Multimodal Annotation of Conversational Data

Firsname Lastname
Affiliation
email@address

Abstract

We propose in this paper a broad-coverage approach for multimodal annotation of conversational data. Large annotation projects addressing the question of multimodal annotation bring together many different kinds of information from different domains, with different levels of granularity. We present in this paper the first results of the OTIM project aiming at developing conventions and tools for multimodal annotation.

1 Introduction

Multimodal annotation is faced with the necessity of encoding many different information, from different domains, with different levels of granularity. The main problem is that all these different pieces of information have to be connected, and if possible aligned onto the signal. Doing this is possible provided that data are prepared in this perspective : for example, a precise phonetic and prosodic description requires that each speaker is recorded in a specific channel. Moreover, the transcription has to be done keeping in mind the requirements of all the different annotations (morphosyntax as well as phonetics).

We present in this paper the first results of the OTIM[1] project aiming at developing conventions and tools for multimodal annotation. We show here how such an approach can be applied in the annotation of a large conversational speech corpus. In the first section, we describe the different step of the annotation process, the conventions and the tools helping it. In the second section, we presents the corpus, its annotations and give first evaluations. The last section gives some information about the XML encoding and its interest in the perspective of information treatment.

2 Multimodal Annotation : Tools and Conventions

We present in this section the different tools and convention that we used in our annotation process. They illustrate the kind of process to implement in the perspective of obtaining rich and broad-coverage annotation. Some of these tools already exists, some others have been developed in the OTIM framework.

2.1 Enriched orthographic transcription (EOT)

As it is often the case with large audio corpora (Koiso98), the speech signal is first automatically segmented in inter-pausal units (IPUs), speech blocks surrounded by 200ms silent pauses (this duration is well suited for French). The transcription process takes as input this set of IPUs. It is done following specific conventions, taking into account some remarkable and frequent phonetic phenomena such as non-standard elisions, phoneme substitutions or additions, etc.

Their annotation is necessary in order to study their frequency, variability and impact on the alignment. Moreover, such annotation allows can help in building a lexicon of non-standard phonetic variant, improving their automatic generation. Our transcription conventions derive from that of the GARS (Blanche-Benveniste87) : transcribers were asked to annotate elision, word truncation, silent pauses, filled pauses, liaisons (absence of a standard liaison, presence of an unusual liaison), assimilation phenomenas as well as some specific phenomena, as the realization of schwas in Southern French, laughing sequences, direct reported speech. Overlaps were not annotated : they are automatically located in detecting overlapping IPUs, each speaker being recorded on a separate chan-

[1] OTIM stands for Outils pour le Traitement de l'Information Multimodale (Tools for Multimodal Annotation). This project in funded by the French ANR agency.

nel. The following example illustrates the EOT of a sequence with elisions :

(1) *j'ai on a j'ai p- (en)fin j'ai trouvé l(e) meilleur moyen c'(é)tait d(e) loger chez des amis*
I've we've I've - well I found the best way was to live at friends'

From the EOT two transcriptions are generated automatically : (1) the standard orthographic transcription from which the *orthographic tokens* are extracted to be used for semantics, syntax and discourse analysis and their related tools (POS tagger, parser, etc.) ; (2) a specific transcription from which the *phonetic tokens* are obtained to be used by the grapheme-phoneme converter.

Obviously, EOT is time consuming and increases the transcription duration and moreover there were one initial transcription and two corrections. The (total annotation time) / (speech time) ratio is about 90. However, it guarantees a faithful transcription, providing facilities for further annotations on phenomena (disfluencies, elisions, etc.). It is moreover necessary in order to iprove the phoneme/signal alignment.

2.2 Speech processing tools

Grapheme-phoneme converter : The grapheme-phoneme converter is a dictionary and rule-based system (Di Cristo01). It takes as input a phonetic tokens sequence extracted from the EOT and outputs a sequence of phonemes, coded in Sampa. For example, if the EOT is :

EOT	*j'ai j'ai p- (en)fin [je sais, ch] p(l)us* I have I have p- well I don't remember
Phonetic tokens	*j'ai j'ai p fin ch pus*
Phonemes	*Z E Z E p f e S p y*

Phoneme/signal alignment : From the phoneme sequence generated by the converter and the audio signal, the aligner outputs for each phoneme its time localization. This aligner is HMM-based (Brun04), it uses a set of 10 macro-classes of vowel (7 oral and 3 nasal), 2 semi-vowels and 15 consonants. It relies on acoustic models based on standard French. The minimal duration allowed for a phoneme is 28ms. The alignment is done for each IPU separately. Finally, from the time aligned phoneme sequence plus the EOT, the orthographic tokens is time-aligned.

EOT	t(u)	sais	[je suis, chui]
phon.tok.	t	sais	chui
ortho.tok.	tu	sais	je_suis

Evaluation : We measured the difference between the automatic and manual vowel boundaries of 2 speakers (about 13,000 vowels). 75% of the absolute differences were less than 20 ms at the beginning of the vowel, and less than 23 at the end. This has to be considered in line with the mean duration of a phoneme in this corpus, about 80 ms (Bertrand08).

2.3 Syllabification

The problem consists in identifying the syllable boundaries starting from a phoneme sequence.

Our RBS phoneme-to-syllable segmentation system (Bigi10) is based on 2 main principles :

1. a syllable contains a vowel, and only one
2. a pause is a syllable boundary[2]

These two principles focus the problem on the task of finding a syllabic boundary between two vowels. The novel aspect of the syllabification work is as follows :

– to propose a generic RBS tool to identify syllabic segments from phonemes ;
– to propose relevant rules in the particular context of French spontaneous speech.

The method : Phonemes are first grouped into classes and rules are then divided in two categories : general rules, that give the boundary between 2 vowels depending on the number of consonants between them, and exception rules operating on the specific sequences found between the two vowels.

Tokens	*non dans les parcs c'est un peu limité* no in the parks it is rather restricted
Phonemes	/n Õ d Ã l e p A R k s e t œ̃ p @ l i m i t e/
Classes	NVOVLVOVLOFVOVOVLVNVOV
Output	nÕ / dÃ / le / pAR / kse / tœ̃ / p@ / li / mi / te

This was compared to expert outputs as for example :

nÕ . dÃ . le . pARk . se . tœ̃ . p@ . li . mi . te

This system obtains very interesting results, outperforming other syllabification tools already existing for French.

The tool : The program *LPL-Syllabeur* proposes a stand-off configuration file which allows users to define their own phoneme encoding, their own

[2] Our corpus was first automatically segmented into inter-pausal units (IPU) delimited by silent pauses of 200 ms and more. Human transcribers marked shorter pauses they perceived. In both cases, a pause signals a syllable boundary.

phoneme classes and all the rules. This choice allows the tool to be adapted to other syllabification tasks or other languages without changing any of the implementation (only the external ASCII configuration file).

2.4 Prosody

The Momel-Intsint protocol allows a semi-automatic annotation of prosodic form. The first assumption underlying it, consists in factoring out two components, a macroprosodic component and a microprosodic component, from the f0 curve [Di Cristo, Hirst, 1986]. The macroprosodic component is further modeled via the application of the Momel algorithm [Hirst, Espesser, 1993] [Hirst05]. This modeling is grounded on the definition of target points in time frequency space : these target points correspond to the inflections in the f0 curve where the slope is null (i.e. the first derivative equals zero). To obtain a smooth curve, the pitch targets are linked by a quadratic spline function. This representation is mainly acoustically oriented, though these points could correspond to the sites where the speaker voluntarily changes the direction of the fundamental frequency to achieve his/her communicative goals (VaissiÃÍre 2002). The Momel algorithm is currently implemented under the Praat software [Hirst07].

At the higher, surface phonological level, the f0 targets receive a symbolic coding in terms of the INTSINT prosodic alphabet [Hirst, Di Cristo, 1998]. The INTSINT alphabet comprises 8 distinct symbols : the tonal targets they define are characterised either globally with respect to the speaker's pitch range (via the long-term parameters of key and range ; the corresponding labels are T(op), B(ottom) and M(edium)) or locally, by the reference to the preceding target (H(igher), L(ower), S(ame)). The H and L labels have the iterative variants D(ownstepped) and U(pstepped). This annotation is done semi-automatically, that is, once the plug-in returns the target points, they should be corrected by human annotator before INTSINT encoding. During the correction stage, the points could be modified as to their time position, added or removed. For other details of correction stage see (Nesterenko 2006, Cho 2009 among others).

2.5 Syntactic annotations

The enriched orthographic transcription gives tokens synchronized with speech signal. The orthographic tokens are obtained by filtering non syntactic objects (e.g. laughts, disfluencies, pauses, ...) and form the basic input of the syntactic treatment. A modified version of the syntactic parser StP1 was applied on the data.

The syntactic parser StP1 (Blache&Rauzy, 2008) is a stochastic parser for written French developed at the Laboratoire Parole et Langage. In a first step, it provides for each POS token an automatic annotation of its morphosyntactic category. In the second step, the tokens are grouped in larger units (chunks) following the EASY flat grammar (Paroubek et al., 2006) described in the PEAS guidelines (Gendner et al., 2003). The EASY grammar is rich of six constituents, GN (Noun Phrase), GP (Prepositional Phrase), GR (Adverbial Phrase), GA (Adjective Phrase), NV (Verbal Nucleus), PV (Verbal group introduced by a preposition), organized in the sentence with a flat structure. The StP1 chunker obtains relatively good results on written texts, a F-measure score of 0.94 for the tagging stage and a score 0.92 for the chunking stage, see (Blache&Rauzy, 2008) for more details. These performances are reduced for speech corpora, a F-measure score of 0.79 for chunks formation, but still remain interesting for providing us with an automatic annotation of the syntactic information.

The StP1 chunker has been modified in order to account for the specificities of speech analysis. Two levels of hierarchy were introduced in the syntactic treatment, corresponding to the strong punctuation marks (e.g. full stops, exclamation marks) and weak or soft punctuation marks (such as commas) that can be found in written text but are not annotated in the transcription. The modified stochastic parser thus automatically insert these two kind of frontiers on the basis of the syntactic context. A second modification concerns the lexical frequencies used by the parser model. Some of them, associated to parts of speech with a phatic function for example, have been modified in order to capture phenomena proper to conversational data.

3 Corpus of Interactional Data

We present in this section the application of the previous conventions, recommendations and tools for the annotation of a conversational corpus, called CID (Corpus of Interactional Data, see (Bertrand08)).

3.1 Annotations

Before entering into the detail of the annotations for different linguistic phenomena, it is necessary to underline the different information domains to be annotated (see above) does not rely on the same granularity level and then require a different amount of work. As a consequence, the state of the annotation of each domain in not the same in the CID, this work is still on progress. However, we already have significant results, as it will be shown hereafter.

Syllables : Like phonemes, syllables are one of fundamental linguistic units. Syllables play an important role in phonology and phonetics as well as in the description of phonotactic constraints or the analysis of synchronization phenomena between tonal events and segmental strings. First our corpus was automatically segmented in syllables (see section syllabifier). Next sub-syllabic constituents, onset, nucleus and coda, are identified within each syllable. Another feature describes the syllable structure (V, CV, CCV, etc.) and we specify the syllabic position in polysyllabic words. All these pieces of information will allow us to explore the distribution of syllabic structures, acoustic properties of sounds in different structural positions (Hawkins, 2003) and prosodic (temporal and tonal) properties in spontaneous speech.

Prosody : In our study prosody is interpreted as an organizational system [Beckman, 1996] that could be exhaustively specified via the analysis of tonal and rhythmical layers as well as that of prosodic phrasing [Di Cristo et al., 2004 ; Selkirk, 1995]. The representations developed for the three layers have recently been formalized in numerous phonological studies [Pierrehumbert, 1980 ; Ladd, 1996 among others].

Prosodic phrasing : Prosodic phrasing refers to the structuring of speech material in terms of boundaries and groupings. A universal prosodic hierarchy was proposed by Nespor and Vogel (1986) : at the same time, authors claim that the relevance of each level in describing phonological organisation in a language should be empirically proved and supported by phonological, sandhi or temporal / tonal processes. Our annotation scheme for French supposes the distinction between two levels of phrasing : the level of accentual phrases (AP, (Jun, 2002)) and the higher level of intonational phrases (IP). Such a choice does not mean that we completely exclude from the analysis of French prosody the level of intermediate phrases. Mean annotation time for IPs and APs was 30 minutes per minute.

Prominence : The annotation of prominences relies on syllabic layer. In annotating prominence status of a syllable we choose to distinguish between accentuability (a possibility for syllable to be prominent in realization) and prominence (a syllable is perceived as prominent if it is realized more salient than others). In French the first and the last full syllables (not containing a schwa) of a polysyllabic word could be prominent, though this actual realization depends on the speaker's choices. Accentuability annotation is automatic while prominence annotation is perceptually based.

Tonal layer : Today, intonational phonology [Pierrehumbert, 1980 ; Ladd, 1996] is a reference framework in analysing and annotating tonal phenomena. The main principles of intonational phonology were incorporated into the ToBI prosodic annotation system [Beckman et al., 2005] and since then, ToBI-inspired annotation systems have been proposed for many languages [Jun, 2005]. Given a lack of consensus on the inventory of tonal accents in French, we choose to integrate in our annotation scheme three types of tonal events : a/ underlying tones (for an eventual FrenchToBI annotation) ; b/ surface tones (annotated in terms of MOMel-Intsint protocol Hirst et al 2000) ; c/ melodic contours (perceptually annotated pitch movements in terms of their form and function). The interest to have both manual and automatic INTSINT annotations is that it allows the study of their links.

Hand gestures : Different typologies have been adopted for the classification of gestures, based on the work by Kendon (1980) and McNeill (1992, 2005). The formal model we use for the annotation of hand gestures is adapted from the specification files created by Kipp (2004) and from the MUMIN coding scheme (Allwood et al., 2005). Both models consider McNeill's research on gestures (1992, 2005). The gesture types we are using are mostly taken from McNeill's work. Iconics present "images of concrete entities and/or actions", whereas Metaphorics present "images of the abstract", they "involve a metaphoric use of form" and/or "of space". (McNeill, 2005 : 39). Deictics are pointing gestures and Beats bear no “discernible meaning” and are rather connected with speech rhythm (McNeill, 1992 : 80). Emblems are conventionalized signs and Butterworths are gestures made in lexical retrieval. Adaptors are non verbal gestures that do not participate directly in the meaning of speech since they are used for comfort. Although they are not linked to speech content, we decided to annotate these auto-contact gestures since they give relevant information on the organization of speech turns.

We used the Anvil tool (Kipp, 2004) for the manual annotations. The changes we made to existing specification files concerned the organization of the different information types and the addition of new values adapted to the CID corpus description. For instance, we added a separate track “Symmetry”. 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 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 10 tracks, enabling to code phases, phrases. We allowed the possibility of a gesture pertaining to several semiotic types using a boolean notation. A gesture phrase (i.e. the whole gesture) can be decomposed into several gesture phases i.e. the different parts of a gesture such as the preparation, the stroke (the climax of the gesture), the hold and the retraction (when the hands return to their rest position) (McNeill, 1992). The scheme also enables us to annotate the gesture lemmas (Kipp, 2004 :237), the shape and orientation of the hand during the stroke, the gesture space (where the gesture is produced in the space in front of the speaker’s body (McNeill92), and contact (hand in contact with the body of the speaker, of the addressee, or with an object). We added the 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), gesture velocity (fast, normal or slow) and gesture amplitude (small, medium and large). A gesture may be produced away from the speaker in the extreme periphery, while having a very small amplitude if the hand was already in this part of the gesture space.

Discourse and Interaction : Our discourse annotation scheme relies on multidimensionalframeworks such as DIT++ (Bunt, 2009) and is compatible with the guidelines defined by the Semantic Annotation Framework (Dialogue Act) working group of ISO TC37/4. These proposals have taken advantage of the experience acquired on several dialogue annotation project and their related schema (in particular DAMSL(**?**) and SWBD-DAMSL (**?**)) as well as a thorough empirical and theoretical description concerning the multifonctionality and multidimensionality of communicative behavior (**?**; **?**).

Discourse units include information about their producer, have a form *(clause, fragment, disfluency, non-verbal)*, a content and a communicative function. The same span of raw data may be covered by several discourse units playing different communicative functions. Two discourse units may even have exactly the same temporal extension,due to the multifonctionality that cannot be avoided (Bunt, 2009).

Compared to standard dialogue act annotation frameworks, three main additions are proposed : *rhetorical function*, *reported speech* and *humor*. The additions are due to the nature of our conversational and narrative data. Describing narrative sequences require coherence relations. Our rhetorical layer is an adaptation of an existing schema developed for monologic written data in the context of the ANNODIS project (**?**). Moreover, the storytelling nature of the data results in a considerable amount of reported speech. This phenomenon is quite difficult to deal with in standard dialogue frameworks that have been usually designed for handling task-oriented dialogues. Finally, humor cannot be overlooked since the same sequence or discourse unit may have very different functions in humorous and non-humorous sequence.

Disfluencies : Disfluencies are considered as ruptures in the syntagmatic flow. They are organized around an interruption point (also called *break*), which can occur almost anywhere in the production. Disfluencies can be purely prosodic (lenghtenings, silent and filled pauses, etc.), but they are mainly lexicalized. In this case, they appear as a word or a phrase truncation, that can either be completed (or repaired) or not. We distinguish three parts in a disfluency (see (Shriberg, 1994), (Blanche-Benveniste87)) :

– Reparandum : what precedes the interruption point. This part is mandatory in all disfluencies. We indicate there the nature of the interrupted unit (word or phrase), and the type of the truncated word (lexical or grammatical) ;

- Break interval. It is optional, some disfluencies do not bear any specific event there.
- Reparans : the part following the break, that is supposed to repair the reparandum. We indicate there type of the repair (no restart, word restart, determiner restart, phrase restart, etc.), and its function (continuation, repair without change, repair with change, etc.).

3.2 Quantitative information

Hand gestures 75 minutes of the CID involving 6 speakers have been coded for hand gestures. Whereas each dependent track has been informed, the annotation yielded a total number of 1477 gestures. The numbers of hand gestures per semiotic type are listed in table 3. The onset for each hand gesture corresponds to the first frame in which the hand(s) moves from its rest position whereas the offset corresponds to the first frame in which the hand returns to its rest position when the gesture is produced in isolation. When the gesture is produced in between two other gestures without any return to rest position, its onset corresponds to the first frame in which the hand changes trajectory from the previous gesture (initiates the preparation or stroke of the gesture). Its offset corresponds to the last frame before the hand changes trajectory for the preparation or stroke of the next gesture. One has to keep in mind that due to the granularity of the videos (24 frames per second), the onset and offset of hand gestures are defined less precisely than the onset and offset of speech.

Face and gaze At the present time, 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 way as for hand gestures.

Body Posture Methods of describing postures are relatively domain-related. They vary from grid-based observational studies to technical measures. Psychologists (Scherer82) distinguish posture behavior from action behavior. The posture behavior refers to overall postures (sitting, standing, lying), frontal orientation of trunk (facing, turned away), trunk lean (forward, straight, backward, sideways), arm and leg position (folded arms, uncrossed legs) and feet (flat on floor, under chair, on other knee). In the Posture Scoring system (Bull, 1987), any movement which is taken up and maintained for at least one second is annotated as a posture. This system covers head, arms, trunk and legs. Bull also proposed a second system using a dynamic approach, which describes the posture in terms of a series of movements rather than static positions.

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

TAB. 1 – Recap of some annotations

The previous annotation scheme for the CID corpus (Bertrand et al., 2008) only considered chest movements at trunk level. Aiming at extending the postural sphere, we added a set of tracks and attributes relevant to sitting positions met in the CID 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. As listed in Table 1 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. At arm level, Bull's system mainly distinguishes whether the hand touches or not ; while Kipp's scheme covers four spatial dimensions to capture it. We made a trade-off decision between the two systems : we kept the four dimensions from Kipp's coding scheme with respect to the height, distance, radial orientation and swivel degree of the arm, and created a new track describing the hand touching objects to get back to the ideas of Bull's system. Also, we added two dimensions to describe respectively the arm posture in the sagittal plane and the palm orientation of the forearm and the hand. With respect to the leg posture, we added three dimensions : height, orientation and the way in which the legs are crossed in sitting position.

As for discourse, reported speech has been annotated for 3 hours while the annotation of communicative functions is just starting with about 30 minutes annotated.

Disfluencies Part of the corpus has been annotated (12 mns). At the moment, this annotation is fully manual (we just developed a tool helping the process in identifying disfluencies, but it has not yet been evaluated). Annotating this phenomenon requires 15 mns for 1 minute of the corpus. The following table illustrates the fact that disfluencies are speaker-dependent in terms of quantity and type. Phrase interruption form the majority of disfluencies in both cases. However, word disfluencies appear more frequently in Speaker_1 (20%) than in Speaker_2 (8%) productions. These figures also shows that disfluencies affect lexicalized words as well as grammatical ones.

	Speaker_1	Speaker_1
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

Syntax The categories and chunks counts for the whole corpus are summarized in the following figure :

Category	Count	Group	Count
adverb	15123	GA	3634
adjective	4585	GN	13107
auxiliary	3057	GP	7041
determiner	9427	GR	15040
conjunction	9390	NV	22925
interjection	5068	PV	1323
preposition	8693	Total	63070
pronoun	25199		
noun	13419	Soft Pct	9689
verb	20436	Strong Pct	14459
Total	114397	Total	24148

Transcription and phonemes The following table recaps 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

In the same way, the following table gives some information as for the phonetic and phonologic domains :

Item	Number
IPU	13069
Orthog. tok.	124051
Syllable	
Vowel	139751
Consonant	152573
Semi-vowel	9554

3.3 First Evaluations

Prosodic annotation : Prosodic annotation, mainly that of prosodic phrasing, of at least 1 dialogue was done by 2 experts. The annotators worked separately using Praat, and the resulting annotations were further processed as to inter-transcriber agreement. Inter-transcriber agreement studies were done for 1-hour conversation between two female participants : we analyzed the observed agreement for the annotation of higher prosodic units (IPs). 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. The level of agreement is thus comparable with one, observed in other studies, mainly done within ToBI paradigm : for example, Syrdal and McGory (2001) communicate the inter-transcriber agreement of 74% for boundary indices but in this study, 5 degrees of boundary strength were distinguished ; for prosodic words and intonational phrases, which roughly correspond to APs and IPs in our study, the agreement was higher than 80%. The corresponding kappa values (5 degrees of boundary strength) were of 0.65 for female speaker and 0.62 for male speakers.

Gesture : The Gesture Space is a "shallow disk in front of the speaker" (McNeill, 1992 : 86) where most gestures are performed. It is divided into four regions (center-center, center, periphery and extreme periphery) and eight coordinates (no coordinate, right, left, left-right, upper right, upper left, lower right, lower left, upper, lower, upper left-right, lower left-right). We used McNeill's (1992 : 89) diagram for the coding. The left-right coordinate was useful whenever a gesture was produced with both hands. 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 (Carletta96).

The kappa coefficient measures pairwise agreement among two coders classifying mutually exclusive categories, correcting for hypothetical chance agreement. Usually, the kappa between .68 and .80 refers to substantial agreement, while k under .68 is considered as fair or moderate agreement. However, applying kappa to nonverbal annotation has not much success. Most of gesture and gaze annotation kappa score is quite low.

Three coders have annotated three minutes for GestureSpace including GestureRegion and GestureCoordinates. Annotations of these two dimensions are based on the point of maximum extension of the gesture. When annotating GestureRegion, the gesture that occurred in the axis of the armrests of the chair is judged in periphery. When it is outside, we annotate extreme periphery. The landmark is the position of the wrist.

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.

4 Information representation

4.1 XML encoding

In our case, the process generates an XML schema for each domain such as phonetics and prosody. Data are then represented in different document structures, which is motivated by two important points : (i) this representation provides a high level of modularity for applicative requirements and enables to modify only the structure / schema needed ; (ii) recent works on multistructured documents open a way for dynamically linking and aggregating various documents with different structure (Bruno06). Within such a framework, it is possible to deal with a TFS-like XML representation and to process data as standard XML documents.

Our XML schema, generated from TFS via UML, can be seen as the ”third component" mentioned as a perspective in (**?**). This component, besides a basic encoding of data following AIF, encode all information concerning the organization as well as the constraints on the structures. In the same way as TFS are used as a tree description language in theories such as HPSG, the XML schema generated from our TFS representation also plays the same role with respect to the XML annotation data file. On the one hand, basic data are encoded with AIF, on the other hand, the XML schema encode all higher level information. Both components (basic data + structural constraints) guarantee against information loss that otherwise occurs when translating from one coding format to another (for example from Anvil to Praat).

The general process is represented in the following schema illustrating the fact that XML schemata are added to the AIF data encoding to ensure full data conservation.

4.2 Querying

To ease the multimodal exploitation of the data, our objective is to provide a set of operators dedicated to concurrent querying on hierarchical annotation. Concurrent querying consists in querying annotations belonging to two or more modalities or even in querying the relationships between modalities. For instance, we want to be able to express queries over gestures and intonation contours (what kind of intonational contour does the speaker use when he looks at the listener ?). We also want to be able to query temporal relationships (in terms of anticipation, synchronization or delay) between both gesture strokes and lexical affiliates.

Our proposal is to define these operators as an extension of the recommended query language for XML, XQuery (http :www.x3.org/xquery). We will extend formal models and languages from the XML universe because they are open standards for which stable and efficient implementations exist (in particular for storage and manipulation). Therefore, from the XML encoding shown in section 5, and the temporal alignment of annotated data, it will possible to express queries to find patterns and to navigate in the structure. We also want to enable a user to check predicates on parts of the corpus using classical criteria on values, annotations and existing relationships (temporal or structural ones corresponding to inclusions or overlaps between annotations). First, we shall rely on one of our previous proposal called MSXD (MultiStructured XML Document). It is a XML-compatible model designed to describe and query concurrent hierarchical structures defined over the same textual data [refLSIS] which supports Allen's relations [refAllen]. Besides, we will study a second approach based on description logics [refDL] to represent annotated data. Our objective is twofold : first to compare the expressiveness of this representation to the TSF one, and second to use the OWL standard (http ://www.w3.org/TR/owl-features/) which relies on description logics. The querying will then be done by means of the standard semantic web query language SPARQL (http ://www.w3.org/TR/rdf-sparql-query/ 2008).

5 Conclusion

Multimodal annotation is often reduced to the encoding of gesture, eventually accompanied with another level of linguistic information (e.g. morpho-syntax). We reported in this paper a broad-coverage approach, aiming at encoding all the linguistic domains into a unique framework. We developed for this a set of conventions and tools making it possible to bring together and align all these different pieces of information. The result is the CID (Corpus of Interactional Data), the first large corpus of conversational data bearing rich annotations on all the linguistic domains.

References

Allen J. (1999) Time and time again : The many way to represent time. International Journal of Intelligent Systems, 6(4)

Allwood, J., Cerrato, L., Dybkjaer, L., Jokinen, K., Navarretta, C., Paggio, P. (2005) The MUMIN Multimodal Coding Scheme, NorFA yearbook 2005.

Baader F., D. Calvanese, D. L. McGuinness, D. Nardi, P. F. Patel-Schneider (2003) The Description Logic Handbook : Theory, Implementation, Applications. Cambridge University Press.

Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B., Rauzy, S. (2008) "Le CID - Corpus of Interactional Data - Annotation et Exploitation Multimodale de Parole Conversationnelle", in revue *Traitement Automatique des Langues*, 49 :3.

Bigi, C. Meunier, I. Nesterenko, R. Bertrand 2010. "Syllable Boundaries Automatic Detection in Spontaneous Speech", in *proceedings of LREC 2010.*

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 proceedings of *TALN 2008* (Avignon, France), pp. 290-299.

Blache P., R. Bertrand, and G. Ferré 2009. "Creating and Exploiting Multimodal Annotated Corpora : The ToMA Project". In *Multimodal Corpora : From Models of Natural Interaction to Systems and Applications*, Springer.

Blanche-Benveniste C. & C. Jeanjean (1987) *Le français parlé. Transcription et édition*, Didier Erudition.

Blanche-Benveniste C. 1987. "Syntaxe, choix du lexique et lieux de bafouillage", in *DRLAV* 36-37

Browman C. P. and L. Goldstein. 1989. "Articulatory gestures as phonological units". In *Phonology* 6, 201-252

Brun A., Cerisara C., Fohr D., Illina I., Langlois D., Mella O. & Smaili K. (2004- "Ants : Le systÃÎme de transcription automatique du Loria", Actes des *XXV Journées d'Etudes sur la Parole*, Fès.

E. Bruno, E. Murisasco (2006) Describing and Querying hierarchical structures defined over the same textual data, in Proceedings of the *ACM Symposium on Document Engineering* (DocEng 2006).

Bull, P. (1987) *Posture and Gesture*, Pergamon Press.

Bunt H. 2009. "Multifunctionality and multidimensional dialogue semantics." In *Proceedings of DiaHolmia'09*, SEMDIAL.

Bürki A., C. Gendrot, G. Gravier & al.(2008) "Alignement automatique et analyse phonétique : comparaison de différents systèmes pour l'analyse du schwa", in revue TAL ,49 :3

Carletta, J. (1996) "Assessing agreement on classification tasks : The kappa statistic", in *Computational Linguistics* 22.

Corlett, E. N., Wilson,John R. Manenica. I. (1986) "Influence Parameters and Assessment Methods for Evaluating Body Postures", in *Ergonomics of Working Postures : Models, Methods and Cases* , Proceedings of the First International Occupational Ergonomics Symposium.

Di Cristo & Hirst D. (1996) "Vers une typologie des unites intonatives du français", XXIème JEP, 219-222, 1996, Avignon, France

Di Cristo A. & Di Cristo P. (2001) "Syntaix, une approche métrique-autosegmentale de la prosodie", in revue *Traitement Automatique des Langues*, 42 :1.

Dipper S., M. Goetze and S. Skopeteas (eds.) 2007. *Information Structure in Cross-Linguistic Corpora : Annotation Guidelines*, Working Papers of the SFB 632, 7 :07

FGNet Second Foresight Report (2004) Face and Gesture Recognition Working Group. http ://www.mmk.ei.tum.de/ waf/fgnet-intern/3rd-fgnet-foresight-workshop.pdf

Gendner V. et al. 2003. "PEAS, the first instantiation of a comparative framework for evaluating parsers of French". in *Research Notes of EACL 2003* (Budapest, Hungaria).

Hawkins S. and N. Nguyen 2003. "Effects on word recognition of syllable-onset cues to syllable-coda voicing", in *Papers in Laboratory Phonology VI.* Cambridge Univ. Press.

Hirst, D., Di Cristo, A., Espesser, R. 2000. "Levels of description and levels of representation in the analysis of intonation", in *Prosody : Theory and Experiment*, Kluwer.

Hirst, D.J. (2005) "Form and function in the representation of speech prosody", in K.Hirose, D.J.Hirst & Y.Sagisaka (eds) *Quantitative prosody modeling for natural speech description and generation* (*Speech Communication* 46 :3-4.

Hirst, D.J. (2007) "A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation", in *Proceedings of the XVIth International Conference of Phonetic Sciences*.

Hirst, D. (2007), Plugin Momel-Intsint. Internet : http ://uk.groups.yahoo.com/group/praat-users/files/Daniel_Hirst/plugin_momel-intsint.zip, Boersma, Weenink, 2007.

Jun, S.-A., Fougeron, C. 2002. "Realizations of accentual phrase in French intonation", in *Probus 14*.

Kendon, A. (1980) "Gesticulation and Speech : Two Aspects of the Porcess of Utterance", in M.R. Key (ed.), *The Relationship of Verbal and Nonverbal Communication*, The Hague : Mouton.

Kita, S., Ozyurek, A. (2003) "What does cross-linguistic variation in semantic coordination of speech and gesture reveal ? Evidence for an interface representation of spatial thinking and speaking", in *Journal of Memory and Language*, 48.

Kipp, M. (2004). Gesture Generation by Imitation - From Human Behavior to Computer Character Animation. Boca Raton, Florida, Dissertation.com.

Kipp, M., Neff, M., Albrecht, I. (2007). An annotation scheme for conversational gestures : how to economically capture timing and form. Language Resources and Evaluation, 41(3).

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

McNeill, D. (1992). Hand and Mind. What Gestures Reveal about Thought, Chicago : The University of Chicago Press.

McNeill, D. (2005). Gesture and Thought, Chicago, London : The University of Chicago Press.

Milborrow S., F. Nicolls. (2008). Locating Facial Features with an Extended Active Shape Model. ECCV (4).

Nesterenko I. (2006) "Corpus du parler russe spontané : annotations et observations sur la distribution des frontières prosodiques", in revue TIPA, 25.

Paroubek P. et al. 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 2006* (Genoa, Italy), pp. 314-320.

Pierrehumbert & Beckman (1988) Japanese Tone Structure. Coll. Linguistic Inquiry Monographs, 15. Cambridge, MA, USA : The MIT Press.

Platzer, W., Kahle W. (2004) Color Atlas and Textbook of Human Anatomy, Thieme. Project MuDis. Technische Universitat Munchen. http ://www9.cs.tum.edu/research

Scherer, K.R., Ekman, P. (1982) Handbook of methods in nonverbal behavior research. Cambridge University Press.

Shriberg E. 1994. *Preliminaries to a theory of speech disfluencies*. PhD Thesis, University of California, Berkeley

Wallhoff F., M. Ablassmeier, and G. Rigoll. (2006) "Multimodal Face Detection, Head Orientation and Eye Gaze Tracking", in proceedings of *International Conference on Multisensor Fusion and Integration* (MFI).

White, T. D., Folkens, P. A. (1991) Human Osteology. San Diego : Academic Press, Inc.