2010). Tu veux couper là faut dire pourquoi. Propositions pour une segmentation syntaxique du français parlé. *Actes du Congrès Mondial de Linguistique Française (CMLF 2010)*, New Orleans.

Berrendonner, A. (1990). Pour une macro-syntaxe. *Travaux de linguistique, 21*, 25-31.

Berrendonner A. (2002). Les deux syntaxes, *Verbum, 1-2*, 23-35.

Berrendonner A. (2011). Unités syntaxiques & unités prosodiques, *Langue Française, 170*, 81-93.

Blanche-Benveniste, Cl., Borel B., Deulofeu J., Durand J., Giacomi A. , Loufrani, Cl., Meziane B. et Pazery N. (1979) Des grilles pour le français parlé. *Recherches sur le français parlé, 2*, 163–205.

Blanche-Benveniste, Cl. (1990). Un modèle d'analyse syntaxique « en grilles » pour les productions orales. *Anuario de Psicologia, 47*, 11-28.

Blanche-Benveniste, Cl., Bilger M., Rouget Ch. et Van den Eyende K. (1990). *Le français parlé. Etudes grammaticales*. Paris: Editions du Centre National de la Recherche Scientifique.

Blanche-Benveniste, Cl., (2002) Phrase et construction verbale, *Verbum 1-2*, 7-22.

Bohmová, A., Hajič J., Hajičová E., Hladká B. (2003). The PDT: a 3-level annotation scenario. In A. Abeillé (ed.) *Treebanks: Building and Using Parsed Corpora*. Kluwer Dordrecht, 103–127.

Bolinger D. (1968). *Aspects of Language.* New York-Chicago, Harcourt, Brace, Jovanovich.

Bonvino E., Masini F., Pietrandrea P. (2009). List Constructions: a semantic network. *Troisième Conférence Internationale de l'AFLiCo*, Nanterre. Accessible à http://francescamasini.caissa.it/Presentations_files/parigi_draft.pdf.

Bourigault D. (2007). *Un analyseur syntaxique opérationnel : SYNTEX*, Habilitation à Diriger les Recherches, Université Toulouse-Le Mirail.

Brown P., Levinson S. (1978). Universals in Language Use: Politeness Phenomena. In E. Goody (ed.), *Questions and Politeness: Strategies in Social Interaction*, Cambridge, Cambridge University Pres, 56-310.

Chafe, W. (1998) Language and the Flow of Thought. *The New Psychology of Language*, M. Tomasello (ed.), New Jersey, Lawrence Erlbaum Publishers, 93-111.

Cresti, E. (2000). *Corpus di italiano parlato*. Florence: Accademia della Crusca.

Cresti E. (2005) Enunciato e frase. Teoria e verifiche empiriche. In Biffi M., Calabrese O., Salibra L. (eds.) *Italia Linguistica: discorsi di scritto e di parlatoEnunciato e frase: teoria e verifiche empiriche, Scritti in onore di Giovanni Nencioni*, Prolagon, Siena.

Degand L., A. C. Simon, On identifying basic discourse units in speech: theoretical and empirical issues, *Discours* , 4 | 2009 URL : http://discours.revues.org/5852 ; DOI : 10.4000/discours.5852

Dehé, N., Kavalova, Y. (2006). The syntax, pragmatics, and prosody of parenthetical *what*, *English Language and Linguistics*, *10*, 289-320.

Delais-Roussarie, E. (2005). *Phonologie et Grammaire: Etudes et modélisation des interfaces prosodiques*, Habilitation à Diriger des Recherches, Université de Toulouse-le Mirail.

Deulofeu, J. (1999). *Recherches sur les formes de la prédication dans les énoncés assertifs en*

français contemporain (le cas des énoncés introduits par le morphème que). Thèse d'état, Université Paris 3.

Deulofeu J., Dufort L., Gerdes K., Kahane S., Pietrandrea P. (2010). Depends on what the French say: Spoken corpus annotation with and beyond syntactic function, *4th Linguistic Annotation Workshop (LAW IV),* ACL, Uppsala, 274-281.

Gerdes K. (2013). Collaborative Dependency Annotation, *Proceedings of Depling*, 88–97, Prague.

Gerdes K., Kahane S. (2009). Speaking in Piles. Paradigmatic Annotation of a Spoken French Corpus. *5th Corpus Linguistics Conference*, http://ucrel.lancs.ac.uk/publications/cl2009, Liverpool, 15 p.

Groupe de Fribourg. (2012). *Grammaire de la période*. Berne, Peter Lang.

Hajič J. (1998). Building a syntactically annotated corpus: The Prague Dependency Treebank. In E. Hajičová (ed*.) Issues of Valency and Meaning, studies in honour of Jarmila Panevová,* 106-132. Prague: Karolinum.

Hasegawa-Johnson M., Chen K., Cole J., Borys S., Kim S., Cohen A., Zhang T., Choi J., Kim, H., Yoon T., Chavarria S. (2005). Simultaneous Recognition of Words and Prosody in the Boston University Radio Speech Corpus, *Speech Communication, 46*(*3-4*), 418-439

Holmes J. (1986). Functions of you know in women's and men's speech*, Language in Society, 15*, 1-21.

Kahane, S. (2012). De l'analyse en grille à la modélisation des entassements. In S. Caddeo, M.-N. Roubaud, M. Rouquier, F. Sabio, *Penser les langues avec Claire Blanche-Benveniste*, Presses de l'université de Provence, 101-116.

Kahane, S. (2013). *Tutoriel codage microsyntaxique*, http://www.projet-rhapsodie.fr.

Kahane S., Pietrandrea P. (2012a) La typologie des entassements en français*, Actes du 3ème congrès mondial de linguistique française (CMLF),* Lyon, 1809-1828.

Kahane S., Pietrandrea P. (2012b). Les parenthétiques comme « Unités Illocutoires Associées » : une perspective macrosyntaxique. *Linx, 61,* 49-70.

Kärkkäinen, E. (2003), *Epistemic stance in English conversation. A description of its interactional functions, with a focus on I think*, Amsterdam, Benjamins.

Kleiber, G., 2003, Faut-il dire *adieu* à la phrase ?, *L'information grammaticale*, *98*, 17-22.

Lacheret A., Beaugendre F. (1999). *La prosodie du français*, Paris, Editions du CNRS.

Lacheret A. (2003). *La prosodie des circonstants en français parlé,* Paris-Leuven, Peeters.

Lacheret A., Kahane S., Pietrandrea P., Avanzi M., Victorri B. (2011). Oui mais elle est où la coupure, là? Quand syntaxe et prosodie s'entraident ou se complètent. *Langue Française,*

170, 61-80.

Loufrani C. (1984). Le locuteur collectif. Typologie de configurations discursives. *Recherches sur le français parlé, 6*, 169-193.

Mel'čuk, I. (1988). *Dependency Syntax: Theory and Practice,* SUNY Press, New-York.

Miller, J. and Weinert, R. (1998/2009). Spontaneous spoken language. Syntax and discourse. Oxford: Oxford University Press.

Moneglia, M. (2011) Spoken corpora and pragmatics. RBLA, Belo Horizonte, v.11 (2) : 479-519.

Nespor, M., Vogel, I. (1986) *Prosodic Phonology*, Foris, Dordrecht.

Nivre J. (2008). Treebanks, in Lüdeling, Anke / Kytö, Merja (eds.) Corpus Linguistics, Mouton de Gruyter 225-24.

Nølke H., Adam J.M., (eds) (1999). Approches modulaires, de la langue au discours, Delachaux et Niestlé, Lausanne.

Nølke, H. (1990). Recherches sur les adverbes : bref aperçu historique des travaux de classification, *Langue française, 88*, 117-122.

Ostendorf M., Shafran I., Shattuck-Hufnagel S., Carmichael L., Byrne W. (2001). A prosodically labeled database of spontaneous speech. In *Proceedings ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding,* Red Bank, NJ.

Östman J.-O. (1981). *'You know'. A Discourse-functional Approach.* Amsterdam/ Philadelphia, Benjamins.

Ross J. R. (1973). Slifting. In M. Gross, M. Halle, M. P Schützenberger (eds.) *The Formal Analysis of Natural Language*, Berlin, Mouton, 133-169.

Rossi, M. (1979). Le français, langue sans accent ? In Fonagy, I., Léon, P. (Eds.), L'accent en français contemporain. *Studia Phonetica* 15, Didier, Paris, pp. 13-51.

Roulet, E. L.Filliettaz. A. Grobet et M. Burger (2001). *Un modèle et un instrument d'analyse de l'organisation du discours*, Bern, Peter Lang (Collection Sciences pour la communication).

Sabio F. (2006). Phrases et constructions verbales : quelques remarques sur les unités syntaxiques dans le français parlé. In D. Lebaud, C. Paulin, K. Ploog (eds.), *Constructions verbales et production de sens*, Presses Universitaires de Franche-Comté, Besançon.

Schelfhout C., Arno Coppen P., Oostdijk N. (2004). Finite Comment Clauses in Dutch: A Corpus-based Approach, *Journal of Germanic Linguistics*, *16*, 331-349.

Schiffrin D. (1987). *Discourse Markers*, Cambridge, Cambridge University Press.

Selkirk, Elisabeth. 2005. Comments on intonational phrasing. In S. Frota, M. Vigario, M. J. Freitas (eds.), *Prosodies*, Berlin: Mouton de Gruyter, 11-58.

Villemonte de La Clergerie E. (2005) DyALog: a tabular logic programming based environment for NLP. *2nd International Workshop on Constraint Solving and Language Processing* (*CSLP'05*), Barcelona, Spain.

Resources

Avanzi M. (2012), *L'interface prosodie/syntaxe en français : Dislocations, incises et asyndètes*, Bruxelles, Peter Lang

Avanzi, M., Simon, A.C., Goldman, J.-P. & A. Auchlin. (2010), C-PROM. Un corpus de français parlé annoté pour l'étude des proéminences, *Actes des 23èmes journées d'étude sur la parole* (Mons, Belgique, 25-28 mai 2010)

Branca-Rosoff S., Fleury S., Lefeuvre Fl., Pires M (2012), *Discours sur la ville. Corpus de Français Parlé Parisien des années 2000 (CFPP2000)* http://cfpp2000.univ-paris3.fr/

Durand, J., Laks, B. & Lyche, C. (2009), Le projet PFC (phonologie du français contemporain): une source de données primaires structurées, in : J. Durand, B. Laks & C. Lyche (eds.), *Phonologie, variation et accents du français*. Hermès, Paris, 19-61.

Eshkol-Taravella I., Baude O., Maurel D., Hriba L., Dugua C., Tellier I., (2012), Un grand corpus oral « disponible » : le corpus d'Orléans 1968-2012., in *Ressources linguistiques libres*, TAL. Volume 52 – n° 3/2011, 17-46

Lacheret A. (2003), *La prosodie des circonstants*, Leuven Peeters.

Mertens P. (1987) , *L'intonation du français : de la description linguistique à la reconnaissance automatique,* Thèse de Doctorat, Université de Louvain