Intonosyntactic Data Structures : The Rhapsodie Treebank of Spoken French

Kim Gerdes, Sylvain Kahane, Anne Lacheret, Arthur Truong, Paola Pietrandrea

▶ **To cite this version:**

Kim Gerdes, Sylvain Kahane, Anne Lacheret, Arthur Truong, Paola Pietrandrea. Intonosyntactic Data Structures : The Rhapsodie Treebank of Spoken French. Sixth Linguistic Annotation Workshop (LAW VI) held in conjunction with ACL-2012, 2012, Jeju, South Korea. halshs-01509071

HAL Id: halshs-01509071
https://shs.hal.science/halshs-01509071

Submitted on 19 Sep 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Intonosyntactic Data Structures: The Rhapsodie Treebank of Spoken French

Kim Gerdes
LPP, Sorbonne nouvelle & CNRS /
NLPR, Institute of Automation, Chinese Academy of Sciences
kim@gerdes.fr

Sylvain Kahane
Anne Lacheret
Arthur Truong
Modyco, Université Paris Ouest Nanterre & CNRS
sylvain@kahane.fr
anne@lacheret.fr
arthur_truong@hotmail.com

Paola Pietrandrea
Modyco, Université Paris Ouest Nanterre & CNRS / Lattice, CNRS
paolapietrandrea@gmail.com

Abstract

In this work, we present the data structures that were developed for the Rhapsodie project, an intonosyntactic annotation project of spoken French. Phoneticians and syntacticians work on different base units: a time aligned sound file for the former, and a partially ordered list of tokens for the latter. The alignment between the sound-file and the tokens is partial and non-trivial. We propose to encode this data with a small set of interconnected structures: lists, constituent trees, and directed acyclic graphs (DAGs). Our query language remains simple, similar to the Annis Query language, as the precedence and including relations are handled in accordance with the requested objects and their type of alignment: The order between prosodic units is time-based, whereas the order between syntactic units is lexeme-based.

1 Introduction

Our work results from a corpus development project of Spoken French, Rhapsodie, set up for the study of the syntax-prosody interface. An intonosyntactic corpus has to allow corpus-based studies on the relation between prosody and syntax, which implies the recognition of prosodic structures, syntactic structures, and the relation between them. In spite of the abundance of work on treebanks, very few attempt to annotate spoken language, and even less spontaneous speech. We are aware of the Switchboard Corpus (Meteer et al. 1995), which is annotated with phrase structures, disfluencies, and illocutionary acts; the CHRISTINE Treebank (Rahman & Sampson 2000) is annotated with phrase structure, like the British component of the International Corpus of English (Nelson et al. 2002); the treebanks of English, German, and Japanese, created within the VERBMOBIL project (Hinrichs et al. 2000) have the Negra-style (Brants 2000) mixed annotation of functionally augmented constituent structures, the Venice Italian Treebank (Delmonte 2009) annotated with dependency and phrase structures, the Ester treebank for French (Cerisara et al. 2010, dependency annotation on radio transcripts), the CNG (Spoken Dutch Corpus; Schuurman et al. 2004, dependency annotation on spontaneous speech, skipping over disfluencies), the Hong Kong corpus (Cheng et al. 2008, prosodic annotation of prominence, tone, key and termination). We may add to this list the C-Oral Rom Corpus (Cresti and Moneglia 2005). The C-Oral Rom does not include an annotation of syntactic constituency or syntactic relations (what we call microsyntactic annotation), but is endowed with a rich macrosyntactic annotation (see below for the micro/macrosyntactic distinction).

1.1 Data-structures

The commonly used structures of spoken data annotation do not allow any complex syntactic annotation and inversely, syntactic formats are difficult to adapt or link to spoken data, because the nature of the data is profoundly different: Syntax is commonly based on a chain of lexemes and spoken data annotation consists of classifying time-aligned segments of the sound-file. Spontaneous spoken language has the additional twist that we have multiple tiers of partially overlapping speech. In this article we present the complete data structure of an intono-syntactic corpus of spontaneous spoken French. We show the complex relations that exist between prosodic and syntactic units and

how we resolved multiple problems arising in the complex process of partial alignment. The goal is to obtain a non-redundant data-structure where the same syntactic or prosodic units can be part of different tree-structures, resulting in a highly complex acyclic graph structure as the common base structure of linguistic annotation. We describe the XML import and export format and the internal SQL representation of the data structure as well as the query language allowing for cross domain queries between syntax and prosody.

The Rhapsodie corpus is the result of a four-year project funded by the French National Research Agency (ANR). The corpus is made up of 57 samples of spoken French (5 minutes on average) mainly drawn from existing corpora of spoken French for a total of 3 hours and 33 000 words and distributed under a Creative Commons licence at http://www.projet-rhapsodie.fr. It synthesizes and formalizes various approaches to the syntactic and prosodic analysis of spoken French, in particular research stemming from the Aix school (Blanche-Benveniste et al. 1990, Deulofeu et al. 2010). The project provides a treebank endowed with both:

- a complete microsyntactic and macrosyntactic annotation (microsyntax: morpho-syntactic and functional (dependency) annotation; macrosyntactic: illocutionary groupings of maximal microsyntactic units, including discourse markers, dislocations, reported speech, parentheses, etc.)
- a rich prosodic annotation including perceptually identified phenomena such as prominences, breaks, disfluencies; phonetic alignments; detailed acoustic measurements; a large range of melodic contours; and annotation of prosodic macrostructures.

Moreover, a number of tools necessary to conduct a complete analysis at the interface of prosody, microsyntax and macrosyntax are provided.

1.2 Existing tools for the annotation of spoken language

The tools commonly used for editing prosodic transcription and aligning them to the signal (Praat (Boersma, Paul & Weenink 2012; Delais-Roussarie et al. 2003), WinPitch (Martin 2000), Exmeralda (Schmidt 2004)) allow for different segmentations of the same sample: different types of segments are stored in different independent tiers. This tier-based approach can simulate constituent structures by time-aligning bigger segments in one tier with smaller segments in another tier. Yet, this does not allow for an explicit encoding of constituent structure, because one segment cannot be linked to another segment. Therefore, neither constituency based nor dependency based syntactic structures can be described in the commonly used tools for prosodic annotation.

1.3 Treebank query tools

Inversely, the plenitude of tools that have been developed for tree-banking (visualization, annotation, correction, and search) are all token based. A well-known versatile tool is Annis (Zeldes et al. 2009) which allows for import, visualization, and search of various annotation formats and multiple annotations of the same text with segments, constituent trees, and dependency trees, all stored in a united XML format called Paula (Chiarcos et al. 2008). Annis is completely token based and although the tokens can be time-aligned, the Paula format is not well-adapted to spoken data, because we would have to choose the phonemic transcription as the base units and define the tokens (called *markables* in Paula) on this base transcription. This implies that all precedence relations are symbolic and not time-based and all order relations are based on the most fine-grained list of tokens. Moreover, as we will see, all lexemes cannot be decomposed into phonemes and the set of lexemes needs an independent (partial) order relation.

2 Linguistic annotation

2.1 Syntactic annotation

We have annotated two cohesive levels of syntax: microsyntax and macrosyntax.

Microsyntax describes the syntactic relations which are usually encoded through dependency trees or phrase structure trees. These relations are annotated in all the major syntactic treebanks such as the Penn Treebank, the Prague Treebank, the French Treebank, the Copenhagen Dependency Treebank, etc.

Macrosyntax can be regarded as an intermediate level between syntax and discourse. This level describes and classifies the sequences that make up one and only one illocutionary act as well as the relations holding between these sequences. We

have identified the macrosyntactic structure of our corpus on purely discursive and syntactic arguments whereas in the C-Oral Rom corpus, macrosyntactic units are regarded as functional interpretations of prosodic units (Cresti 2005).

The annotation of macrosyntax is essential to account for a number of cohesion phenomena typical of spoken discourse and in particular of French spoken discourse, because of the high frequency of paratactic phenomena that characterize this language. See for example (1)

(1) moi < ma mère < le salon < c'est de la moquette //
me < my mother < the living room < it's carpet
'My mother's living room is carpeted'

The microsyntactic and macrosyntactic phenomena have been encoded independently from one another in a modular, partially computer-aided approach relying on collaborative online tools (Deulofeu et al. 2010). The annotation provides an analysis of all linguistic utterances of the samples and includes a complete annotation and a functional tagging of what we call *pile* structures: By piles we intend the multiple realization of the same structural position, which occurs in coordination (2), reformulation (3), disfluency (4), and correction (5) phenomena (Gerdes & Kahane 2009):

(2) nous avons été sous très gros bombardements { américains | ^**puis** anglais } // (D003)
we have been under very heavy bombing { American | ^then English }

(3) tu arrives place aux Herbes avec { une | une } sorte **{ de halle | "quoi" { de |de |de } structure métallique }** // (M001)
you arrive Square of Herbs with {a | a } kind { of hall | "like" {of | of | of } metalic structure }

(4) alors < **{ { j'a~ | j'avais } beaucoup | j'avais beaucoup }** trop peur de m'installer (comme ça) seule **{ d~ | dans }** la brousse // (D204)
well < {{ I had | I had } too | I had too } much fear to settle (like that) alone {i~ | in } the jungle

(5) c'est la crise générale { { **des | des } Français** |} //+ {(**"enfin" des Français**//) | (pas simplement des Français "hein"//) | { { des | de } l'humanité | ^et de la lecture } } // (D004)
it's the general crisis {{ of | of } the Frenchmen|} //+ {("well" of the Frenchmen//) | (not simply of the Frenchmen "huh"//) | {{ of the | of } humanity | ^and of reading }}//

Albeit extremely frequent in spoken language, pile relations, which can be seen as a particular type of microsyntactic relation, are often disregarded in corpus annotation. By extensively annotating and tagging pile phenomena we could guarantee an exhaustive microsyntactic annotation of all our data, including disfluencies, repetitions, reformulations generally considered as performance errors and commonly not analyzed in spoken languages treebanks (see for instance the CNG).

In more general terms, we have provided a complete categorical and functional tagging for every word of the corpus, including discourse markers, which are integrated into the syntactic representation at the macrosyntactic level.

2.2 Prosodic annotation

As for prosody, we built on the theoretical hypothesis formulated by the Dutch-IPO school ('t Hart et al. 1990) stating that, out of the whole of 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 only three perceptual phenomena characterizing real productions: prominences, pauses and disfluencies (Avanzi et al. 2010, Smith 2011).

We have annotated perceptual syllabic salience in speech context by using a gradual labeling distinguishing between strong, weak, and zero prominences. Strong prominences mark intonation packages and weak prominences mark rhythmic groups. Metrical feet are marked by prominences outside words.[1] Periods (Lacheret-Dujour & Victorri 2002) are ended by an occurrence of a pause of at least 300 ms, detection of an F0 pitch movement reaching a certain amplitude and of a "jump" and the absence of disfluency or a "uh" in the vicinity of the pause.

Various studies have shown the usefulness of seeing prosody as a tree structure (Tepperman & Narayanan 2008, Gibbon 2003) consisting of prosodic constituents of different levels. Building on the syllabic salience labeling, we were able to generate the totality of the prosodic tree structure made up of a hierarchy of prosodic segments characterized by more or less prominent frontiers. We

[1] Words are inflected forms of lexemes or amalgams of two lexemes (see section 3.2). A word has an orthographic, a phonetic, and a syllabized form. For example in *une extraordinaire aventure* 'an extraordinary adventure', the word *extraordinaire* has the phonetic form /ekstrordinEr/ and the syllabized form /nekstrordinE/, due to the liaisons between the words. Words are linked to two types of child nodes: the phonemes and the syllables, providing two, possibly different, time alignments. These links may not be one to one: In the example table 1, we see the word *il* 'it' and *y* 'there' sharing the same phoneme.

<table>
<tr><th>per</th><td colspan="30">do~ k@ ja a y n2 z9n Fi j@ a bi je tu ta~ nwaR</td></tr>
<tr><th>pkg</th><td colspan="15">do~ k@ ja a y n2 z9n Fi</td><td colspan="9">j@ a bi je tu</td><td colspan="6">ta~ nwaR</td></tr>
<tr><th>rhg</th><td colspan="4">do~ k@</td><td colspan="6">ja a y n2</td><td colspan="5">z9n Fi</td><td colspan="9">j@ a bi je tu</td><td colspan="6">ta~ nwaR</td></tr>
<tr><th>feet</th><td colspan="4">do~ k@</td><td colspan="6">ja a y n2</td><td colspan="5">z9n Fi</td><td colspan="3">j@ a</td><td colspan="6">bi je tu</td><td colspan="6">ta~ nwaR</td></tr>
<tr><th>pree</th><td colspan="2"></td><td colspan="2">W</td><td colspan="3"></td><td></td><td colspan="2">W</td><td colspan="3"></td><td colspan="2">S</td><td colspan="2"></td><td>W</td><td colspan="2"></td><td colspan="2"></td><td colspan="2">W</td><td colspan="2"></td><td colspan="4"></td></tr>
<tr><th>syl</th><td colspan="2">do~</td><td colspan="2">k@</td><td colspan="3">ja</td><td>y</td><td colspan="2">n2</td><td colspan="3">Z9n</td><td colspan="2">Fi</td><td colspan="2">j@</td><td>a</td><td colspan="2">bi</td><td colspan="2">je</td><td colspan="2">tu</td><td colspan="2">ta~</td><td colspan="4">nwaR</td></tr>
<tr><th>ph</th><td>d</td><td>o~</td><td>k</td><td>@</td><td colspan="2">j</td><td>a</td><td>y</td><td>n</td><td>2</td><td>Z</td><td>9</td><td>n</td><td>F</td><td>i</td><td>j</td><td>@</td><td>a</td><td>b</td><td>i</td><td>j</td><td>e</td><td>t</td><td>u</td><td>t</td><td>a~</td><td>n</td><td>w</td><td>a</td><td>R</td></tr>
<tr><th>w/ort</th><td colspan="3">donc</td><td>euh</td><td>il</td><td>y</td><td>a</td><td colspan="3">une</td><td colspan="3">jeune</td><td colspan="3">fille</td><td>euh</td><td colspan="5">habillée</td><td colspan="3">tout</td><td>en</td><td colspan="4">noir</td></tr>
<tr><th>w/ph</th><td colspan="3">do~k</td><td>@</td><td colspan="2">j</td><td>a</td><td colspan="3">yn2</td><td colspan="3">Z9n</td><td colspan="3">Fij</td><td>@</td><td colspan="5">abije</td><td colspan="3">tut</td><td>a~</td><td colspan="4">nwaR</td></tr>
<tr><th>w/syl</th><td colspan="3">do~</td><td>k@</td><td colspan="2"></td><td>ja</td><td colspan="3">yn2</td><td colspan="3">Z9n</td><td colspan="3">Fi</td><td>j@</td><td colspan="5">abije</td><td colspan="3">tu</td><td>ta~</td><td colspan="4">nwaR</td></tr>
<tr><th>english</th><td colspan="3">so</td><td>uh</td><td colspan="3">there is</td><td colspan="3">a</td><td colspan="3">young</td><td colspan="3">girl</td><td>uh</td><td colspan="5">dressed</td><td colspan="3">all</td><td>in</td><td colspan="4">black</td></tr>
</table>

Table 1: Structure of the phonetic and orthographic tiers:
per:period, pkg: packages, rhg: rhythmic groups, feet: metrical feet, pree: preeminences (on syllables), syl: syllables, ph: phonemes, w: word (with three forms: ort: orthographic, ph: phonetic, syl: syllabized)

identified global macro-prosodic units called periods, which are iteratively divided into intonational packages, rhythmic groups, feet, and syllables.

We generate prototypical-stylized melodic contours for all prosodic and syntactic units. The availability in the Rhapsodie Treebank of these various contours will allow the user to build various lexicons of intonational contours in an extremely flexible way according to his or her research goals. In more general terms, it should be highlighted that these annotation choices have allowed us on the one hand to identify the primitives of prosodic structure independently from any reference to syntax or pragmatics, and, on the other hand, to provide all the elements needed for a complete prosodic analysis of linguistic units.

3 Formal properties of the structure

3.1 Prosodic structure

Our prosodic structure consists of a hierarchy of segments of various levels: In the general case, a sample consists of speech turns that are segmented into periods. The periods are composed of prosodic packages that in turn are divided into intonational packages, which are divided into rhythmic groups. The rhythmic groups have two types of incompatible segmentations: feet and words. However, each foot and word is composed of syllables. The syllables are the smallest prosodic units, they are composed of phonemes and can combine to form (the syllabized form of) words.[2]

Table 1 shows the decomposition of a period in its prosodic components.

This structure corresponds to a non-recursive constituency-like representation of prosody (as the number of levels is predetermined in our prosodic model), however, the fact that feet and phonological words are segmentations of the same level implies that we have in fact two constituent trees, one including feet, the other including words. All other nodes are shared between the two constituent trees. Thus, our structure should be seen as a constituent DAG rather than a set of constituent trees.

The terminal nodes of those structures are generally phonemes, but the structures have terminal nodes at various levels, because pauses are not further developed, for example a pause between two rhythmic groups, is not developed all the way down to the foot level.

Another complication stems from the segmentation of the sound into speech turns, where in the case of overlaps (i.e. two people speaking at the same time), for technical reasons, a segmentation into units with a higher granularity is possible for at most one of the speakers. If one of the speakers is analyzable, this segment is handled just as any other non-overlapping part where the periods and subsequent segments are children of the analyzable speech turn. The unanalyzable parts of the overlaps have no further segmentation into finer grained segments. In case the sound of both speakers cannot be further analyzed, the unanalyzable segments of both speakers will share the same "overlap" node, which again gives a DAG structure and not a simple tree.

Each point of the time line is thus included in at most one element of each level (or is exactly at the border of two) with the exception of speech turns, where overlaps can occur.

[2] In the example table, *fille* /Fij/ 'girl' is monosyllabic, but due to liaison with the following word, /j/ forms a syllable with the following vowel and is a child of the next word. In the same way, the words *il* and *y* corresponding to the phonetic form /j/ have an empty syllabized form, because /j/ forms a syllable with the vowel of the next word *a* /a/.

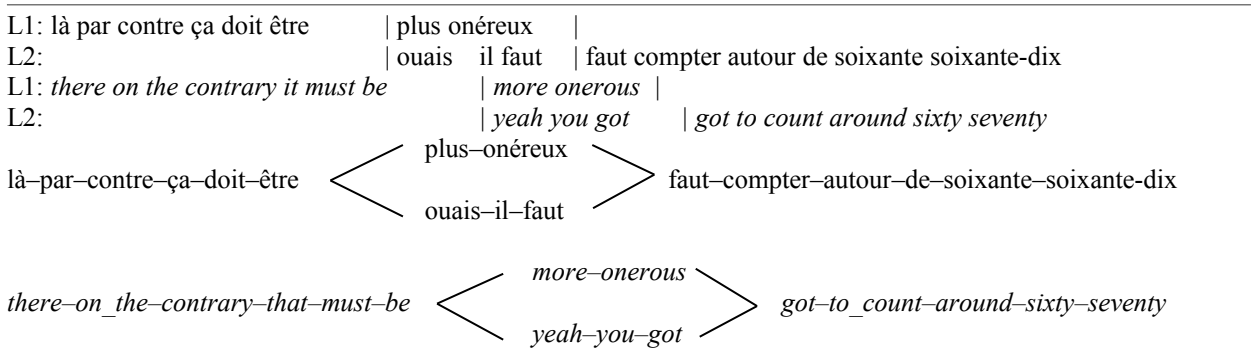


Figure 1: Transcription and lexical graph in case of overlap of speech turns

3.2 Syntactic structure

Just as phonemes are the base units of the prosodic structure, lexemes are the base units of syntax. All our syntactic structures are aligned on the lexemes. Most lexemes are time-aligned, i.e. we can determine the beginning and end of the utterance on the timeline because most lexemes correspond to words for which we have the time-alignment via the alignment of their phonemes. Some lexemes, however, are not time-aligned, for two different reasons:

Contrarily to the prosodic side of the data, generally, overlaps can be transcribed for all speakers. Nevertheless, the alignment of the words is not systematic and we do not access any time-alignment for lexemes contained in the overlap.

Secondly, we have lexemes that are not in a one-to-one relation with words. The most prominent case for French are porte-manteau words like *au* /o/ that are composed of two lexemes (À + LE, 'to + the') and contraction like *il y a* 'there is' pronounced /ja/ rather than /ilia/.[3]

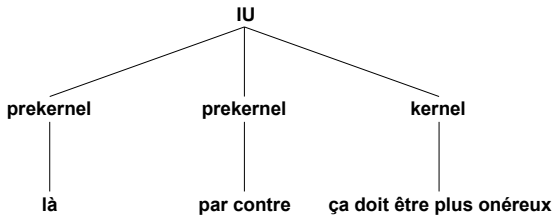
This implies that the order of lexemes is nondeductible from the alignment of the lexemes on the time line and has to be provided independently. Contrarily, to the order of time-aligned prosodic structures, the order of the lexemes is a partial order, i.e. we do not have a precedence relation between any pair of lexemes. This is due to overlaps, where two speakers can produce lexemes at the same time.

The fact that a spoken corpus needs two orders for the annotation, a temporal order and a structural order, partially aligned, was anticipated and formalized by Bird & Liberman (2001). Their formalization was implemented for the AN.ANA.S corpus by Voghera & Cutugno 2009, however, without addressing our central problem of the duality of time-aligned and non time-aligned items. Contrarily to Bird & Liberman, we prefer to introduce an order on lexemes rather than to introduce abstract points with only structural order relations with the relevant time points (that is the start and end points of the time aligned units).

Figure 1 shows an extract from the Rhapsodie corpus containing a speech turn overlap: The transcription is followed by a lexical graph where some lexemes have been produced in parallel and thus have no mutual order, for example *onéreux* 'onerous' has no order relation with *ouais* 'yeah'.

The syntactic annotation consists of various constituent and dependency structures. The macrosyntactic structure is a constituent tree. The maximal macrosyntactic unit we consider is the illocutionary unit (IU), which is divided into one central component, the kernel, bearing the illocutionary force, and some peripheral components. The next examples are annotated following the conventions exposed in Deulofeu et al. (2010), which are equivalent to the associated constituent tree.

(6) là < par contre < ça doit être plus onéreux // (D005)
there < however < it got to be more onerous

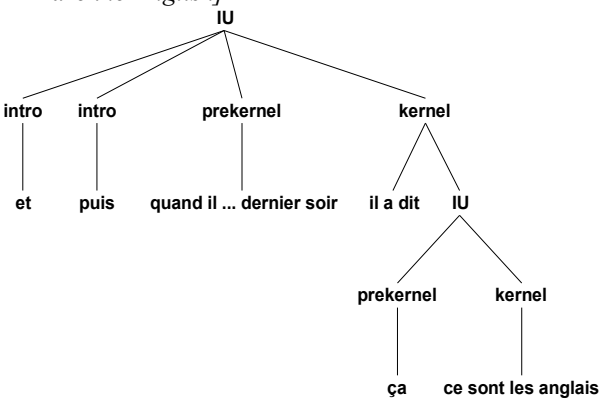


The macrosyntatctic tree is recursive because IUs can themselves contain other IUs, for instance in

[3] Due to technical reasons of our alignment process where lexemes were aligned to tokens, we encounter a similar situation when a token contains two lexemes due to elisions (*c'est* = CE + ÊTRE 'this + be' and *l'ami* = LE + AMI 'the + friend').

the following case of reported speech, *ça ce sont les anglais* forms an embedded illocutionary unit.

(7) ^et ^puis quand il a entendu le bombardement anglais { le de~ | le dernier soir } < il a dit [ça < ce sont les anglais //] //
^and ^then when he heard the English bombing { the la~ | the last evening } < he said [that < this are the English]

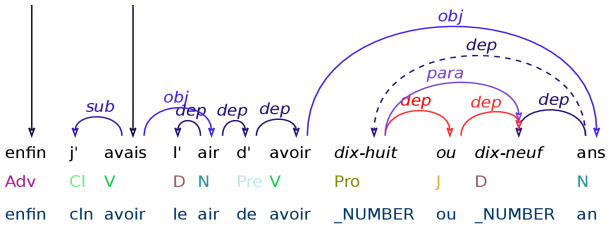


A dependency structure is commonly not a tree but a DAG (Tesnière 1959, Hudson 1990, Gerdes & Kahane 2011) because in some constructions, certain nodes are assigned multiple heads. For example, coordinations and other pile structures can have a symmetrical and an asymmetrical analysis. In the symmetrical analysis of (8) *dix-huit* and *dix-neuf* are coheads of the determiner phrase *dix-huit ou dix-neuf* and thus both dependent of *ans,* whereas in the asymmetrical analysis, *dix-huit* governs *ou*, which governs *dix-neuf*. We want a structure that subsumes both of these analyses which implies a graph (and not a tree) structure of our dependency analysis. We also add a paradigmatic link between the two conjuncts.

(8) enfin < j'avais l'air d'avoir {dix-huit | ^ou dix-neuf } ans (D201)
so < I had the appearance to have {18|^or 19} years
'Well, I looked like 18 or 19'

More precisely, we can type the two dependency links between *ans* 'years' and its determiners in a way to privilege one of the two links.[4]

The dependency analysis of Rhapsodie has no projectivity constraint, but we disallow directed cycles.[5] A dependency DAG canonically induces a



constituent tree, the tree of maximal projections of each lexeme. Note that in case of a node with multiple governors, the projection does not induce a partition of the lexemes, as the resulting constituents overlap and the projection of the node with governors appears (at least) twice in the resulting constituent tree. This constituent tree has discontinuous constituents if and only if the dependency DAG is non-projective (Lecerf 1961).

Note that, next to the macrosyntactic constituent tree and the constituency induced by dependency, we consider a third constituent structure, consisting of the piles with their layers, because many piles are not microsyntactic constituents (like for instance the disfluencies in (4)).

4 Implementation

4.1 Internal data exchange and export format

We developed an XML format for internal data exchange during the annotation process and for export of the final treebank. The format is well-adapted to our specific needs: It allows for time-aligned, partially time-aligned and only indirectly time-aligned DAGs of tokens, and two types of tree structures, dependency-like and constituency-like trees, on any of the token DAGs. For example, the syntactic annotation is based on the (incomplete, see above) time-alignment of the transcription. This is linked to another DAG of tokens where precedence relations are added based on the order of the transcription. This list is then linked to the lexemes. This linking is non trivial as it contains two-to-one relations (e.g *parce que* 'because') and one-to-two relations (e.g. *au* = À + LE 'to the').

The development of the format was guided by the Paula format and existing TEI norms. On the

[4] Some authors like Mel'čuk 1988 consider that the first conjoint is systematically the head. However, in the case of left headed coordinative structures we have good reasons to consider the second conjoint as preeminent because of closer prosodic and agreement ties between the latter conjoint and the head.

[5] For instance, some analyses of wh-words, like Tesnière (1959), see a double syntactic position of the relative pronoun, one as the complementizer and a second as the pronoun saturating a syntactic position inside of the relative clause. The dependency representation of this analysis causes a directed cycle.

one hand, the multiple token DAGs make the format slightly more complex than Paula, on the other hand, the single file structure and the limitation to the two types of tree-structure allowed us to slightly simplify Paula: A dependency graph is a simple list of relations on a token DAG, and constituency is encoded using directly the XML-inherent constituency structure (whereas in Paula, higher nodes have an explicitly encoded governor relations with the nodes it is composed of).

4.2 The structure of the database

Our treebank consists of three types of structures: different sets of segments on which hold precedence relations (partially induced from the time-alignment), constituent trees, and dependency graphs. We have decided to pre-compute and store in the Rhapsodie database the whole of acoustic correlates associated with each syntactic unit and each perceived prosodic event. It is therefore possible to search not only F0 tracks but also durations of segments, speech rates, temporal characteristics of a melodic contour, speech rate, intonational register, etc.

Each unit is stored in a "unit" table, with an attribute specifying its type ("syllable", "iu", ...). All the details of a unit (start time, end time, textual value, flow, particular attributes ...) are stored in an "attribute" table, with columns notably referring to the name of the attribute, its value, and the identifier of the corresponding unit. Relational tables store the different links between the units.[6]

[6]A first table refers to the "parent to child" relations, and includes two columns giving respectively the identifier of the parent and the identifier of the child. This table also provides two columns which indicate the place of the child unit among the other children of the parent, from the left and from the right. A second table stores the transitive closure of the "parent to child" table. We also have two tables storing the "direct" dependency relations between lexemes and their transitive closure (the "indirect" relations). The data base also provides tables to store the precedence and succession relations between units. In such tables, there are two columns corresponding to the identifiers of the left and right units. We also have to distinguish the temporal order (for the prosodic units) and the lexeme based order (for the syntactic units). Many segments of our annotation can thus appear twice with an identical span; once as a phonological group, once as a syntactic group. The database stores them separately, while keeping the identical time-alignment via the linking to the phoneme tier.

4.3 Rhapsodie QL

We developed the Rhapsodie Query Language based on the Annis Query Language (Zeldes et al. 2009). Our goal was to keep the full descriptive power of AnnisQL while adapting the language to the needs of intonosyntactic corpus searches. The differences are mainly caused by the multiple partial precedence relations that our trees are built upon and the rather numerical than symbolic character of the prosodic queries. Even queries crossing prosody and syntax are particularly simple as both annotations share the same constituency tree encoding. Moreover, we wanted to provide directly in the query language some simple mathematical functions for statistical studies of the corpus.

The Rhapsody QL covers the whole intono-syntactic structure of the corpus and allows to specify constraints on every level.

A query in RQL is composed of three parts:

- The **definition of variables** and their unit types:
  - $x1 = phone; $x2 = ui; ...
- The **constraints** to be applied on these variables. Such constraints correspond to "paths" through the structure to attain the nodes on which we want to define restrictions. To each level corresponds a unit type and a tree depth, where we want to define a restriction, or several restrictions with Boolean operators. A level is described between brackets.
  For example, we define a "group", and we want this group to be "rhythmic strong" or "rhythmic weak", or we want the group to be included in a period whose duration exceed 5 seconds:
  $gr = group; CONSTRAINT ([$gr.type = "rhythmic strong" | $gr.type = "rhythmic weak"] | [$gr] in [period.duration>5])*[7]
- The **results**: the specific attributes of the specific units we want to get.
  If we take the last example and we want to return the duration of the groups which satisfy the constraint, and also we want to take the textual value of the last phone of these groups

[7] We have defined a variable "$gr" whose type is "group", and we have two paths separated by the "OR" boolean: [$gr.type="rhythmic strong" | $gr.type = "rhythmic weak"]. [$gr] in* [period.duration>5] consists in starting from the "group" level, to go up to the parent level "period" with the large child to parent relation "in*", and to restrict the level "period".

when this phone is not an "a", we continue the request by:*RETURN [$gr].duration; [$gr] in*(rl1) [phone.text!="a"].text;*
We asked for two results:
[$gr].duration: the duration of the selected groups
[$gr] ni(rl1) [phone.text="!a"].text*: for each group, we search the last phone child ("rl1": the first from right to left), we precise we do not want an "a", and we ask to return the attribute "text" (*ni* is the parent to child relation, the converse relation of *in*).

RhapsodieQL also provides functions, which will take as argument numbers, strings, queries, or recursive function calls. For instance, if we want the ratio between the mean of the duration of the rhythmic groups and the mean of the duration of all the groups, we will ask:

- *ratio(mean(RETURN[group.type="rhythmic %"].duration;!),mean(RETURN [group].duration;!))*

5 Conclusion

The development of both the prosodic and the syntactic annotation schemata was guided by the objective of modeling the interplay between prosody and syntax in discourse structuring. In order to achieve this goal we decided not to constrain the complexity of spontaneous speech productions within the limits of a given model of linguistic representation selected a priori. Rather we borrowed and formalized general representation principles from various compatible data-oriented models – such as the Dutch-IPO school for prosody, dependency grammars, and the macrosyntactic theory of syntax. Building on the difficulties we encountered in the annotation task, we induced and refined our formal models of syntactic and prosodic representations. We incrementally adapted our annotation to these emerging models. Traditional annotation schemata could not be applied due to our choice of not neglecting what is usually considered as "performance" phenomena: hesitations, disfluencies, incomplete utterances, dialogical completion of syntactic structures, parentheses, overlaps, grafts, etc.

The choice of giving a unified representation of prosodic and syntactic phenomena has raised a number of new theoretical and practical issues.

1. We found that whereas prosodic units are all time-aligned, syntactic constituents are aligned on lexemes, i.e., on units which are only partially time-aligned.
2. Our database considers therefore two types of orders for our structures: time on the one hand and partial order of lexemes on the other hand. These two orders are partially aligned to one another.
3. Several constituent structures are considered: two hierarchies for prosody (rhythmic groups can be independently partitioned into words or metric feet) and three hierarchies for macrosyntax, microsyntax, and piles.
4. Our dependency structure is represented through a directed acyclic graph rather than a tree. This representation has been chosen to account for the various possible analyses of the syntactic structures of a pile.

RhapsodieQL, the query language developed for parsing our data structures, extends previous query languages as it allows the user to simultaneously explore time-based and lexeme-based structures and to cross-search prosody and syntax.

This corpus allows to answer some important questions concerning spoken language in general and spoken French in particular:

- The hypothesis of the dependency connectivity of prosodic constituents (Mertens 1987)
- The prosodic structure of cleft sentences.
- The prosodic contours of left and right dislocated elements (pre- and post-kernels)
- The frequency in spoken language of non-canonical sentences, i.e. illocutionary units which are not realized by complete verbal dependency units.
- The study of prosodic differences between coordination and reformulation (Kahane & Pietrandrea 2012)

The free status of the Rhapsodie corpus and the corresponding tools as well as the existing prosodic and deep syntactic annotations provide a good basis for additional and competing levels of syntactic, prosodic, and semantic analyses. Further levels of annotation on the corpus could for instance include a complete discourse structure and coreference annotation, which will allow for a deeper study of the prosodic realizations of information packaging (the communicative structure).

References

Avanzi M., Simon A.-C., Goldman J.-Ph., Auchlin A. 2010. C-PROM. An annotated corpus for French prominence studies. Prosodic Prominence: Perceptual and Automatic Identification, Proc. Speech Prosody, Chicago.

Bird S., Liberman M. 2001. A formal framework for linguistic annotation. Speech Communication, 33(1,2), 23-60.

Boersma P., Weenink D. 2012. Praat: doing phonetics by computer [Computer program]. Version 5.3.10, http://www.praat.org/.

Brants Th. 2000. Inter-annotator agreement for a German newspaper corpus, Proc. LREC.

Cerisara C., Gardent C., Anderson C. 2010. Building and Exploiting a Dependency Treebank for French Radio Broadcast. Proc. 9th international workshop on Treebanks and Linguistic Theories (TLT9), Tartu : Estonie

Cheng W., Greaves C., Warren M. 2008. Discourse Intonation systems. A Corpus-driven Study of Discourse Intonation The Hong Kong Corpus of Spoken English (Prosodic). Benjamins.

Cresti E. 2005. Notes on lexical strategy, structural strategies and surface clause indexes in the C-ORAL-ROM spoken corpora. In E. Cresti, M. Moneglia (2005), 209-256.

Cresti E., Moneglia M. (eds.) 2005. C-ORAL-ROM. Integrated reference corpora for spoken romance languages, DVD + vol., Benjamins.

Chiarcos C., Dipper S., Götze M., Leser U., Lüdeling A., Ritz J., Stede M. 2008. A Flexible Framework for Integrating Annotations from Different Tools and Tag Sets. Traitement Automatique des Langues, 49.

Delais-Roussarie E., Meqqori A., Tarrier J.-M. 2003. Annoter et segmenter des données sous Praat. In E. Delais-Roussarie, J. Durand (eds.). Corpus et Variation en Phonologie. Presses Universitaires du Mirail.

Delmonte R. 2009. Treebanking in VIT: from Phrase Structure to Dependency Representation, in S. Nirenburg (ed.), Language Engineering for Lesser-Studied Languages, IOS Press, Amsterdam, The Netherlands, 51-80.

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, The Fourth Linguistic Annotation Workshop (LAW IV), 8 p.

Gerdes K., Kahane S. 2009. Speaking in Piles. Paradigmatic Annotation of a Spoken French Corpus. Proc. Corpus Linguistics Conference, Liverpool.

Gerdes K., Kahane S. 2011, Defining dependencies (and constituents), Proceedings of Depling, Barcelona.

Gibbon D. .2003. Corpus-based syntax-prosody tree matching. Proc. Eurospeech 2003, Geneva.

Hinrichs E. W., Bartels J., Kawata Y., Kordoni V., Telljohann H. (2000) The VERBMOBIL Treebanks. In W. Zuehlke, E. G. Schukat-Talamazzini (eds.), Proc. KONVENS 2000 Sprachkommunikation, ITG-Fachbericht 161, 107-112. VDE Verlag.

Hudson R. 1990. English Word Grammar. Oxford: Blackwell.

Kahane S., Pietrandrea P. 2012. Les parenthétiques comme « Unités Illocutoires Associées » : un approche macrosyntaxique, in M. Avanzi & J. Glikman (éd.), Les Verbes Parenthétiques : Hypotaxe, Parataxe ou Parenthèse ?, Linx, 61, 19 p.

Lacheret-Dujour A., Victorri B. 2002. La période intonative comme unité d'analyse pour l'étude du français parlé : modélisation prosodique et enjeux linguistiques. Verbum, M. Charolles (ed), Nancy, 55-72

Lacheret A., Obin N., Avanzi M. 2010. Design and evaluation of shared prosodic annotation for spontaneous French speech: from expert knowledge to non-expert annotation", Proc. 4th Linguistic Annotation Workshop (LAWIV), Uppsala, 265-273.

Lecerf Y. 1961. Une représentation algébrique de la structure des phrases dans diverses langues naturelles", C. R. Acad. Sc., Paris.

Martin Ph. 2000. WinPitch 2000: a tool for experimental phonology and intonation research, Proc. Prosody 2000 Workshop, Kraków, Pologne.

Mel'čuk I. 1988. Dependency Syntax: Theory and Practice. The SUNY Press, Albany, N.Y.

Mertens, P. 1987. L'intonation du français. De la description linguistique à la reconnaissance automatique. Unpublished Ph.D., Univ. Leuven, Belgium.

Meteer M. et al. rev. by A. Taylor. 1995. Dysfluency Annotation Stylebook for the Switchboard Corpus.

Nelson G., Wallis S., Aarts B. (eds.) 2002. Exploring Natural Language: Working with the British Component of the International Corpus of English, John Benjamins Publishing Company, Varieties of English Around the World G29.

Pulgram E. 1970. Syllable, Word, Nexus, Cursus. La Haye: Mouton.

Rahman A., Sampson G. R. 2000. Extending grammar annotation standards to spontaneous speech. In J.M. Kirk (ed.), Corpora Galore: Analyses and Techniques in Describing English, Amsterdam: Rodopi, 295-311.

Schmidt T. 2004. Transcribing and annotating spoken language with EXMARaLDA. Proc. LREC-Workshop on XML-based richly annotated corpora.

Schuurman I., Goedertier W., Hoekstra H., N. Oostdijk, R. Piepenbrock, Schouppe M. 2004. Linguistic annotation of the Spoken Dutch Corpus: If we had to do it all over again ..., Proc. LREC, Lisbon, 57-60.

Silverman K., Beckman M., Pitrelli J., Ostendorf M., Wightman C., Price P., Pierrehumbert J., Hirschberg, J. 1992. TOBI: a standard for labeling English prosody", Proc. ICSLP-1992, 867-870.

Smith, C. 2011. Naïve Listeners' Perceptions of French Prosody Compared to the Predictions of Theoretical Models. Proc. 3rd IDP conference, 349-335

't Hart J., Collier R., Cohen A. 2006. A perceptual study of intonation, an experimental phonetic approach to speech melody, Cambridge University Press.

Tepperman J., Narayanan S. 2008. Using articulatory representations to detect segmental errors in nonnative pronunciation. IEEE Transactions on Speech, Audio and Language Processing, 16(1):8-22.

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

Voghera M., Cutugno F. AN.ANA.S.: aligning text to temporal syntagmatic progression in Treebanks. In Proceedings of the fifth Corpus Linguistics Conference, Liverpool 20-23 July. 2009

Zeldes A., Ritz J., Lüdeling A., Chiarcos C. 2009. ANNIS: A Search Tool for Multi-Layer Annotated Corpora. Proc. Corpus Linguistics 2009, Liverpool, UK.