The Corpus of Interactional Data: a Large Multimodal Annotated Resource

Philippe Blache, Roxane Bertrand, Gaëlle Ferré, Berthille Pallaud, Laurent Prevot, Stéphane Rauzy

The Corpus of Interactional Data: a Large Multimodal Annotated Resource

Philippe Blache, Roxane Bertrand, Gaëlle Ferré, Berthille Pallaud, Laurent Prevot, Stéphane Rauzy

The *Corpus of Interactional Data*: a Large Multimodal Annotated Resource

Philippe Blache, Roxanne Bertrand, Gaëlle Ferré, Berthille Pallaud, Laurent Prévot, Stéphane Rauzy

Abstract

1 Introduction

Studying language in its natural context is one of the new challenges for natural language processing as well as linguistics in general. Much work have been done in the perspective of spoken language processing, even though the issues in this domain remains largely unsolved (disfluencies, ill-formedness, etc.). But the problem becomes even harder when trying to take into account all the aspects of natural communication, including pragmatics and gestures. In this case, we need to describe many different sources of information (let's call them linguistic domains) coming from the signal (prosody, phonetics), the transcription (morphology, syntax, lexical semantics), as well as the behavior of the conversation partners (gestures, attitudes, etc), the contextual background, etc. Taking into account such a rich environment means that language is seen in its multimodal dimension which necessitates a full description of each verbal or non-verbal domain as well as their interaction. Such a description is obviously a pre-requisite before the elaboration of a multimodal theory of language. It is also a basis for the development of parsing tools or annotating devices. Both goals rely on the availability of annotated resources, providing information on all the different domains and modalities. This is the goal of the project described here, that led to the development of a large annotated multimodal corpus called CID (*Corpus of Interactional Data).*

In this article, the context of multimodality and the issues we are faced with when building multimodal resources will first be presented. The second part, we will present more precisely the organization of the project during which the CID corpus was built. The rest of the paper will describe the solutions we propose to what we consider as the main issues for multimodal annotation, namely the annotation scheme, the alignment between the different domains and the interoperability of the different sources of information.

2 Multimodal Interaction and its Annotation

Our work aims at collecting data in natural situations, with audio and video recordings of human interaction, focusing then on language and gestures, to the exclusion of the other kinds of modalities be they natural (smell, touch) or artificial (related to human-machine interaction for example). More specifically, what we are interested in when studying such domains is the interaction that exists between the different sources of information. Indeed, we think that (1) meaning comes from the interplay of different dimensions such as prosody, lexicon, gestures,

attitude, etc and (2) these dimensions are directly related one to another, independently from the modality they come from. In other words, they are not hierarchically organized.

2.1 Situation

Different types of resources making possible a multimodal description are now available for the description of natural interaction. However, only few of them propose an adequate level for information representation. The particularity of multimodal linguistics being the study of the verbal and non verbal aspects of the interaction, corpora use video as primary data. Enriching such data relies on a precise orthographic and phonetic transcription, a precise alignment of transcription onto the signal and the representation of all the different levels of linguistic information. Unfortunately, only few resources provide such precise annotations of the different domains: prosody, syntax, pragmatics, gestures etc. One of the difficulty comes from the fact that for many of these domains (typically gestures), the annotation process is so far entirely manual. As a consequence, the existing projects addressing multimodality usually focus on some of these domains, mainly those with existing tools providing automatic annotation (e.g. POS tagging, segmentation, etc.).

Only few initiatives try to build broad coverage annotated corpora, with a good level of precision in the annotation of each domain. The AMI project is one of them [Carletta, 2006], even though the annotations do not seem to be at the same precision level in the different domains. For its part, the LUNA project [Rodriguez et al., 2007] focuses on spoken language understanding. The corpus is made of human-machine and human-human dialogues. It proposes, on top of the transcription, different levels of annotation, from morphosyntax to semantics and discourse analysis. Annotations have been done by means of different tools producing different formats. Another type of project in the context of human-machine interaction is presented in [Kruijff-Korbayova et al., 2006]). Annotations are done using the Nite XML Toolkit [Carletta et al., 2003]; they concern syntactic and discourse-level information, plus indication about the specific computer modality used in the experiment. A comparable resource, also acquired following a Wizard-of-Oz technique, has been built by the DIME project [Pineda et al., 2002] for Spanish. In comparison with previous ones, this resource mainly focuses on first-level prosodic information as well as dialog acts.

Our goal with the OTIM project is to bridge the gap towards large and richly annotated multimodal corpora. Such resources are required in order to understand what kind of information is encoded in each domain and, moreover, what kind of interaction exists between the different domains. This means that all the different domains have to be annotated in such a way that some alignment between them can be stipulated. Figure 1 illustrates an example of annotations associated with a segment of the corpus CID we created during OTIM. Each line represents a type of information, several lines being possibly grouped, according to the structuration level of the domain.

Annotating a multimodal corpus is a two-stage process: first, building an abstract knowledge representation scheme and second, creating the annotation, following the scheme. Several coding frameworks have been proposed by different initiatives such as MATE, NIMM, EMMA, XCES, TUSNELDA, etc. However, their application to corpus annotation usually focuses on one or two modalities. Our goal with the OTIM framework is a fine-grained annotation covering all the different modalities. This means first a precise representation of the different domains has to be built, and this is done in terms of typed feature structures (see next section). The result is a unique and homogeneous formal framework, offering in particular the

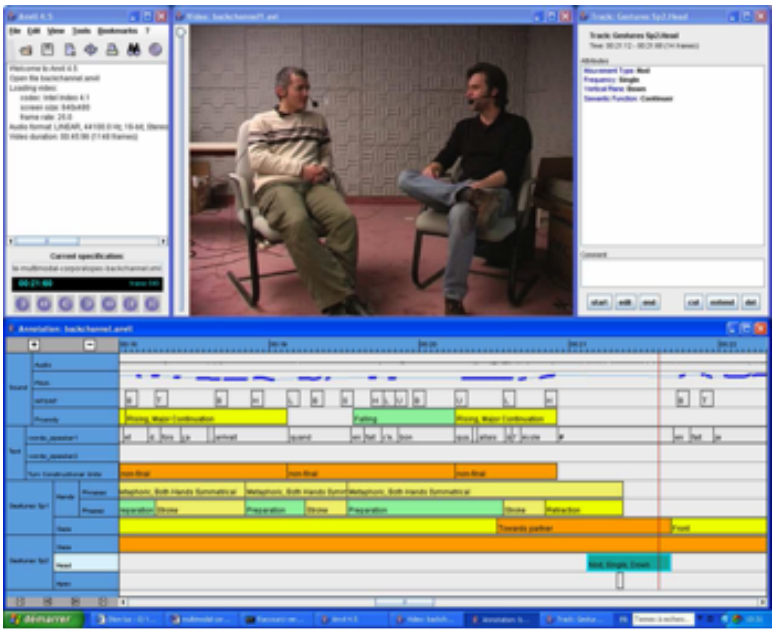


Figure 1: Snapshot from the Anvil annotation window, CID corpus

possibility to represent relations between the different domains directly. This aspect is of deep importance: as explained above, one of the problems with multimodally annotated corpora is the possibility to retrieve or manipulate complex information, made of subsets of annotations coming from these different domains. For example, some studies can focus on specific gestures (e.g. pointing gestures) associated with certain morphosyntactic categories (e.g. pronouns) and possibly some particular intonation pattern. This amounts to the querying of different (and separate) annotations and to calculate how they can be related to each other. Encoding information by means of a homogeneous abstract model is an element of answer to this issue: all information can be encoded into the same language (whatever the encoding tool or platform initially used). The problem in querying such complex information mainly consists in identifying the alignment (or synchronization) relations.

2.2 Description of the corpus

So far, the CID comprises 8 hours of video recordings. Each hour is a recording of a conversation between two participants, all the participants being French and either from the south-east region of France or having lived there for several years. Each recording involves either two male or two female participants, which makes a total of 10 women and 6 men. The corpus type is a compromise between genuine interactions and corpora such as MAPTASK [Anderson et al., 1991] also called task oriented corpora.

Before the recording, the participants were suggested one of the following two topics of conversation: either to speak about conflicts in their professional environment or about funny situations in which they may have found themselves involved. These however were suggestions and the participants were free to speak of any topic which may have come to their mind and indeed if all participants tried to stick to the task they also had bits of interaction in which

they were clearly speaking of something else.

All the participants were quite familiar with the lab (they were all either permanent members of the lab or doctoral students – the familiarity was indeed a prerequisite condition since it reduced the level of stress induced by the recording itself). They were also familiar with each other: this second condition aimed at obtaining more spontaneous conversations which would involve a certain conversational background. The result was very satisfying since the recordings sound like spontaneous conversations: speech is at some moments extremely fluent and at others on the contrary quite hesitant (which appears in the numerous filled pauses and false starts for instance). Speaking turns do correspond to the conversation structure described in [Sacks et al., 1974]. Transitions are sometimes smooth, that is speaker change does not take a long time (silent pause) or occur with too much overlap. At other times, many overlaps are observed revealing non-smooth transitions [Koiso et al., 1998]. The recordings were made in a studio and the two participants were seated in separate fixed armchairs next to each other and slightly oriented towards each other. They were filmed by a digital camera (Canon XM2) adopting a fixed frame and were equipped with head microphones so as to record each voice on a separate track. Thanks to this device, the speech recordings could be treated at the phonetic level since it enables for instance to process the signal in its integrity, even when the two participants speak in overlap. It would not have been possible otherwise to analyze the overlapping speech segments with a sound analysis tool such as PRAAT, since no such tool has the capacity to separate voices.

3 Transcription

The question of transcription, in spite of its apparent simplicity, is of deep importance. The way the audio signal is transcribed can indeed condition the rest of the annotation. This section presents the general guidelines that have been elaborated by a group of several teams involved in such a task. The goal is to propose a way to homogenize (and then to insure interoperability) the transcription conventions of spoken languages. Several recommendations have already been made in this perspective (EAGLES, TEI, etc.). However, despite these attempts, a large number of specific conventions still exist, each research group offering its own recommendations based on its specific needs (scientific perspective, linguistic domain, etc.).

The convention adopted here is not exhaustive, but rather provides a basic ground that could be shared by different transcription conventions. this proposal is based on the conventions elaborated by different laboratories[1].

Context, principles: One recommendation we follow is that a standard orthographic transcription is always preferable: no transcript will therefore use a spelling trick, for example when transcribing specific pronunciations e.g. `/chui/` instead of `/je suis/` (*I am*). Using correct lexical entries offers several advantages, on top of being respectful of the linguistic production. First, it ensures the possibility of a good alignment with the signal. Several tools make such a process automatic, provided that a good grapheme-phoneme conversion is possible. It is known that their results are better when using correct lexical items. Moreover, such

[1]This is a French-speaking initiative, the partners of the project are ATILF, ICAR, LI, LIMSI, LLL, LPL, SYLED, VALIBEL.

inputs also enable an automatic processing of transcription (POS-tagging, syntactic parsing, etc.).

Moreover, other principles have guided our approach:

- *Independence from the editing tools*: the types of information encoded as well as their representation should not depend on the tools or features they offer. For example, the fact that a system like Transcriber provides specific tags to encode information (e.g. laughters) or specific devices (e.g. overlaps) does not replace the need of an explicit encoding.

- *Distinction transcription vs. annotation*: the primary transcript is limited to the encoding of information which cannot be generated automatically. In general, we do not encode any information that does not come directly from the utterance or requires particular interpretations.

- The transcription is intended to be "*inline markup*": we encode information together with the transcribed speech, at the same level. This is a clear distinction with annotations (for example discourse-level, gestures or syntactic annotations) that are typically standoff.

The different types of annotation are detailed in the table below. The first type concerns information associated to sets of words in the transcription. This is the case for example for information such as the type of sequence (toponyms, acronyms, etc.) or the way the words are pronounced (without entering into a prosodic analysis: whispering, laughing, etc.). It can also give higher level information such as code switching. This information may be of great help in particular when parsing the transcription automatically.

Phenomenon	Encoding	Example	TEI correspondence
Acronyms	$ ... A/$	$LREC A/$	<seg type="acronym">
Code switching	$ ... C/$	$ partie en Français C/$	<foreign xml:lang="de">
Whispering	$ ... W/$	$blowing in the wind W/$	
Laughing	$ ... L/$	$I am happy L/$	
Spelled words	$ ... S/$	$ h a p p y S/$	
Untranscribed parts	$ ... X/$	$ comment X/$	
Patronyms	$ ... P/$	$ Mark P/$	<seg type="patronym">
Laughter	$ R/$	$ R/$	<desc>laugh</desc>
Titles	$ ... O/$	$East of Eden O/$	<seg type="title">
Toponyms	$ ... T/$	$London T/$	<seg type="toponym">

A second type of information encoded during transcription concerns specific realizations: missing elements (elision) or addition of phonemes (non standard liaison), disfluencies (truncated words, filled pauses).

Phenomenon	Encoding	Example	TEI correspondence
Partial words	-	courti-	
Elision	()	i(l) y a d(é)jà	
Hesitation	list	euh, mmh	<desc>mmh</desc>
Specific liaisons	=...=	donne moi =z= en	
Unclear words	?	?	<gap>
Onomatopoeia	list	meow, oink	<desc>meow</desc>
Breaks	#	#	<pause>

Different other kinds of information are encoded during transcription, in particular about pronunciations (foreign words, specific pronunciations, direct phonetic transcription). This information is helpful both in the perspective of automatic alignment and phonetic studies.

Phenomenon	Encoding	Example	TEI correspondence
Noise	... /	noise_type /	<incident>
Overlap	< ... >	< ... >	<who + trans=overlap>
Foreign words	[... , ...]	[B&B, biEnbi]	<foreign xml:lang="de">
Multiple transcriptions	/ ... /	/des, les/ acides	<choice>
Specific pronunciations	[... , ...]	[aéroport, aReopOR]	
Phonetic transcription	[? , ...]	[?, sampa]	

A transcription following such requirements facilitates (1) the generation of the phonetic transcription and (2) an automatic alignment with the sound signal. This has been done using the ANTS4 system developed at the LORIA-INRIA by [Fohr et al., 2004]. The alignment has been checked manually and errors (principally due to schwa deletion and sound assimilations in ordinary speech) are corrected to provide a precise phonemic transcription of the speech signal which constitutes the transcription basis for every annotation.

The following table summarizes the main figures about the different specific phenomena annotated in the EOT. To the best of our knowledge, these data are the first of this type obtained on a large corpus. This information is still to be analyzed.

Phenomenon	Number
Elision	11,058
Word truncation	1,732
Standard liaison missing	160
Unusual liaison	49
Non-standard phonetic realization	2,812
Laugh seq.	2,111
Laughing speech seq.	367
Single laugh IPU	844
Overlaps > 150 ms	4,150

4 Annotation scheme

Elaborating an annotation scheme that involves many different domains is a difficult task. It consists first in identifying the kind of information each domain is supposed to encode and second to choose a formalism encoding the information. Each domain corresponding to a different linguistic subfield, it comes with its own history and habits in terms of information representation. Moreover, our goal being generic, we want to have a general and, if possible, exhaustive representation of all the information for each domain. Concretely, the project started by the identification of the information to encode in each subfield. This resulted in defining a set of features coming with all possible values. In a second step, we discussed how to encode in an homogeneous way the information coming from the different domains. The outcome was the adoption of typed feature structures [Carpenter, 1992], which offer the advantage to encode precisely each type of information and, when necessary, to structure it into a hierarchical organization.

At this stage, we tried to remain as independent as possible from any linguistic theory, even though an organization in terms of constituents is often implicit when using hierarchical

structures. The result of this work was the elaboration of an annotation scheme consisting in a set of feature structures for all the different domains we wanted to annotate.

4.1 Notes on the OTIM scheme presentation

All annotated information (which we call an *object*) corresponds to a type, that can be organized into type hierarchies. Each type usually corresponds to a set of *appropriated* features. Moreover, in some cases, the objects contain constituents. For example, a prosodic phrase is a set of syllables, each one being a set of phonemes.

It is important to distinguish type hierarchy on one hand from constituency hierarchy on the other hand. It is clear for example that a *word fragment* is a kind of *lexicalized disfluency*, the difference between them being the level of precision of the object, both of them belonging to type hierarchy rooted by *disfluency*). It is also clear that a *phoneme* is part of a *syllable*, but a phoneme is not a specific type of syllable. In this case, a phoneme is a *constituent* of a syllable. More generally, it is important to distinguish clearly between type hierarchy and constituent hierarchy. The first can be represented by a relation *is-a*, the second by a relation *belongs-to*.

A constituent is then an object with the particularity that it has to be aligned with an upper-level one. Concretely, when using for example a tool like Anvil [Kipp, 2001], an object and its constituents will be represented respectively as primary and secondary tracks.

Before presenting types and feature structures, we propose an overview of the objects and their constituents. Here is a list of some abbreviations:

tcu	turn constructional unit
pros_phr	prosodic phrase
ip	intonational phrase
ap	accentual phrase
syl	syllable
const_syl	syllable constituent
sent	sentence
synt_phr	syntactic phrase
word	word

This list is not exhaustive and only contains complex objects (those with constituents). The constituency organization can be represented with a simple grammar (note that the grammar is not complete in the sense that not all non terminals correspond to a left-hand side of a rule). The following schema presents the constituent hierarchy of the main objects described in this presentation:

TCU ::= PROS_PHR^{+}
IP ::= AP* AP ::= SYL^{+} SYL ::= CONST_SYL^{+} CONST_SYL ::= PHON^{+}
SENT ::= SYNT_PHR^{+} SYNT_PHR ::= WORD^{+} WORD ::= PHON^{+}
DISFLUENCE ::= REPARANDUM; BREAK_INTERVAL; REPARANS

In such a representation, the operator '+' indicates that the constituents appear at least once and can be repeated, the operator '*' is the Kleene star (means 0 to n).

This grammar specifies that TCUs are formed by one or several prosodic phrases. We will see farther that prosodic phrases can be of two types: *ip* (intonational phrase) or *ap* (accentual phrase). In turn, IPs are sets of APs, that are made with one or several syllables.

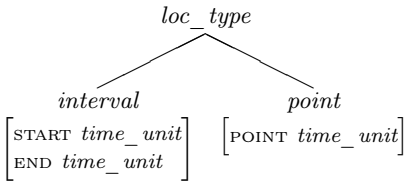
4.2 The *object* supertype

At the most general level, all objects need an index in order to be referred to (for example as target of a grammatical relation, or constituent of another higher level object). An index is simply an integer assigned to the object. Moreover, objects are defined in terms of positions in the signal. The following feature structure presents these two pieces of information, that will be conveyed by all objects:

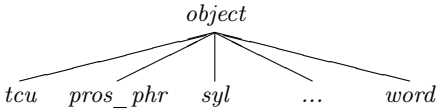
$$_{object}\begin{bmatrix}\textsc{index}\ integer\\ \textsc{location}\ loc\_type\end{bmatrix}$$

Note that by convention, types are noted in italics. A typed feature structure represents types in italics as subscript of the feature structure.

In terms of location, an object can be situated by means of two different kinds of position, depending on the fact that they correspond to an interval (for example a syllable), or a point (e.g. a tone). In the first case, interval boundaries are represented by the features START and END, with temporal values (usually milliseconds). The following hierarchy presents the location type and its two subtypes (*interval* and *point*), together with their appropriated features. Remind that a type inherits from all the properties of its supertypes. Concretely, a property being represented by a feature, the feature structure of an object of a certain type is the sum of the appropriated features of this type and that of all its supertypes.



As for typing aspects, object being the most general type, all other objects are subtypes, as represented in the following type hierarchy:



This means that tcus, words, syllables are all specific instances of the type *object*. As a consequence, they inherit its structure: all kinds of objects, whatever their subtype, will have the LOCATION feature.

4.3 Concrete encoding

When a certain type of information is specified in terms of feature structures, it is necessary to describe how such information is concretely encoded during the annotation process. Typically, when annotation is done using editors such as Praat, it is necessary to indicate what information correspond to an annotation track (called a tier in Praat), what the possible values and

their labels are. Usually, a feature corresponds to a tier name, whereas the feature value is the label of the corresponding interval in the tier.

Different types of encoding are possible, depending on the fact that the information is factorized or not. For example, one can choose to create one tier per feature. Another solution is to factorize all the features into a feature vector, creating then a unique tag.

In some situations, the factorized representation is the recommended option. This is in particular the case of recursive objects. Typically, syntactic objects have constituents that are also syntactic objects.

Below is an example of the differences between the options: one decentralized, with one tier per feature the other factorized, with one feature vector encoding all the features.

- *Example*: The encoding of intonational phrases (see infra) can be done as follows:
  - Decentralized:

Tier name	*Tag value example*
`ip.label`	IP
`ip.contour.direction`	falling
`ip.contour.position`	final
`ip.contour.function`	conclusive

  - Factorized:

Tier name	*Tag value example*
`ip`	IP.falling.final.conclusive

Note that the feature vector has to have a canonical structure. This means that each position is fixed and corresponds to a feature. In this example, the first position gives the value of the label feature, the second, that of the direction of the contour, etc.

Vector representations can be simplified by using notation conventions which encode each value with two characters. In this paper, we give a proposal for each object.

5 Annotations

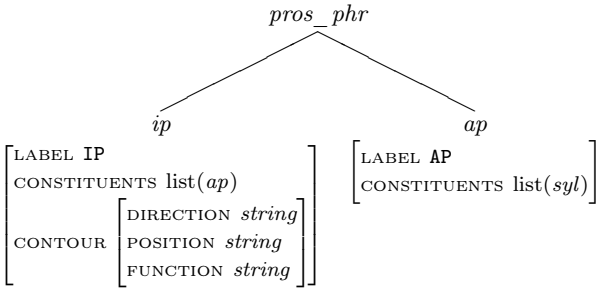
This section presents the representation of some of the domains annotated in the CID. The idea here is to show how the annotation scheme can be instantiated for each domain. For each domain presented here, we first specify its abstract organization in terms of typed feature structures. Such a representation enables to define on one hand the entire set of features involved in the description of the domain and on the other hand (when necessary) their hierarchical organization (some features being possibly constituents of others).

The second part of the description of domain annotation consists in proposing a concrete encoding for each feature (in other words how the feature is encoded by the annotator). Indeed, most of the annotations being manual, it is not possible to encode feature structures directly. We then propose to encode information in terms of vectors gathering one or several feature values.

Concretely, such an annotation process amounts to the encoding of two kinds of information separately: the general structure and its organization in an abstract schema (the TFS) plus the set of feature vectors instantiating the specific values of a given object. This two-dimensional annotation process ensures not only the implementation of a clear feature organization without the use of any ad hoc structuration mechanism (typically dependencies between tiers), but also makes it possible to generate automatically their generic XML representation from the concrete annotation (as described in the last section).

5.1 Prosody

For prosody, we adapted the model proposed in [Jun and Fougeron, 1995] in taking into consideration only two types of prosodic phrases (among the three possible units): *ip* or *ap* (remember that, by convention, types are noted in lowercase). The corresponding type hierarchy is represented as follows:



Accentual phrases (type *ap*) bear two features: the label, which value is simply the name of the corresponding type, and the list of constituents, in this case a list of syllables.

The feature structure of *ip* objects contains, on top of the label, the list of its constituents (a set of *ap*s) as well as the description of its contour. A contour is a prosodic event, situated at the end of the *ip* and is usually associated to an *ap*.

- *Example 1*: The following FS presents a complete AP structure, in which index and location feature have been added thanks to inheritance:

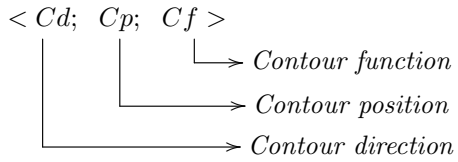
$$ap\begin{bmatrix}\text{LABEL } \texttt{AP}\\ \text{INDEX } 25\\ \text{LOCATION}\begin{bmatrix}\text{START } 192.28\\ \text{END } 204.21\end{bmatrix}\end{bmatrix}$$

- *Example 2*: This example illustrates an IP containing one AP (at its end) and characterized by a conclusive contour:

$$ip\begin{bmatrix}\text{LABEL } \texttt{IP}\\ \text{INDEX } 18\\ \text{LOCATION}\begin{bmatrix}\text{START } 83.11\\ \text{END } 204.21\end{bmatrix}\\ \text{CONSTITUENTS}\left\{\left(\text{AP}\begin{bmatrix}\text{LABEL } \texttt{AP}\\ \text{INDEX } 25\\ \text{LOCATION}\begin{bmatrix}\text{START } 192.28\\ \text{END } 204.21\end{bmatrix}\end{bmatrix}\right)\right\}\\ \text{CONTOUR}\begin{bmatrix}\text{DIRECTION falling}\\ \text{POSITION final}\\ \text{FUNCTION conclusive}\end{bmatrix}\end{bmatrix}$$

The concrete encoding of prosodic information by annotators follows the general TFS organization. The AP being terminal, it only bears the type indication, the beginning and the end of the interval, which is directly encoded into a tier. The same is valid for IPs. Besides labels, contour types is the second important kind of information to be encoded. A feature

vector can be proposed in order to encode the different possible values. By convention (unless explicitly mentioned), each feature is encoded in a vector by means of two characters, the first being uppercase. The following figure explains the vector associated to contour description:



The following table gives the generic representation of the possible contour feature values:

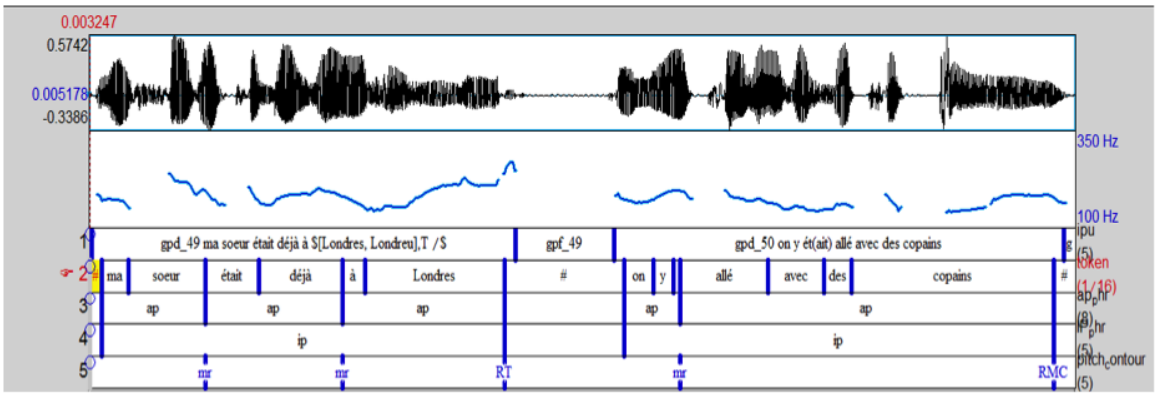
Label	`Ip`			
Index	*integer*			
Contour direction	`Rr` *rising*	`Ff` *falling*	`Rf` *rising-falling*	`Un` *unspecified*
Contour position	`Fi` *final*	`Pn` *penultimate*	`Un` *unspecified*	
Contour function	`Cc` *conclusive*	`Ct` *non conclusive*		

A specific contour encoding has been proposed in [Bertrand et al., 2007], mixing these different aspects into a compact feature encoding.

Contour type	Encoding
Falling	F
Falling from the penultimate	RF2
Rising-falling	RF1
Questioning rising	RQ
Terminal rising	RT
List rising	RL
Rising major continuation	RMC
Minor	m

In the CID, we also encoded the intonation information, following the INTSINT representation [Hirst et al., 2000] which codes the intonation by means of symbols that constitute a surface phonological representation of the intonation: T (Top), H (Higher), U (Upstepped), S (Same), M (mid), D (Downstepped), L (Lower), B (Bottom). The INTSINT annotation has been done automatically thanks to the tool presented in [Hirst, 2007].

The following figure illustrates the encoding of the prosodic information in the CID in three tiers: prosodic phrases, pitch contours and Momel-Intsint's phonological representation.



The prosodic annotation has been done by 2 experts. The annotators worked separately using Praat. Inter-transcriber agreement scores were calculated for the annotation of higher prosodic units. First annotator marked 3,159 and second annotator 2,855 Intonational Phrases. Mean percentage of inter-transcriber agreement was 91.4% and mean kappa-statistics 0.79, which stands for a quite substantial agreement.

5.2 Disfluencies

Disfluencies can occur at different levels (Shriberg, 1995 et 1999)Dister, 2008; Henry et Pallaud, 2003; 2007), we focus in this section on morpho-syntax. Disfluencies are organized around an interruption point (the break), and can occur almost anywhere in the production. These breaks and variations in the verbal fluency are related, in most of the cases, with one or several kinds of events or items inserted in the middle of a phrase or even a word. Most of the time, the statements are just hung up but in some cases these ruptures are followed by disturbances in the morpho-syntactic organization of verbal flow, the most frequently quoted being the resumptions after a break, such as auto-repairs, and incomplete phrases or words (Clark et Wasow, 1998; Guenot, 2005; Pallaud et Henry, 2006).

We propose to distinguish between two kinds of disfluencies:

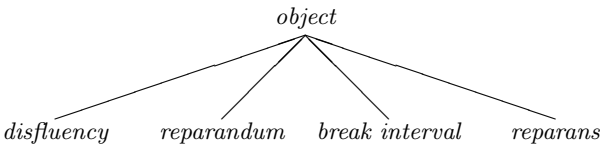
- *non lexicalized*: those without any lexical material. Typically lengthening, silent pauses or filled pauses (hm, euh, etc.)
- *lexicalized*: characterized by a non-voluntary break in the syntagmatic flow, generating a word or a phrase fragment.

According to the Shriberg's typology (1994), we separate linguistic material preceding the interruption point (the Reparandum) and those following it. In the latter, we distinguish between the content of the final utterance of the disfluency (Reparans) and the elements that can take place between the interruption point and the Reparans (Break_Interval). While the Reparandum is mandatory in these constructions, the break interval is optional, and the Reparans is forbidden in incomplete disfluencies.

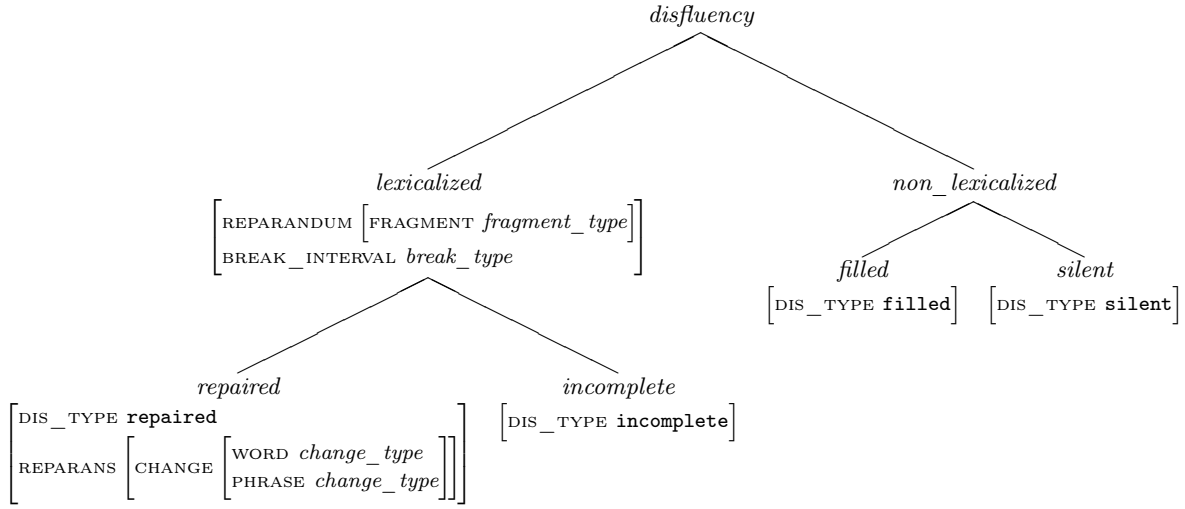
Lexicalized disfluencies reveal a particular organization:

- `reparandum`: the word or phrase fragment, in which the break occurs. Are indicated the nature of the interrupted unit (word or phrase), and the type of the truncated word (lexical or grammatical).
- `break`: a point (the break is empty) or an interval. We indicated a list of the filling elements that appear, among which: silent or filled pause, discursive connector, parenthetic statement;
- `reparans`: all that follows the break and recovers the Reparandum in continuing the statement (ie without any resumption of the Reparandum items) or in modifying or completing it (after a partial or total resumption of the Rerandum). We can indicate the position of the repair (no restart, word restart, determiner restart, phrase restart or other), and its functioning (simple continuation of the item, repair without change, continuing through repeating, repair with change in the truncated word, or repair with multiple changes).

The different objects involved in the description of disfluencies are: disfluency, reparandum, break interval, reparans. This corresponds to the general type hierarchy:



Remember that a distinction has to be made between type and constituent hierarchies. As for the latter, the following structure shows that reparandum, break interval and reparans are constituents of the disfluency.

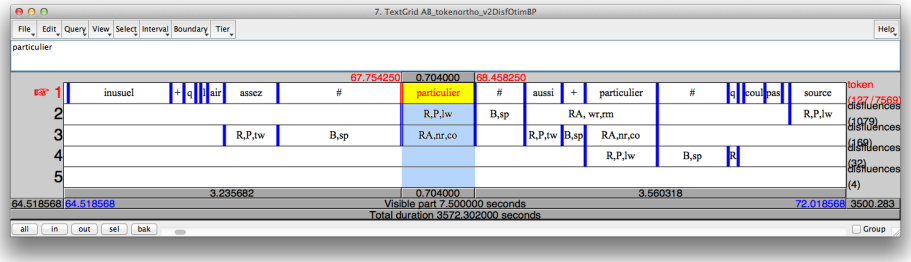


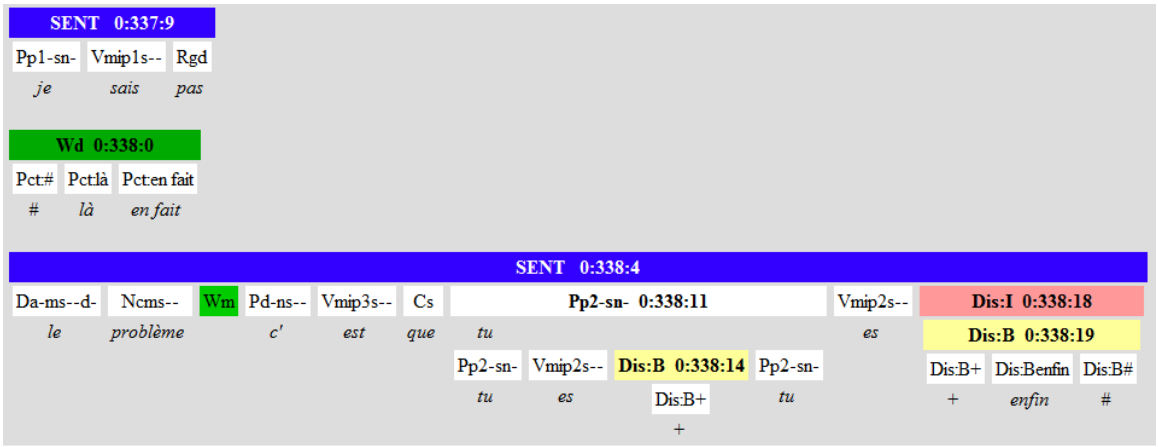
The general feature structure of a disfluency is represented in the following figure:

$$\begin{bmatrix} \textsc{reparandum} \left[\textsc{fragment}\ \textit{fragment\_type}\right] \\ \textsc{break\_interval}\ \textit{break\_type} \\ \textsc{dis\_type}\ \textit{dis\_type} \\ \textsc{reparans} \left[\textsc{change} \begin{bmatrix} \textsc{word}\ \textit{change\_type} \\ \textsc{phrase}\ \textit{change\_type} \end{bmatrix}\right] \end{bmatrix}$$

The different feature values are encoded with the following labels:

Reparandum		
Reparandum Type	`R`	*Temporary interruption*
	`I`	*Definitive Interruption*
Reparandum category	`W`	*Word reparandum*
	`P`	*Phrase reparandum*
Lexical type	`tw`	*Tool word*
	`lw`	*Lexical word*
Break type B		
	`no`	*no interval*
	`sp`	*silent pause (> 200ms)*
	`fp`	*filled pause*
	`dc`	*discursive connector*
	`ps`	*parenthetical statement*
	`rt`	*truncation repetition*
Reparans RA		
Reparans position type	`nr`	*no restart*
	`wr`	*word restart*
	`dr`	*determinant restart*
	`pr`	*phrase restart*
	`or`	*other restart*
Reparans type	`co`	*continuing the item*
	`wc`	*repairing without change*
	`rp`	*Repairing through repeating*
	`rc`	*repair with change in the truncated word*
	`rm`	*repair with multiple change*





The annotation of disfluencies is at the moment fully manual. We have developed a tool which facilitates the process in identifying such phenomena, but it has not yet been evaluated. This manual annotation requires 15mns for 1 minute of the corpus. The following table

illustrates the fact that disfluencies are speaker-dependent in terms of quantity and type. These figures also show that disfluencies affect lexicalized words as well as grammatical ones.

	Speaker_1	Speaker_2
Total number of words	1,434	1,304
Disfluent grammatical words	17	54
Disfluent lexicalized words	18	92
Truncated words	7	12
Truncated phrases	26	134

Some results

We used for disfluency annotations a semi-automatic method (detection of all Interregnum spaces; Shriberg 1994) which made it possible to identify 81% of the breaks, the 19% remainder, were manually identified.

On average, it is possible to find one rupture in the syntagmatic flow every 7.4 words (from 6.2 to 9.8 words, depending on the speakers). However, when the syntagmatic flow is stopped, it is not always broken: half of these ruptures are just hung up i.e. the statement is going on as if it had not been suspended. The other half causes a morpho-syntatic disturbance (unfinished or resumed statements); also their frequency strongly varies from one speaker to another: on average, it is one every 15.9 words.

5.3 Syntax

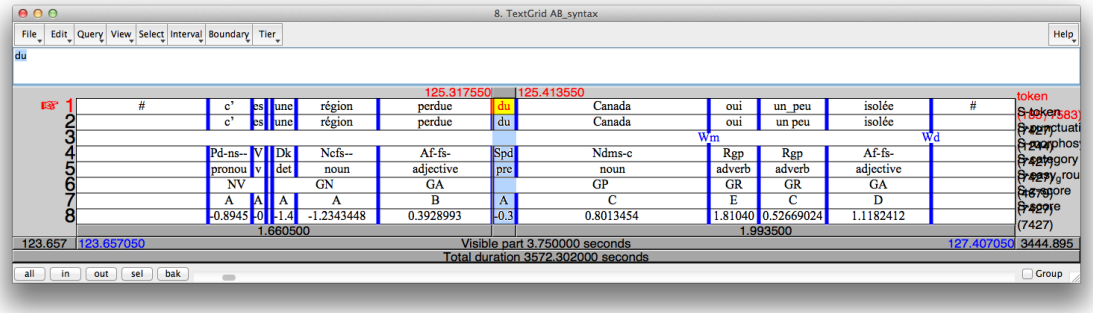
Parsing spoken languages remains problematic at a large scale and for unrestricted material such as the one we are faced with in this project. The first stage consists in encoding POS tags. The tagger we use has been originally developed for written texts with a good efficiency (F-Score 0.975) and adapted to spoken language (in particular in modifying the category distribution of the forms). It uses a precise tagset of 51 categories. The results are very good, the adapted POS tagger obtaining a 0.948 F-Score. The CID tagging has been manually corrected (about 6,000 errors for 115,000 tokens). These results show that the tagger could be used even without any correction with a good reliability.

In order to propose a broad coverage syntactic annotation, we chose to annotate three levels: chunks, trees and specific constructions. The lowest syntactic annotation, namely chunks, has been done automatically thanks to a stochastic parser developed at the LPL [Blache and Rauzy, 2008]. This tool performs at the same time POS-tagging, chunk bracketing and sentence segmentation. This last operation consists in identifying the largest syntactically homogeneous fragments, that could correspond to pseudo-sentences (this notion not being relevant with spoken language).

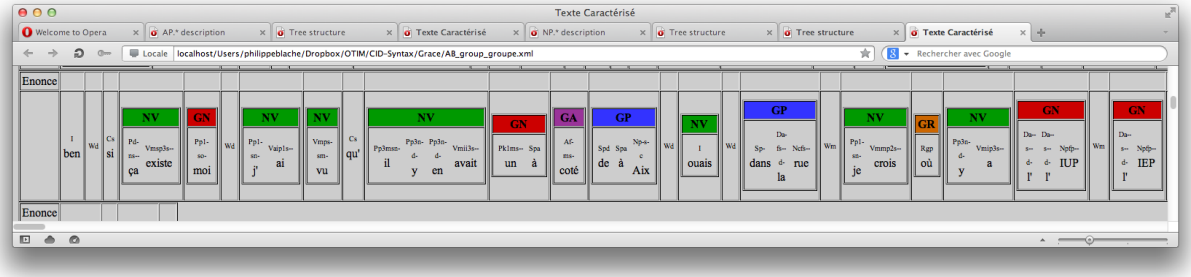
The category and chunk counts for the whole corpus are summarized in the following table:

Category	Count	Group	Count
Adverb	15123	AP	3634
Adjective	4585	NP	13107
Auxiliary	3057	PP	7041
Determiner	9427	AdvP	15040
Conjunction	9390	VPn	22925
Interjection	5068	VP	1323
Preposition	8693	*Total*	63070
Pronoun	25199		
Noun	13419	Soft Punctuation	9689
Verb	20436	Strong Punctuation	14459
Total	114397	Total	24148

The following example illustrates such an encoding in Praat (this format being generated automatically by the chunker). The first tier shows tokens as they have been transcribed, the second one corresponds to tokens as they can be found in the lexicon (especially for locutions, compounds, etc.). The third tier indicates the pseudo-punctuation: weak punctuation, playing the role of a comma, is indicated with "`Wm`", strong punctuation with "`Wd`". The next tier encodes POS tagging: one can see the kind of morpho-syntactic feature vector used here, the category itself being represented in a human-readable format in the tier right after.

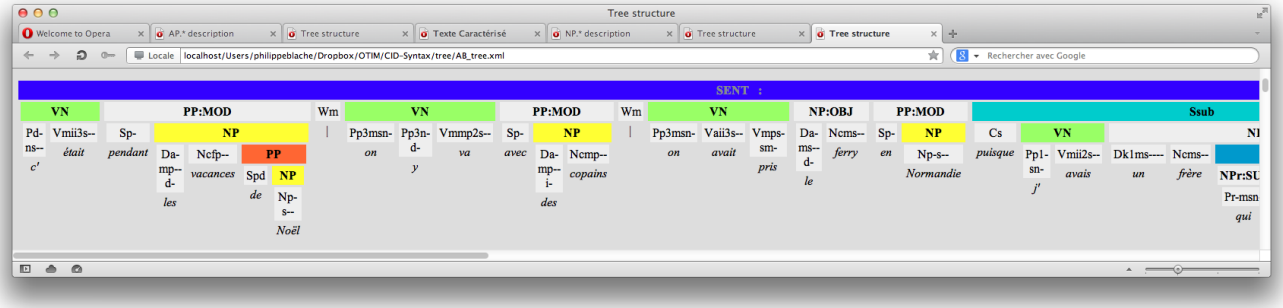


The next figure represents chunks in an html format (they are also directly encoded in textgrid, as shown in the figure above). This representation follows the PEAS convention [Gendner et al., 2003], used during the chunking evaluation campaign Easy [Paroubek et al., 2006]. Chunks, especially when parsing spoken language, are usually short, but their advantage is that they enable to identify the main syntactic constituents. In particular, they are useful when building the syntactic relations, that are not necessarily specified between words, but set of words.



The corpus has also been parsed in order to build a deeper representation in terms of trees. At this stage, no specific pre-processing having been done, in particular for disfluencies

which are automatized, the result is only indicative but can be useful for the utterances with a sufficient level of syntactic construction (which is not always the case). However, disfluencies have been extensively (manually) annotated in a large part of the CID. We benefited from of this information in the parsing process. The following figure gives an example.



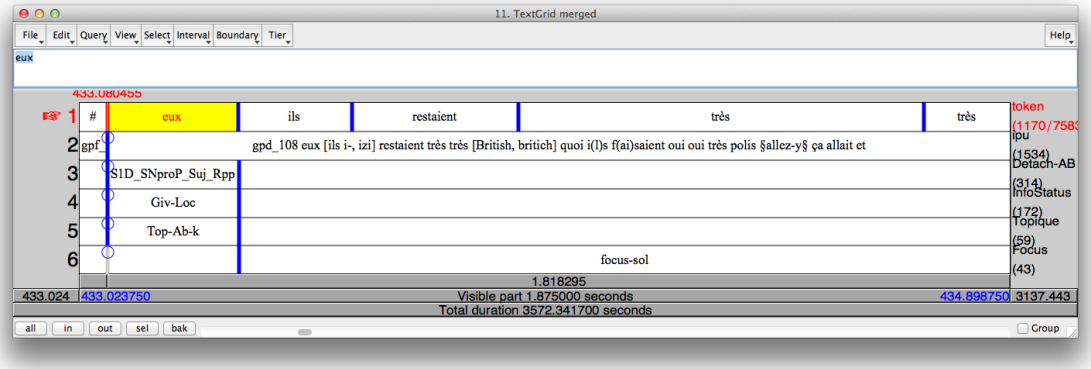
Besides automatic syntactic annotations, syntactic description also relies on the annotation of specific constructions. We worked on one of them: detachments. This annotation has been done for three of the dialogues in the CID. The phenomena annotated here are of different types:

- *Dislocation*: one element has been extracted to the right or the left of the sentence. It can be expressed in an anaphoric relation with a resumptive clitic in the sentence, agreeing with it (ex: "*Chocolate, I hate that*").
- *Cleft*: the extracted element appears to the left of the sentence within a "it ... wh-" structure (ex: "*It is John who married Ann*").
- *Pseudo-cleft*: of the form wh-clause + be + X (ex: "*What he wanted to do was to travel*").
- *Binary constructions*: one element is realized before the sentence, semantically related with it, but not syntactically directly built (ex: "*Being sick, I don't like*").

We use the following feature values to encode these different phenomena:

Detachment type	Dislocation	D
	Non dislocation	nD
	Cleft	CV
	Pseudo-cleft	PSCV
	Binary relation	B
Detached category	SN, SNrel, SNproP, SNproD, SNproQ, SP, SA, SAdv, SV, Ph	
Function	Suj, Odir, Oind, Loc, Adj	
Resumptive element	nR (no resumptive), Rxx (xx : type of the res. element)	

The example below illustrates a dislocation. The feature vector indicates that the dislocated element "eux/themselves" is a personal pronoun, subject and with an anaphoric relation with a clitic (in this case "ils/they")).



5.4 Gesture

The formal model we used for the annotation of hand gestures in Anvil is adapted from the specification files created by [Kipp, 2006] and the MUMIN coding scheme [Allwood et al., 2005]. Both models already integrated McNeill's research on gesture [McNeill, 1992, McNeill, 2005]. The changes we made concerned rather the organization of the different information types and the addition of a few values for a description adapted to the CID. For instance, we added a separate track 'Symmetry' to be able to say if the gesture was single or two-handed. In case of a single-handed gesture, we coded it in its 'Hand_Type': left or right hand. In case of a two-handed gesture, we coded it in the left Hand_Type by default if both hands moved in a symmetric way or in both Hand_Types if the two hands moved in an asymmetric way. For each hand, the scheme has a number of 10 tracks, enabling to code phases, phrases for which we allowed the possibility of a gesture pertaining to several semiotic types using a boolean notation, lexicon (gesture lemmas, [Kipp, 2006]), shape and orientation of the hand during stroke, gesture space (where the gesture is produced in the space in front of the speaker's body [McNeill, 1992] and contact (hand in contact with the body of the speaker, of the addressee, or with an object). At last, we added three tracks to code the hand trajectory (adding the possibility of a left-right trajectory to encode two-handed gestures in a single Hand_Type, and thus save time in the annotation process), quality (fast, normal or slow) and amplitude (small, medium and large), as a gesture may be produced away from the speaker in the extreme periphery, but have a very small amplitude if the hand was already in this part of the gesture space during the production of a preceding gesture.

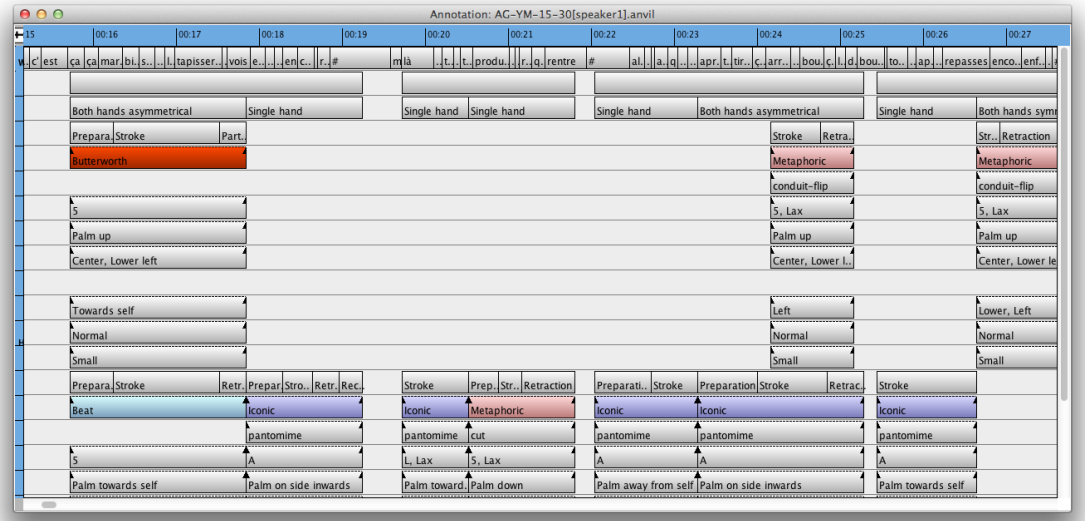
5.4.1 Hands

$$
\textit{Hands}\begin{bmatrix}
\textsc{Symmetry}\ \textit{symmetry} & \\
\textsc{hands\_type} & \begin{bmatrix}
\textsc{Phase}\ \textit{phase} & \\
\textsc{Gesture}\ \textit{gesture} & \\
\textsc{HandShape} & \begin{bmatrix} \textsc{Shape} \\ \textsc{Laxness}\ \textit{boolean} \end{bmatrix} \\
\textsc{HandOrientation}\ \textit{orientation} & \\
\textsc{Space} & \begin{bmatrix} \textsc{Region}\ \textit{region} \\ \textsc{Coordinates} \end{bmatrix} \\
\textsc{Contact} & \\
\textsc{MovementQuality} & \begin{bmatrix} \textsc{Trajectory}\ \textit{trajectory} \\ \textsc{Velocity}\ \textit{velocity} \\ \textsc{Amplitude}\ \textit{amplitude} \end{bmatrix}
\end{bmatrix}
\end{bmatrix}
$$

symmetry:	*{Both hands symmetrical, Both hands asymmetrical, Single hand}*
phase:	*{Preparation, Stroke, Hold, Retraction, Partial Retraction, Recoil, Beat}*
gesture:	*{Adaptor, Iconic, Metaphoric, Deictic, Emblem, Butterworth, Beat, ...}*
orientation:	*{ Palm up, Palm down, Palm towards self, Palm away from self, ...}*
region:	*{ Center center, Center, Periphery, Extreme periphery}*
coordinates:	*{ Right, Left, Upper, Lower, Upper right, Lower right, Upper left, ...}*
contact:	*{ Forehead, Hair, Cheek, Chin, Eyes, Eyebrow, Nose, Ear, Mouth, Neck, ...}*
trajectory:	*{ Upper, Lower, Right, Left, Upper right, lower right, Upper left, Lower left, ...}*
velocity:	*{ Normal, Fast, Slow}*
amplitude:	*{ Small, Medium, large}*

Whenever the value in the Symmetry track has been assigned "Both hands symmetrical", the description has been made for the left hand by default, assuming that the right hand would have similar values in terms of hand shape, movement velocity or amplitude, for instance. Some values like "upper left-right" allow the notation of mirror movements. Both hands were encoded when they were asymmetrical. The movements were annotated for the corresponding hand when the value was "Single hand". All the tiers listed in the theoretical description have been annotated for the following files (each speaker annotated independently):

So far, 75 minutes involving 6 speakers have been annotated, yielding a total number of 1477 gestures. The onset and offset of gestures correspond to the video frames, starting from and going back to a rest position. The example below illustrates the encoding of hand gestures in Anvil:



5.4.2 Head

$$
\textit{head}\left[\begin{array}{l}
\textsc{Facial Expression}\left[\begin{array}{l}
\textsc{Eyebrows}\left[\begin{array}{l}\textsc{Eyebrows\_movements}\ \textit{Eyebrows\_movements}\\ \textsc{GestureSpeed}\ \textit{Gesture\_speed}\end{array}\right]\\
\textsc{Eyes}\left[\textsc{Shape}\ \textit{Eyes\_shape}\right]\\
\textsc{Lips}\left[\textsc{Shape}\ \textit{Lips\_shape}\right]\\
\textsc{Global\_Expression}\ \textit{Global\_expression}
\end{array}\right]\\
\textsc{Gaze}\left[\begin{array}{l}\textsc{Direction}\ \textit{Gaze\_direction}\\ \textsc{Side}\ \textit{Gaze\_side}\\ \textsc{Global}\ \textit{Gaze\_global}\end{array}\right]\\
\textsc{Gesture}\left[\begin{array}{l}\textsc{Phrase}\left[\begin{array}{l}\textsc{Head\_movement}\ \textit{Head\_movement}\\ \textsc{Head\_direction}\ \textit{Head\_direction}\\ \textsc{Head\_side}\ \textit{Head\_side}\\ \textsc{Speed}\ \textit{Gesture\_speed}\\ \textsc{Global}\ \textit{Head\_global}\end{array}\right]\\ \textsc{Phase}\ \textit{Movement\_phase}\end{array}\right]\\
\textsc{Chin\_movement}\ \textit{Chin\_movement}
\end{array}\right]
$$

Eyebrows_movements:	{ *Frowning, rising*}
Gesture_speed:	{ *Slow, Fast*}
Eyes_shape:	{ *ExtraOpen, ClosingBoth, Closing One, Closing Reapeated, Other*}
Lips_shape:	{ *Circle, Drawn, Smile, Laughter*}
Global_expression:	{ *Faint Smile, Smile, Large Smile, Laughter*}
Gaze_direction:	{ *Up, Down, Sideways, Wandering, Towards adressee, Towards object*}
Gaze_side:	{ *Left, Right*}
Gaze_global:	{ *All Gaze Directions, Most Frequent Gaze Poses, ...*}
Head_movement:	{ *Nod, Jerk, Tilt, Turn, Waggle, Pointing, Other*}
Head_direction:	{ *Up, Down, Sideways, Wandering, Towards adressee, Towards object*}
Head_side:	{ *Left, Right*}
Head_global:	{ *All Head Directions, All Head Poses*}
Movement_phase:	{ *Preparation, Stroke, Hold, TurnRepeted, Retraction*}
Chin_movement:	{ *Pointing*}

At the moment, head movements, gaze directions and facial expressions have been coded in 15 minutes of speech yielding a total number of 1144 movements, directions and expressions, to the exclusion of gesture phases. The onset and offset of each tag are determined in the same way as for hand gestures.

5.4.3 Posture

$$
postures\left[\begin{array}{l}
\textsc{Arms}\left[\begin{array}{l}
\textsc{ArmsHeight}\ \textit{ArmsHeight}\\
\textsc{ArmsDistance}\ \textit{ArmsDistance}\\
\textsc{ArmsRadialOrientation}\ \textit{ArmsRadialOrientation}\\
\textsc{ArmsRadialZ}\ \textit{ArmsRadialZ}\\
\textsc{ArmsSwivel}\ \textit{ArmsSwivel}\\
\textsc{ForearmHandOrietation}\ \textit{ForearmHandOrietation}\\
\textsc{ArmsTouch}\ \textit{ArmsTouch}
\end{array}\right]\\
\textsc{Shoulder}\left[\textsc{ShoulderType}\ \textit{ShoulderType}\right]\\
\textsc{Trunk}\left[\textsc{TrunkType}\ \textit{TrunkType}\right]\\
\textsc{Legs}\left[\begin{array}{l}
\textsc{LegsHeight}\ \textit{LegsHeight}\\
\textsc{LegsDistance}\ \textit{LegsDistance}\\
\textsc{LegsSwivel}\ \textit{LegsSwivel}\\
\textsc{LegsRadialOrientation}\ \textit{LegsRadialOrientation}\\
\textsc{LegToLegDistance}\ \textit{LegToLegDistance}\\
\textsc{CrossedLegs}\ \textit{CrossedLegs}
\end{array}\right]
\end{array}\right]
$$

ArmsHeight	{ *Above head, Head, Shoulder, Chest, Abdomen, Waist, Hip/Buttock, ...}*
ArmsDistance	{ *Far, Normal, Close, Touch}*
ArmsRadialOrientation	{ *Behind, Out, Side, Front, Inward, Inside}*
ArmsRadialZ	{ *Forward, Obverse, Downward, Reverse, Backward, Upward}*
ArmSwivel	{ *Touch, Normal, Out, Orthogonal, Raised}*
ForearmHandOrient	{ *Palm up, Palmdown, Palm toward self, Palm away ...}*
ArmTouch	{ *Head, Arm, Trunk, Leg, Furniture, Clothes, Nottouching}*
ShoulderType	{ *Raise left shoulder, Riase right shoulder, Raise shoulders, Lower left ...}*
TrunkType	{ *Lean forward, Lean backward, Turn toward person, Turn away from , ...}*
LegsHeight	{ *Chest, Abdomen, Belt, Buttock, Thigh}*
LegsDistance	{ *Feet behind Knee, Feet in fron of Knee}*
LegsSwivel	{ *Feet outside Knee, Feet inside Knee }*
LegsRadialOrientation	{ *Behind, Out, Side, Front, Inward}*
LegToLegDistance	{ *Knees apart Ankles together, Knees together ...}*
CrossedLegs	{ *Ankle over thigh, At knees, At ankles, Feet over feet, Gross legged }*

Our annotation scheme considers, on top of chest movements at trunk level, attributes relevant to sitting positions (due to the specificity of our corpus). It is based on the *Posture Scoring System* (Bull, 1987) and the *Annotation Scheme for Conversational Gestures* (Kipp et al., 2007). Our scheme covers four body parts: arms, shoulders, trunk and legs. Seven dimensions at arm level and six dimensions at leg level, as well as their related reference points we take in fixing the spatial location, are encoded.

Moreover, two dimensions were added to describe the arm posture in the sagittal plane as well as the palm orientation of the forearm and the hand respectively. Finally, we added three dimensions for leg posture: height, orientation and the way in which the legs are crossed in sitting position.

We annotated postures on 15 minutes of the corpus involving one pair of speakers, leading to 855 tags with respect to 15 different spatial location dimensions of arms, shoulder, trunk and legs.

Annotation	Time (min.)	Units
Transcript	480	-
Hands	75	1477
Face	15	634
Gaze	15	510
Posture	15	855
Reported Speech	180	
Commmunication Function	6	229

We performed a measure of inter-reliability for three independent coders for Gesture Space. The measure is based on Cohen's corrected kappa coefficient for the validation of coding schemes [Carletta, 1996].

Three coders annotated three minutes for *GestureSpace* including *GestureRegion* and *GestureCoordinates*. The kappa values indicated that the agreement is high for *GestureRegion* of right hand (kappa = 0.649) and left hand (kappa = 0.674). However it is low for *GestureCoordinates* of right hand (k= 0.257) and left hand (k= 0.592). Such low agreement of *GestureCoordinates* might be due to several factors. First, the number of categorical values is important. Second, three minutes might be limited in terms of data to run a kappa measure. Third, GestureRegion affects GestureCoordinates: if the coders disagree about GestureRegion, they are likely to also annotate GestureCoordinates in a different way. For instance, it was decided that no coordinate would be selected for a gesture in the center-center region, whereas there is a coordinate value for gestures occurring in other parts of the GestureRegion. This means that whenever coders disagree between the center-center or center region, the annotation of the coordinates cannot be congruent.

5.5 Discourse

Concerning discourse units, the annotation campaign involved naive annotators that have segmented the whole corpus. This was realized thanks to a discourse segmentation guidelines, inspired from [**?**] but largely adapted to our interactional spoken data and simplified to be used by naive annotators. The guidelines combined semantic (eventualities identification) and discourse (discourse markers) and pragmatic (recognition of specific speech acts) instructions to create the segmentation. Such a mixture of levels has been made necessary by the nature of the data featuring both rather monologic narrative sequences and highly interactional ones. Manual discourse segmentation with our guidelines has proven to be reliable with κ-scores ranging between 0.8 and 0.85.

Discourse units definition We took a rather semantic view on the definition of a discourse unit. A discourse unit is a segment describing an eventuality (1) or a segment bearing a clear and proper communicative function (2). Discourse markers are also used in the guidelines.

(1) **Eventualities**

a. [on y va avec des copains]$_{du}$ [on avait pris le ferry en Normandie]$_{du}$ [puisque j'avais un frère qui était en Normandie]$_{du}$ [on traverse]$_{du}$ [on avait passé une nuit épouvantable sur le ferry]$_{du}$
[we going there with friends]$_{du}$ [we took the ferry in Normandy]$_{du}$ [since I had a

brother that was in Normandy$]_{du}$ *[we cross*$]_{du}$ *[we spent a terrible night on the ferry*$]_{du}$

(2) **Clear Communicative Function**

a. [Locuteur1: Tu vois où c'est?]$_{du}$ [Locuteur2: oui]$_{du}$
Speaker 1: You know where it is? Speaker 2: Yes

b. [Locuteur1: Je ne voulais pas les déranger]$_{du}$ [Locuteur2: oui bien sûr]$_{du}$
Speaker 1: I did not want to disturb them ; Speaker 2: Yes of course

We distinguished between several units in discourse: *discourse units* and *abandoned discourse units*. The later are units that are so incomplete that it is impossible to attribute them a discourse contribution. They are distinguished from *false starts* (that are included in the DU they contributed) by the fact that the material they introduced cannot be said to be taken up in the following discourse unit.

(3) Abandoned discourse units
[et euh mh donc t(u) avais si tu veux le sam- + le]$_{adu}$ [pour savoir qui jouait tu (v)ois]$_{du}$
[and err mm so tu had if you want the sat- + the$]_{adu}$ *[in order to know who play you see*$]_{du}$

Annotation process The creation of the guidelines had been an iterative process. Starting from [**?**], a discourse annotation manual for written text, we modified the manual by removing rare cases in spoken language and adding specific spoken phenomena (such as turn alternation that plays a role in the definition of the units). We used this first version of the manual to segment 10 minutes of conversation. We then updated it and run a first annotation round with four annotators working on 15 minutes of 2 different files. A debriefing session was organized and the segmentations were checked. Mostly this session provided the annotators with much more examples they will use intuitively later. A second annotation round was performed on one hour of data. Again a long debriefing session was organized. After that, the annotators worked independently on the data. The annotation period was about 2 months for annotating a little more than 4 conversation of one hour. All the data is at least double-segmented and some parts have up to 4 concurrent annotations.

The segmentation was performed on time-aligned data from both participants to the conversation but without access to the signal. This decision was made because we wanted prosody and discourse annotation to be as independently as possible. Ideally, we wanted to perform the segmentation based on orthographic transcripts only. However, after running a short pilot based on transcripts only we realized that the conversation were simply impossible to follow without timing information (mostly because complex intertwinement of speaker's contributions). The segmentation was therefore done with a tier-based tool (Praat, [Boersma and Weenink, 1996]) but without providing the signal itself. The tiers provided were the IPUs from both speakers, the corresponding tokens and two empty tiers for performing the annotation. The Discourse Units (DU) boundaries were instructed to be anchored on token boundaries. As a consequence, IPU can be seen as a superfluous potential source of bias, however simply reading the tokens sequences is rather tiring and time consuming over large period of time because need of constantly adapting the zoom level to be able to read the

$$disc\_unit \begin{bmatrix} \text{LOCATION } interval \\ \text{TYPE } \{\text{DU, ADU}\} \\ \text{PROPERTY } \{\text{NORMAL, PARENTHETICAL}\} \\ \text{CONSTITUENTS } list(tokens) \end{bmatrix}$$

tokens. IPUs on the other hand, with their bigger size are relatively convenient for reading. The segmentation time has been measured to be 10 to 15 times the real time.

Description of the tiers We used two tiers for the annotation: one for the base discourse units and one for handling discontinuities generated by parentheticals and disfluencies. Indeed, these phenomena are able to be inserted within a discourse without necessarily splitting it functionally. A single tier is not able to represent such structure (at least if no mechanism such as joining relation is provided). Theoretically, two tiers are therefore necessary. However, in practice coders used rarely the possibility of discontinuous units and with poor agreement.

6 Application: backchannels

Backchannel signals (BCs) provide information both on the partner's listening and on the speaker's discourse processes: they are used by the recipient to express manifest attention to the speaker in preserving the relation between the participants by regulating exchanges. They also function as acknowledgement, support or attitude statement, and interactional signals in punctuating/marking specific points or steps in the elaboration of discourse. At last, if ten years ago they were still considered as fortuitous (i.e. they were supposed not to be acknowledged by the speaker), other studies showed that they have a real impact on the speaker's discourse [Fox Tree, 1999].

Although they can be verbal ("*ouais*", "*ok*", etc.), vocal ("*mmh*") or gestural (nods, smiles), most of the studies on BCs only concern one modality. Our own general aim in studying BCs is to integrate the different nature of BCs and to analyze them in two complementary approaches of BCs: firstly to draw up a formal and functional typology (to recognize and automatically label BCs in a database, as well as understand more accurately the human-human and human-machine communication strategies ([Allwood and Cerrato, 2003]) secondly to have a better understanding of the "context of occurrence" which can also inform the function of BCs and contribute to the study of the turn-taking system.

The following example is a particularly good illustration of the interest of a multimodal study of BCs. This passage from a conversation between two male speakers is situated at the very beginning of the interaction and at this point each speaker is particularly concerned with the task given to them, i.e. tell something funny which happened to you. Speaker 1 comes up with a story out of the blue but it takes some time before he can find a proper formulation (until the end of IPU_60). Among the many levels of annotation, we focalized on prosody, conversation organization (TCUs) and some gestures which were relevant to the particular study of BCs.

Previous studies showed that backchannels tended to appear after a complete syntactic unit. However, it would be more adequate to say that backchannels usually appear after a point of syntactic, prosodic and pragmatic completion which is why we decided to consider TCUs (as described in section 3.2.3, see [Portes and Bertrand, 2006]) rather than syntactic units in

this particular study. At the prosodic level, several studies have shown that pitch contours are used not only in the composition of turn-constructional units (TCUs), but also as turn-holding and turn-ending resources ([Ward, 1996]; [Ward and Tsukahara, 2000]; [Caspers, 2003]). More specifically, [Portes and Bertrand, 2006] have shown that the rising major continuation contour in French is one of the contours regularly associated with BCs. Besides, other studies showed that the gestures associated with BCs are typically head movements (the most frequent gestures observed), facial expressions, as well as gaze direction – whether there is or is no mutual gaze between the participants –, such as [Allwood and Cerrato, 2003].

In the example transcribed below, we are especially interested in Speaker 1's non final TCU "*quand j'allais à l'école (when I used to go to school)*" and Speaker 2's gestural backchannel (head nod).

	quand enfin fait souvent enfin quand j'(é)tais (en)fin moins
	main(te)nant mais quand j'(é)tais je faisais souvent (en)fin bref
	c'était un rêve (en)fin pas ouais c'était un rêve + et des fois ça
Sp1	m'arrivait quand # en fait c'est bon **quand j'allais à l'école** en fait
	je mh sur le trajet au bout d'un (m)o(m)ent je me d(i)sais p(u)tain
	d'merde j- j'ai oublié d'enlever les chaussons ou a(l)ors j'ai euh j- en
	/p-/ en pantalon de pyjamas quoi tu vois
Sp2	ou(ais) ouais

```
    <track name=''Gestures Sp1.Hands.Phrases''
type=''primary''>
    ...
    <el index=''4'' start=''19.52'' end=''21.36''>
    <attribute name=''Semiotic Type''>Metaphoric</attribute>
    <attribute name=''Hand''>Both Hands
Symmetrical</attribute></el>
    ...
    <track name=''Gestures Sp1.Hands.Phases''
type=''span'' ref='' Gestures Sp1.Hands.Phrases''>
    ...
    <el index=''8'' start=''19.52'' end=''20.56''>
    <attribute name=''Phase''> Preparation</attribute></el>
    <el index=''9'' start=''20.56'' end=''20.84''>
    <attribute name=''Phase''>Stroke</attribute></el>
    ...
    <el index=''10'' start=''20.84'' end=''21.36''>
    <attribute name=''Phase''> Retraction</attribute></el>
    ...
    <track name=''Gestures Sp1.Gaze''
type=''primary''>
    <el index=''2'' start=''20.52'' end=''21.68''>
    <attribute name=''Direction''>Towards
partner</attribute></el>
    ...
    <track name=''Gestures Sp2.Gaze''
type=''primary''>
    ...
    <el index=''7'' start=''9.92'' end=''28.48''>
    <attribute name=''Direction''>Towards
partner</attribute></el>
    ...
    <track name=''Gestures Sp2.Head''
type=''primary''>
    <el index=''0'' start=''21.12'' end=''21.68''>
    <attribute name=''Movement Type''>Nod</attribute>
    <attribute name=''Semantic Function''>Continuer</attribute>
    <attribute name=''Vertical Plane''>Down</attribute>
    <attribute name=''Frequency''>Single</attribute
></el>
    ...
```

The TCU noted in bold print in the orthographic transcription of the example could have been the end of non-final TCU "*et des fois ça m'arrivait quand (and sometimes what happened to me when)*". The syntactic structure is apparently abandoned and Speaker 1 changes his course by pronouncing an unexpected "*en fait c'est bon (oh yeah right)*" similar to what is currently considered as a self-correction to put an end to all previous hesitations and start anew. The TCU in bold print is then a non-final one framing the story to come and pronounced with a Major Continuation Rise. It is followed by a silent pause during which the speaker ends an interesting metaphoric gesture. This gesture was preceded by other metaphoric gestures (the speaker holds both hands in a spherical shape in front of his torso moving them symmetrically from one side to the other during the whole hesitant part of his speech). Right at the beginning of "*en fait c'est bon*" both hands come back in front of his torso and he lowers them: the gesture is metaphoric in the sense that it represents the speaker's ideas moving from one side to the other, meaning he hesitates, and then putting both hands down (putting the *idea* down) as if to say "*that's it, I know what I'm going to say now*". At the gaze level, during the whole hesitant part, he doesn't look at his partner. Instead he is looking right in front of him yet not at his gesture, and his gaze returns to his partner towards the end of the stroke of the metaphoric gesture, just before the retraction phase of the gesture (when both hands return to a rest position on the speaker's lap). In the meanwhile, Speaker 2 – who is during the whole story in the listener position – is gazing constantly at Speaker 1 [Kendon, 1967], for the correlation between gaze and participant status). The backchannel is a gestural one in the shape of a single slight head nod. It is produced by Speaker 2 during the silent pause and its apex (moment of maximal extension of the gesture before retraction) coincides precisely with the end of the metaphoric gesture produced by Speaker 1. The semantic function of the nod is that of a continuer with a double function of *acknowledgement* of the story and "*go on*" meaning. Immediately after the gestural backchannel, Speaker 1 turns again his gaze away from his partner and resumes his story. Only at the end of the story with a final TCU does Speaker2 produce the vocal backchannel "*ouais ouais*" (lit. "*yeah yeah*" meaning "*oh yeah*") which semantic function is this time that of an acknowledgement.

What can be generalized from this particular example is that the interaction between the different levels considered informed us on the "occurrence context" of the BC which production was encouraged by the Major Continuation Rise together with gaze oriented towards partner and retraction phase of the previously initiated hand gesture. This particular gesture sets down an idea and the continuer nod allows Speaker 1 to elaborate on his story. We can deduce that if one of these conditions had been missing, there wouldn't have been a backchannel here, as shown by the preceding example of Major Continuation Rise which is not accompanied by any backchannel since there is no mutual gaze between the participants. However, one has to keep in mind that the head nod does not have the unique function of continuer. In another context, it may have had another function such as acknowledgement or assessment for example. This is also true of verbal or vocal backchannels. The example shows the importance of an analysis which takes into account as many layers of annotation as possible in several linguistic fields since all the information contributes to the constitution of "context" and of the collective construction of discourse. It also shows that the existing functional categories do not explain every occurrence of backchannels since the multimodal analysis reveals a subcategory which has not been described by the traditional dichotomy between the functions of continuer and assessment [Schegloff, 1982].

7 Genericity, Interoperability

One of the main interests in using an abstract annotation scheme encoded in typed feature structures is that it provides a efficient tools for maintaining the coherence of the annotaitons both at the theoretical and the practical level. First, as already underlines, this scheme proposes an homogeneous framework for representing information coming from the different linguistic domains. Our proposal is one of the rare attempts to build a general scheme covering all the domains. Moreover, this scheme can be also used in order generate interoperable annotations. We propose in this section some preliminary steps in this direction.

As it is usually the case in such broad-coverage projects, annotations can be generated either automatically or manually. In this last case, annotations are created by means of different tools, depending on the domain to annotate, the experience of the annotator, etc. In our project, most of the annotations have been created using Praat, Anvil or Elan. In such an environment, maintaining a coherent and consistent annotation system becomes difficult, not to say impossible, due to lack of interoperability between the systems, each one using its own encoding system. Even if it is possible in some cases to import annotations from different formats (for example importing a Praat tier into Anvil), it remains globally impossible to export all modifications. For example, if we add a new phoneme into the phonetic transcription, this modification has to be propagated to all the other annotations linked in some way to this level: syllables, prosodic units, but not token or discourse relations.

This is a very complex problem. A first experience has been proposed by the different software developers [Schmidt et al., 2009], starting from the AIF format. The idea was there to propose a "greatest common denominator" between the different formats. However, this attempts remained theoretical, mainly because of the difficulty in implementing concretely such an exchange. One of the main problems indeed come from not all informations are encoded into the different format and translating a representation from one to another can lead to an information loss. However, the idea to specify an exchange format seems to be the right direction to explore.

Our proposal does not consist in a specific tool nor even in a the elaboration of a generic format. We simply underline the fact that knowing the overall organization of knowledge representation thanks to the TFS abstract scheme, it becomes possible to generate easily an XML representation of the annotations, whatever their original domain. The mechanisms consists in associating an XML description to the TFS scheme. As an example, the following figure illustrates a (partial) xml schema associated to the intonational phrase description:

$$\begin{bmatrix} \textsc{label}\ \texttt{IP} \\ \textsc{constituents}\ \text{list}(ap) \\ \textsc{contour}\ \begin{bmatrix} \textsc{direction}\ \textit{string} \\ \textsc{position}\ \textit{string} \\ \textsc{function}\ \textit{string} \end{bmatrix} \end{bmatrix}$$

```
<xs:complexType name=''IntonationalPhrase''>
  <xs:complexContent>
    <xs:extension base=''ProsodicPhrase''>
    <xs:sequence>
      <xs:element name=''constituents''>
        <xs:complexType>
          <xs:sequence>
            <xs:element name=''accentual\_phrase'' type=''AccentualPhrase''/>
          </xs:sequence>
        </xs:complexType>
      </xs:element>
      <xs:element name=''contour'' type=''Contour''/>
    </xs:sequence>
    </xs:extension>
  </xs:complexContent>
</xs:complexType>
```

Thanks to such description, we can propose straightforward translation from the original

encoding (for example in Praat) into an xml form, as in the example below:

```
  class = ''IntervalTier''
  name = ''at_phrasing''
  xmin = 0
  xmax = 3573.6
  intervals: size = 2782
...
  intervals [2]:
  xmin = 0.78
  xmax = 1.7559754641684542
  text = ''ip''
...
  intervals [5]:
  xmin = 2.6703535937364578
  xmax = 3.329971301020408
  text = ''ip''
```

```
  class = ''TextTier''
  name = ''at_ctr''
  xmin = 0
  xmax = 3573.6
  points: size = 2118
  points [1]:
  time = 1.7559754641684542
  mark = ''RT''
...
  points [3]:
  time = 3.329971301020408
  mark = ''F''
```

```
<IntonationalPhrase index=0>
   <localisation start=0.78 end=1.7559 />
   <contour type=RT time=1.7559  />
</IntonationalPhrase>
   ...
<IntonationalPhrase index=5>
   <localisation start=2.6703 end=3.3299 />
   <contour type=F time=3.3299 />
</IntonationalPhrase>
```

Encoding the entire annotations in XML following the XML schema (then the TFS description) ensures not only an homogenous encoding of the entire annotation set, but also offers the possibility to use XML-based querying tools (such as XQuery) in order to extract information from the entire annotated corpus.

8 Conclusion

The case study presented in this paper addresses the entire annotation workflow, starting from raw data (speech and video) until highly enriched resources. We propose for each annotation step different tools or methods making it possible to homogenize the annotation process. The particularity of multimodal corpora is that information comes from different sources, not strictly synchronized or aligned. It is then necessary to specify precisely first the kind of information to be encoded and second how to represent it. We propose to do this by means of an abstract schema encoded with types feature structures. This schema is not only an efficient way to precisely organize knowledge representation, but also makes it possible to represent heterogenous sources of information in a homogeneous framework. Moreover, it enables to translate automatically proprietary format (for example associated to a specific editors such as Praat) into a generic one (following an XML abstract scheme corresponding to the TFS representation).

We have experimented this annotation workflow in the building of the "fully-annotated" CID corpus, gathering precise annotations at many different linguistic levels. The CID is now one of the largest existing resources proposing manually validated annotations for phonetics, prosody, morpho-syntax, discourse, gesture as well as specific phenomena such as disfluencies. The CID is available through the SLDR (Speech and Language Data Repository, `http://www.sldr.org`).

References

[Allwood and Cerrato, 2003] Allwood, J. and Cerrato, L. (2003). A study of gestural feedback expressions. In *First Nordic Symposium on Multimodal Communication*, pages 7–22.

[Allwood et al., 2005] Allwood, J., Cerrato, L., Dybkjaer, L., Jokinen, K., Navaretta, C., and Paggio, P. (2005). *The MUMIN Multimodal Coding Scheme*, pages 129–157. NorFA yearbook 2005.

[Anderson et al., 1991] Anderson, A. H., Bader, M., Gurman Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., and Weinert, R. (1991). The hcrc map task corpus. *Language and Speech*, 34:351–366.

[Bertrand et al., 2007] Bertrand, R., Portes, C., and Sabio, F. (2007). Distribution syntaxique, discursive et interactionnelle des contours intonatifs du français dans un corpus de conversation. *Travaux neuchâtelois de linguistique*, 47.

[Blache and Rauzy, 2008] Blache, P. and Rauzy, S. (2008). Influence de la qualité de l'étiquetage sur le chunking : une corrélation dépendant de la taille des chunks. In *Actes de Traitement Automatique des Langues Naturelles*, pages 290–299, Avignon, France.

[Boersma and Weenink, 1996] Boersma, P. and Weenink, D. (1996). Praat, a system for doing phonetics by computer, version 3.4. Technical Report 132, Institute of Phonetic Sciences of the University of Amsterdam.

[Carletta, 2006] Carletta, J. (2006). Announcing the ami meeting corpus. *The ELRA Newsletter*, 11(1).

[Carletta et al., 2003] Carletta, J., Evert, S., Heid, U., Kilgour, J., Robertson, J., and Voormann, H. (2003). The nite xml toolkit: flexible annotation for multi-modal language data. *Behavior Research Methods, Instruments, and Computers*, 35(3).

[Carletta, 1996] Carletta, J. C. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

[Carpenter, 1992] Carpenter, B. (1992). *The Logic of Typed Feature Structures.* Cambridge University Press.

[Caspers, 2003] Caspers, J. (2003). Local speech melody as a limiting factor in the turn-taking system in dutch. *Journal of Phonetics*, 31:251–276.

[Fohr et al., 2004] Fohr, D., Mella, O., Cerisara, C., and Illina, I. (2004). The automatic news transcription system: Ants, some real time experiments. In *INTERSPEECH-2004*, pages 377–380.

[Fox Tree, 1999] Fox Tree, J. E. (1999). Listening in on monologues and dialogues. *Discourse Processes*, 27(1):35–53.

[Gendner et al., 2003] Gendner, V., Illouz, G., Jardino, M., Monceaux, L., Paroubek, P., Robba, I., and Vilnat, A. (2003). PEAS, the first instantiation of a comparative framework for evaluating parsers of french. In *Research Notes of EACL 2003*, Budapest, Hongrie.

[Hirst, 2007] Hirst, D. (2007). A praat plugin for momel and intsint with improved algorithms for modelling and coding intonation. In *ICPhS XVI*.

[Hirst et al., 2000] Hirst, D., Di Cristo, A., and Espesser, R. (2000). Levels of representation and levels of analysis for intonation. In Horne, M., editor, *Prosody : Theory and Experiment*, pages 51–87. Kluwer Academic Publishers.

[Jun and Fougeron, 1995] Jun, S. and Fougeron, C. (1995). The source ambiguity the accentual phrase and the prosodic structure of french. In *13th ICPhS*, pages 722–725.

[Kendon, 1967] Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica*, 26:22–63.

[Kipp, 2001] Kipp, M. (2001). Anvil - a generic annotation tool for multimodal dialogue. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, pages 1367–1370.

[Kipp, 2006] Kipp, M. (2006). An annotation scheme for conversational gestures : How to economically capture timing and form. In *Proceedings of the Workshop on "Multimodal Corpora" at LREC 2006*.

[Koiso et al., 1998] Koiso, H., Horiuchi, Y., Ichikawa, A., and Den, Y. (1998). An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs. *Language and Speech*, 41.

[Kruijff-Korbayova et al., 2006] Kruijff-Korbayova, I., Gerstenberger, C., Rieser, V., and Schehl, J. (2006). The sammie multimodal dialogue corpus meets the nite xml toolkit. In *proceedings of LREC06*.

[McNeill, 1992] McNeill, D. (1992). *Hand and Mind. What Gestures Reveal about Thought*. University of Chicago Press.

[McNeill, 2005] McNeill, D. (2005). *Gesture and Thought*. University of Chicago Press.

[Paroubek et al., 2006] Paroubek, P., Robba, I., Vilnat, A., and Ayache, C. (2006). Data annotations and measures in EASY the evaluation campaign for parsers in french. In *Proceedings of the 5th international Conference on Language Resources and Evaluation*, pages 314–320, Genoa, Italy.

[Pineda et al., 2002] Pineda, L. A., Massé, A., Meza, I., Salas, M., Schwarz, E., Uraga, E., and Villaseñor, L. (2002). The dime project. In *Proceedings of MICAI2002*, volume 2313. LNAI.

[Portes and Bertrand, 2006] Portes, C. and Bertrand, R. (2006). Some cues about the interactional value of the «continuation» contour in french. In *Discours et Prosodie comme Interface Complexe*.

[Rodriguez et al., 2007] Rodriguez, K., Dipper, S., Götze, M., Poesio, M., Riccardi, G., and Raymond, C. andRabiega-Wisniewska, J. (2007). Standoff coordination for multi-tool annotation in a dialogue corpus. In *proceedings of Linguistic Annotation Workshop*.

[Sacks et al., 1974] Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50:696–735.

[Schegloff, 1982] Schegloff, E. (1982). Discourse as an interactional achievement: Some uses of "uh huh" and other things that come between sentences. In Tannen, D., editor, *Analyzing discourse: Text and talk*. Georgetown University Press.

[Schmidt et al., 2009] Schmidt, T., Duncan, S., Ehmer, O., Hoyt, J., Kipp, M., Loehr, D., Magnusson, M., Rose, T., and Sloetjes, H. (2009). *An exchange format for multimodal annotations*, pages 207–221. Springer.

[Ward, 1996] Ward, N. (1996). Using prosodic clues to decide when to produce back-channel utterances. In *4th International Conference on Spoken Language Processing*, pages 1724–1727.

[Ward and Tsukahara, 2000] Ward, N. and Tsukahara, W. (2000). Prosodic features which cue back-channel responses in english and japanese. *Journal of Pragmatics*, 23:1177–1207.