An XML Coding Scheme for Multimodal Corpus Annotation

Philippe Blache,[1] Gaëlle Ferré[1] and Stéphane Rauzy[1]

Introduction

Multimodality has become one of today's most crucial challenges both for linguistics and computer science, entailing theoretical issues as well as practical ones (verbal interaction description, human-machine dialogues, virtual reality etc…). Understanding interaction processes is one of the main targets of these sciences, and requires to take into account the whole set of modalities and the way they interact.

From a linguistic standpoint, language and speech analysis are based on studies of distinct research fields, such as phonetics, phonemics, syntax, semantics, pragmatics or gesture studies. Each of them have been investigated in the past either separately or in relation with another field that was considered as closely connected (e.g. syntax and semantics, prosody and syntax, etc.). The perspective adopted by modern linguistics is a considerably broader one: even though each domain reveals a certain degree of autonomy, it cannot be accounted for independently from its interactions with the other domains. Accordingly, the study of the interaction between the fields appears to be as important as the study of each distinct field. This is a pre-requisite for an elaboration of a valid theory of language.

However, as important as the needs in this area might be, high level multimodal resources and adequate methods in order to construct them are scarce and unequally developed. Ongoing projects mainly focus on one modality as a main target, with an alternate modality as an optional complement. Moreover, coding standards in this field remain very partial and do not cover all the needs in terms of multimodal annotation.

One of the first issues we have to face is the definition of a coding scheme providing adequate responses to the needs of the various levels encompassed, from phonetics to pragmatics or syntax. While working in the general context of international coding standards, we plan to create a specific coding standard designed to supply proper responses to the specific needs of multimodal annotation, as available solutions in the area do not seem to be totally satisfactory.

1. The specific needs of multimodal studies

1.1. Different tag sizes and specification types

A multimodal analysis is based on two theoretical groundings: first it is a study based both on the audio and video signals and which underlying principle is that the audio and visual information both play a part in communication. Many studies (Kogure, 2007; Norris and Jones, 2005; Ekman, 1999 among others) have shown that apart from telephone conversations where compensatory means are developped by the

[1] Laboratoire Parole et Langage, Aix-Marseille Universités and CNRS, Aix en Provence, France
e-mail: philippe.blache@lpl.univ-aix.fr, gaelleferre@yahoo.fr, stephane.rauzy@lpl.univ-aix.fr

participants, a study of natural interactions which would be based on speech only (that is the audio recording) would be lacking a certain number of phenomena in the communication process as will be shown below in concrete examples. The second theoretical grounding deriving from the first is that a multimodal analysis whose aim is to contribute to the elaboration of a theory of communication through the study of spontaneous interactions is bound to establish links between different linguistic fields (as many fields indeed as possible) since communication itself can be envisaged from different points of view.

The challenge when trying to establish links between the different linguistic dimensions is that we are faced with heterogeneity considering the size and the specification of the units taken into account in each dimension. Talking about unit size first, one is faced with heterogeneity: for instance, prosodic phenomena may be larger units such as intonational contours or smaller ones like stress occurring on a syllable or tones which are coded as points in the sound signal. Thus, the initial transcription for all prosodic phenomena will need to be a phonemic transcription since a larger transcription would not allow punctual annotations linked to a precise location in the speech signal. Now, the minimal unit size in other linguistic areas such as discourse analysis for instance may be much larger than a prosodic unit. The same is true for the annotation of gesture phenomena, considering as well the fact that they are not linked with the sound signal but with the video signal (that is there may be a gesture during a silent part of the interaction and this gesture may well play a role in communication, see Kogure, 2007) which renders things even more complicated when one wants to establish a link between gesture phenomena and speech in its prosodic dimension.

Now talking about specification, we are faced with yet another challenge in that the labels used in some linguistic fields are under-specified (from a theoretical point of view), whereas in other fields, labels have a rich specification format and may moreover be hierarchically organized. As an example, we annotated part of the corpus in the framework of the enunciation theory which stipulates that spoken French is organized into two constituent types: the theme and rheme. The theme is in turn organized around several constituents, not the rheme, both theme and rheme forming a larger unit, the oral paragraph. One can see in this example that such a segmentation involves some hierarchy but also that some components of the hierarchical structure are poorly specified, especially the rheme part of the oral paragraph. The same remark holds for some gesture phenomena: hand gestures may be decomposed into different phases (as described by Kendon, 1980) such as a preparation phase, during which the hand is placed in the proper configuration for the gesture, the stroke, that is the gesture itself, the hold, an optional phase during which the movement is stopped but the hand remains in the configuration it had during the stroke, and finally a retraction phase during which the hand returns to a rest position. Some links have been established between these different gesture phases and verbal content for instance. Loehr (2004) also found a co-occurrence of the apex of the stroke phase, i.e. the point of maximal extension of the gesture, and some prosodic accent types. These studies contributed much in establishing the validity of the different phases postulated by Kendon. It seems however difficult or even impossible to describe other body movements such as eyebrow raising and frowning in terms of the different phases postulated by Kendon. The same holds for facial expressions, gaze direction and head movements. Kendon argued that these body movements/postures were precisely movements, not gestures, yet, it has been shown in numerous studies (see the large bibliography on backchannelling given in Bertrand et al., forthcoming 2007) that these types of movements play a role in the communication process and if a head nod may be interpreted differently from a vocal backchannel, it is nevertheless perceived as a minimal response from the interactant. Therefore, it makes sense to annotate these body movements, which are under-specified as compared to hand gestures. It may also make sense to include in the annotation physical objects present on the scene of recording and to which participants may refer to in the course of their interaction by means of deictic pointing and referential verbal expression since the physical object suddenly enters the communication process. This phenomenon occured in our corpus as in the following example:

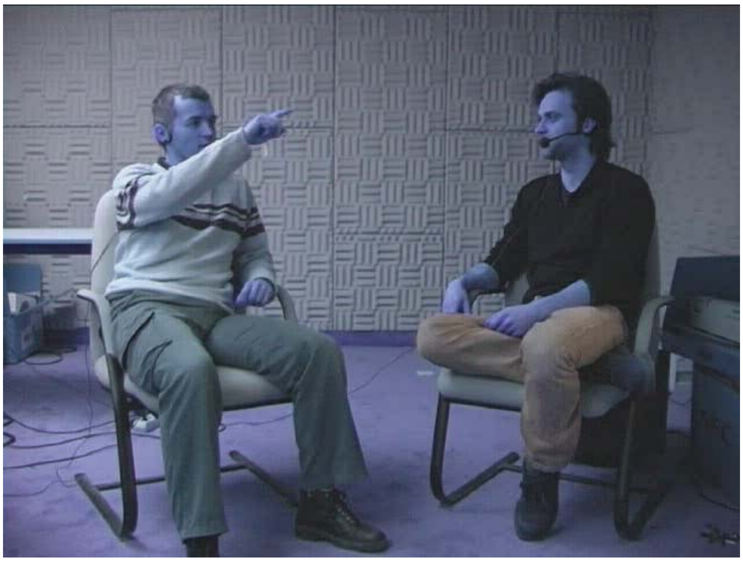


Figure 1: *Introducing a physical object into the communication process by means of pointing.*

The speaker on the left of the picture is describing a hospital room equipped with a small window from which one can watch an operation being done in an adjacent room. He points to the same kind of arrangement in the anechoid room (the technicians may watch the recording setting from a window) saying: “en fait tu as une lucarne un peu comme là tu vois” [in fact you have a small window a bit like here you see]. Doing this, he introduces a part of the anechoid room setting into the description process and this element then has to be taken into account in a multimodal analysis. Yet, to come back to the question of specification, it is obvious that there will be a discrepancy between this kind of video annotation and other types of linguistic information which will have to be addressed at some point of the treatment. One may argue that the more the setting is controlled, the less the participants will add external elements to the communication process itself and consequently the easier it will be to annotate the video recording and establish relationships between for instance movements and prosody or syntax. We do not quite subscribe to this point of view: in parts of our corpus, some speakers used the relative flexibility of the chairs to impress movement on them and use the movement in a communicative way, the point being to determine whether such movements are relevant linguistically and we think they are. What we may conclude from this is that even in a very controlled recording setting such as the one in which we filmed the participants, the environment is also part of the communication process and sometimes has to be taken into account.

To summarize this part, two points have been raised which represent the major challenges of multimodal analyses. Firstly, a multimodal analysis involves the taking into account of the relevant information in various linguistic fields and the units considered in each of these fields may be extremely fine-grained such as phonemes which are needed in the phonetic and prosodic dimensions as opposed to larger units, such as gaze direction, like when the listener gazes at the speaker over a long lapse of speech. It will be difficult to study phenomena taking into account such different unit types (in terms of occurrence in time for instance or in terms of co-occurrence since many smaller units will co-occur with larger units).

The second point raised was the degree of specification of each label used in the annotation. Some labels are richly described like for instance the morphological categories (a verb will be specified in terms of singular/plural, person, tense, aspect, etc., see the example given in section 3.2.) whereas other labels will be underspecified like eyebrow movements which can only be raising or frowning. In addition, eyebrow movements are not linked to any other gesture or to any verbal content and cannot be described as an element of a hierarchical structure, which is not the case of other units like syntactic clauses for example. They don’t have the same degree of reality either, as will be shown in the following subsection.

1.2 The reality of the units used in the different linguistic fields

Another issue raised by the annotation of a multimodal corpus concerns the alignment of the tags with the speech/image signal. Depending on the type of annotation, tags are or are not time-aligned: for instance, a prosodic contour or a gesture is directly linked to the timeline of the video/audio signal. Some references to contextual elements not physically present in the recording will be a bit more tricky to encode. Even more problematic are categorizations in morphology for example: because of assimilation phenomena in spoken language, two morphemes may be pronounced into what sounds like a unique phonetic form. Then, an arbitrary decision will have to be taken as to the segmentation and alignment with the speech signal of the two morphemes. A most typical example of this question is the current pronunciation of “je sais” [I know] [Z@se] in French which is produced as [Zse][2] or even [Se] because of the assimilation after schwa deletion of initial [Z] of “je” and initial [s] of “sais”. In the resulting form [Se] there are clearly two morphemes (first person singular and verb “savoir” at the present tense), yet it sounds arbitrary to attribute [S] to the first morpheme rather than to the second one. The solution we adopted for this kind of problem consists in a symbolic annotation of the two morphemes: the time stamp of the phonetic form is divided into two equal morphological labels. Some annotation tools provide this solution in their structure (such as ELAN developed at the MPI for the annotation of videos). Symbolic annotations are not related to the time line.

This issue leads us to a more general discussion of the annotation type which can be either descriptive: in the example given in the preceding paragraph, a descriptive annotation would describe the assimilation phenomenon; or it can be categorial: the same example would be assigned two morphemes indexed to the unique phonetic form and only the phonetic form is time-aligned. We will discuss later in more detail the utility of such a complex anchorage.

2. The annotation type: descriptive vs. categorial

Before entering the detail of the general coding scheme we used for the annotation of our corpus, we would like to make some more general remarks concerning the issue raised in the preceding section. The issue was about the two different types of

[2] Phonemic annotation in SAMPA for French: The *Speech Assessment Methods Phonetic Alphabet* is a machine-readable phonetic alphabet. http://www.phon.ucl.ac.uk/home/sampa/index.html

annotations which can be made on a corpus and we will start with the description of the labels used for gestures and then show in which way they are congruent with the annotations in the other linguistic fields.

Before starting to work in a multimodal perspective, the corpus had already been annotated in different linguistic fields and especially at the prosodic level with the annotation tool Praat (Boersma and Weenink, 2005). This means that the audio signal had been entirely transcribed in Praat and that prosodic information such as intonational contours, stress, etc. had been annotated and aligned with the signal. Now, when we started to annotate the gestures made by the participants in the interactions, we had to decide which annotation tool would be the most appropriate among a choice of different tools with different functions which we now review in order to explain our choice. This was guided by two requirements:

- the compatibility of the input and output files with other tools since all the annotations are not edited in the same environment;
- the possibility of editing both fine and large labels

The tool that answered these requirements better for the annotation of the video was ANVIL, other tools being ELAN, EXMARALDA and TRANSANA (working in the CLAN environment). Below, we compare ANVIL with the other tools also used by the community working in a multimodal framework and explain our choice addressing first the issue of compatibility, and then the question of label size.

2.1 Compatibility of the video annotation tool with other environments and label size

Concerning the compatibility of the video annotation tool with the other tools used for the annotation of the corpus, we had two initial requirements: the tool used would have to be compatible with Praat which has been used for the initial transcription of the corpus and the phonemic and prosodic annotation (since all the linguistic fields need to be put into relation for a multimodal analysis, it would be a considerable loss of time to have to do the transcription in each tool used); we also needed to be allowed to express hierarchies.

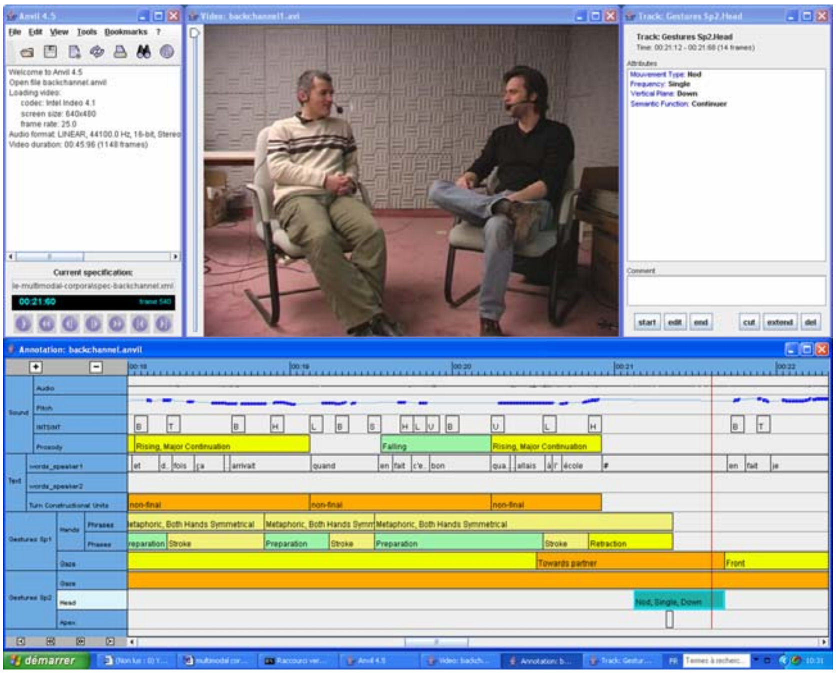
Of the four tools mentioned above, only ELAN doesn't allow import of the annotations made in Praat, which for us ruled this tool out although it presents other interesting features like hierarchical structuring of the annotations and multiple video handling (which is not possible with the other tools). We didn't choose EXMARALDA either because, although it is compatible with Praat, it is not possible to organize the annotations in terms of hierarchies, and we needed this in certain fields, especially syntax. In addition to the compatibility with Praat concerning annotation TextGrids, ANVIL presents another advantage: it is possible to import the waveform, pitch and intensity from Praat which can be useful when relating visual with prosodic phenomena. Lastly, ANVIL's output annotation files are in XML which means that they may be edited with other tools for punctual finer-grained annotations.

TRANSANA is also compatible with Praat annotation TextGrids (though you cannot import the waveform and pitch) and works with the CLAN environment which is used by a certain number of annotators in corpus linguistics. Yet, we didn't choose this tool for reasons of label size: the environment allows a large transcription (for instance in terms of speech turns) but is quite inappropriate for a finer annotation (in phonemes or even morphemes/words). The annotations themselves are bounded by bullets placed at the beginning and end of the corresponding transcription which is itself aligned with the sound signal. So the annotation is presented differently from the other annotation tools which rather adopt an annotation in partition. The absence of partition in Transana renders even more problematic the finding of relationships between tags of different size.

Our choice then fell on ANVIL for all the reasons just mentioned and this has an impact on the coding scheme we elaborated for the annotation of video files as will be explained in the following section.

2.2 ANVIL's environment: a categorial annotation

To better illustrate our point on the type of annotation created in ANVIL (and the coding scheme used), we will compare this tool with ELAN (another tool for the annotation of video phenomena) whose physical environment resembles much that of ANVIL but whose functioning is different.



Welcome to Anvil 4.5
Open file backchannel.anvil
Loading video:
codec: Intel Indeo 4.1
screen size: 640x480
frame rate: 25.0
Audio format: LINEAR, 44100.0 Hz, 16-bit, Stereo
Video duration: 00:45:96 (1148 frames)

Current specification:
le-multimodal-corporalspec-backchannel.xml

00:21:60 frame 540

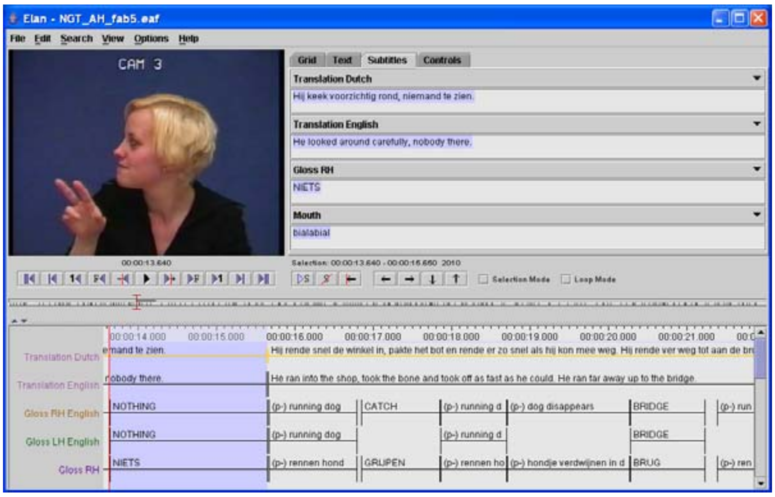
Track: Gestures Sp2.Head
Time: 00:21:12 - 00:21:68 (14 frames)

Attributes
Mouvement Type: Nod
Frequency: Single
Vertical Plane: Down
Semantic Function: Continuer

Comment

Figure 2. ANVIL's working environment.

Both tools (ELAN and ANVIL) are similar in environment aspect: they both show the video to which the annotations are linked (time-aligned), they both have a command window to read the video file at different speeds and both have an annotation window organized in the shape of a partition, each track being dedicated to a certain type of phenomenon (this being decided upon by the annotator). The tracks may be totally independent or organized in groups (for instance a group for syntactic annotations, another for prosodic events and yet another one for gestures). They also may be dependent upon one another: for instance, a syntactic unit like a clause contains a certain number of words which in turn are uttered with certain phonemes, i.e. a syntactic clause is dependent on words which are themselves dependent on phonemes. The annotations in each track appear in the environment as a tag bounded in time with a beginning and an end: playing one tag will play the segment of video where the phenomenon observed starts until the



Elan - NGT_AH_fab5.eaf
File Edit Search View Options Help
CAM 3
Grid Text Subtitles Controls
Selection Mode
Loop Mode

Translation Dutch
Hij keek voorzichtig rond, niemand te zien.

Translation English
He looked around carefully, nobody there.

Gloss RH
NIETS

Mouth
bialabial

00:00:13.640
Selection: 00:00:13.640 - 00:00:15.650 2010

Translation Dutch	emand te zien.	Hij rende snel de winkel in, pakte het bot en rende er zo snel als hij kon mee weg. Hij rende ver weg tot aan de bri				
Translation English	nobody there.	He ran into the shop, took the bone and took off as fast as he could. He ran far away up to the bridge.				
Gloss RH English	NOTHING	(p-) running dog	CATCH	(p-) running d	(p-) dog disappears	BRIDGE
Gloss LH English	NOTHING	(p-) running dog		(p-) running d		BRIDGE
Gloss RH	NIETS	(p-) rennen hond	GRIJPEN	(p-) rennen ho	(p-) hondje verdwijnen in d	BRUG

phenomenon's end. So far, there is no real difference between the two editors. The major difference lies in how tags are created: in ELAN, you enter new tags as you are doing the annotation and they may well be descriptive: for instance, if a speaker is moving in his seat, you may enter a tag with the following gloss "speaker moving in seat". In ANVIL, things are done differently. You cannot enter any new tag or any annotation at all unless you write an XML file which describes in advance all the tracks and tags which will be used during the annotation process.

Figure 2: *Screenshot of ELAN.*[3]

This means that you cannot invent a new tag during this process, although you may add new elements to the specification file itself. There is also the possibility of adding a "token track" in which you may enter any character string but one uses this possibility as less as possible since it is not quite in the spirit of XML files.

For this reason, instead of describing each gesture precisely, we used categories in our coding scheme which is very much an adaptation from the MUMIN coding scheme used by Allwood et al. (2005) who used ANVIL for their annotation of videos and their studies of backchannels. Hand gesture annotations are based on the categories proposed by McNeill (1992) such as iconics, metaphorics, beats, deictics, adaptors. These categories are largely used in the field of gesture studies and have been tested by a high number of researchers. The hand gestures are also decomposed into the phases described by Kendon (1980) already mentioned earlier in this paper. Added to this is a quite complete set of more descriptive values like which hand is performing the gesture, what is the shape of the fingers (based on a simplified description of hand and finger shape for Sign Languages for which these features are highly relevant), what type of movement is performed. The annotation, when completed is precise, although it cannot be as precise as a gloss describing a gesture. If this type of precision is required, one may add a comment to the tag. Yet, this information could hardly be used with automatic queries on the output file.

Movements from other parts of the body are also annotated much in this perspective: for instance as far as gaze is concerned we code gaze direction but also its role in the interaction as regards floor control. Again, the role of gaze in floor control has been largely described since the pionneer work of Kendon (1967). We proceeded in the same way for the description of head and eyebrow movements as well as facial expression. The first step in the annotation process consists in describing the type of movement and then in determining its function in the interaction. This latter step is done by several annotators which guarantees its neutrality. The inter-annotator agreement (on a scale going from 0-disagreement to 4-strong agreement) is also coded.

The drawback of this type of annotation is that there is a certain lack of precision as compared to a descriptive annotation (ELAN type), and the advantage is that only the relevant features of gestures are coded thus allowing a systematic treatment of the data which is necessary if one wants to show the links existing between different annotation types in different linguistic fields.

Such an annotation for gestures is also particularly interesting for us since we proceeded in the same way in other fields. For example in prosody, what is coded is not every micro-variation of the pitch, which are not even perceived by the listener, but rather the general contour of the F0 curve and the relevant prominences. The same

[3] Image from the ELAN homepage at http://www.lat-mpi.eu/tools/elan/elan-screenshot/image_view

applies at the phonetic/phonemic level: we precisely chose a phonemic transcription rather than a phonetic one (with all the intra- and inter-speaker variations in pronunciation) simply because a phonetic transcription would not add anything to a multimodal study, whereas a phonemic transcription is useful for time-alignment. For other fields like morphology, the corpus is finely annotated, yet in a multimodal perspective, we only use the largest categories like noun, verb, adverb (without any mention of person, tense, adverb type, etc.) since the details would probably not be relevant at this stage of our studies.

These were general remarks on the type of annotation we used at the LPL. We are now going to enter the coding scheme used into more detail after a short review of the existing annotation projects.

3. Coding Schemes

3.1 Overview

Several international projects proposed standards for linguistic information encoding of annotation forms. Since the end of the Eighties, the Text Encoding Initiative proposed an exhaustive set of markers (regularly updated) to label all kinds of information being able to enrich a text. Most of the standards suggested respect the directives provided by the TEI; in particular the project CES (Corpus Encoding Standard), or the group EAGLES (Expert Advisory Group on Language Engineering Standards) and its XML evolution called XCES. To be more precise this project provides some encoding specifications for linguistic annotation as well as a data structure for corpus linguistics. The XCES tagging format will be re-used and completed if necessary in our project. This project consists in the developing of a multimodal platform for the processing and annotation of video corpora.

Within the framework of the network of excellence Humaine (2004-2008), automatic and manual annotations of emotions and their form in various modalities are collected from broadcast video news (Martin and al., 2005, 2006). The MUMIN network (MUMIN, 2006) federates research in multimodality in the north countries and proposes an annotation of non-verbal behavior in broadcast news oriented towards the study of speech turns. Bird and Liberman (1999) also present a general specification body for multi-level annotations.

However, in the field of multimodal corpus analysis, there has not yet been any real standardization initiative. Each coding scheme has been decided on individually in each modality. The most used coding schemes are for example FACS for facial expressions (see Ekman and Friesen, 1978), and structural and functional descriptions of hand gestures (see Efron, 1941; McNeill, 1992; Kipp, 2004). These schemes are not computerized and are often adapted solely to the needs of the researchers and the annotation tools they use. They are studied in relation to higher annotation level like communicative functions or emotions which belong to separate annotation schemes (see Pelachaud, 2005). A state of the art of annotation schemes was proposed in (Knudsen and al, 2002).

3.2 Multimodal Coding Scheme

A multimodal coding scheme is necessary to reach a level of genericity allowing the representation of the encoded information independently from the theoretical options chosen. It implies the use of models with a sufficient expressive power. The redundancy of information across several linguistic fields will be possibly accepted.

We found it necessary to use a track-by-track representation, i.e. a representation for which each linguistic annotation level is independent from the others and is annotated following a hierarchical model (as proposed by standard XML tools). This coding scheme is intended to allow a global exploitation of corpora annotated at different levels: each level is independent in the representation, the global view comes from the possibility to interface them for edition and manipulation thanks to the knowledge of the structure. We propose to combine the existing schemes and to extend them so as to obtain an XML coding scheme as complete as possible in all the following domains:

- *Corpus metadata*: we use a TUSNELDA-like coding scheme (see Tusnelda, 2005) in which all the information such as speaker name, sex, region, etc is noted.
- *Morphology and syntax*: we adapted the Maptask coding scheme to the French language in the morphological dimension, completed with syntactic relations and properties.
- *Phonetics and prosody*: some annotations have been inspired by MATE (Carletta and Isard, 1999) and completed. The phonemic representation is coded in SAMPA and we used the INTSINT and MOMEL algorithms for the phonological representation of intonation.
- *Gesture analysis*: we adapted the MUMIN coding scheme (Allwood et al., 2005; Mumin, 2006) by coding separately gestures and discourse tags. In practice, the coding scheme concerning facial expressions and head movements is based on the FACS standards. As for gestures, the coding scheme is derived from existing propositions (Kendon, 2004; Kipp, 2004; McNeill, 2005). Gestures typology is encoded following the scheme proposed in (McNeill, 2005). In this scheme, a gesture may inherit more than one category, e.g. some gestures can share iconic and deictic properties. A gesture lexicon has been compiled from the existing descriptions found in the literature (Kipp, 2004; Krenn and Pirker, 2004) and on the basis of our own experience. In particular, the scheme describes relevant aspects of emotional and individual profiles (handedness) in terms of motion’s quality (Martin, 2006). Gesture expressiveness (Pelachaud, 2005) and perceptual motion’s quality (Wallbott, 1998) is encoded by adapting the LIMSI manual annotation protocol.
- *Pragmatics and discourse analysis*: we use the Maptask (Isard, 2001) and DAMSL coding schemes, extended to other discourse types such as narration, description, etc.

This coding scheme is organized by modality and covers a much broader range of domains than what has been done so far. This organization lies on a system of complex anchorage (Blache, 2003) so as to combine different unit types both in nature and size and thus allow a multimodal processing of the corpus.

It is obviously not possible to present in this article the entire coding scheme. The complete coding scheme is described and can be consulted on the CRDO[4]. We only sketch here some of its aspects. The presentation is given under a TEI-like formalism, describing for each tag its arguments and if necessary its contents. Arguments of the tags correspond to arguments of an XML element. They bear information local to the tag. In some cases, tags correspond to empty elements. If so, the description only contains its list of attributes. Otherwise, tags can correspond to complex elements, with non empty content list. This list corresponds to the constituent set of the tag. They are encoded to their turn into elements. Here is the example of the description of our morphosyntactic encoding:

```
Token::
     attributes: form
     content: Lex*
```

```
Lex::
     attributes: id cat lemma rank prob freq phon ref
     content: Msd

  cat: {Adjective Determiner Noun Pronoun Adverb Preposition Auxiliary Verb
Conjunction           Interjection Ignored Punctuation Particle Filled pause
Unknown}
```

```
Msd::
     attributes: gender number pers mood tense conjug defection verbal_type subcat
                          pronominalization personal reform
```

Morphosyntax is encoded by means of three main tags: `Token`, `Lex` and `Msd`. The first tag has an attribute (`form`) containing the lexical form. It is a complex element, with possible several sub-elements of type `Lex`. Each tag `Lex` correspond to a possible analysis of `form`. It bears several attributes, among them, the category, the lemma, the form frequency, its probability in the context, etc. In our representation, types of attribute values are, when necessary, précised, eventually as a list of possible values, as for the attribute `cat` in the example. Each element `Lex` contains one sub-element, `Msd`, describing precisely its morpho-syntactic characteristics.

The next example illustrates a tag for gesture encoding. It concerns heads movements (based on MUMIN proposal) and shows how to describe them precisely. We will give in the last part of the paper an illustration of the interest of such encoding.

```
Head::
  attributes: movement_type frequency horizontal_plane vertical_plane side_type

  movement_type: {Nod, Jerk, Tilt, Turn , Shake , Waggle , Other}
  frequency: {Single, Repeated }
  horizontal_plane: {Forwards, Backwards, Sideways}
  vertical_plane: {Up, Down}
  side_type: {Left, Right}
```

[4] *Centre de Ressources de Données Orales*. The coding scheme is at the following address: http://crdo.fr/phpwiki/index.php?pagename=CIDcoding. The aim of the CRDO is to host and distribute different types of corpus, together with transcriptions or annotations corresponding to the raw files (either audio or video files). Each resource hosted by the CRDO is thoroughly described and documented and parts of the corpora already are or will soon be shared with the research community.

A coding scheme, in particular those with a generic perspective, are subjects to constant evolution, making it possible to integrate new kinds of information for the description of specific phenomena. This is the case for the two next tags, which we propose to add to the scheme. The former describes backchannels (vocal or gestural; the vocal BC types being given for French), the latter humor information.

```
Simple_vocal_and_gestural_BCs::
  attributes: token frequency type bc_function understanding_value

  token:{ouais, mh, oui, non, eh, ouais, ah, ouais, ah bon, d’accord, voilà, ok,
Other}
  frequency: {Single, Repeated}
  bc_function: {ct udg ack as crt rt}
  understanding_value: {Expected Unexpected}
```

```
Humor_type::
  attributes: Enunciations Enunciator Irony Sarcasm Mockery Deadpan_type Joke Target
                    Face_work

  Enunciations: Number(0,4)
  Enunciator: {Speaker 1, Speaker 2, Both speakers}
  Irony: Boolean
  Sarcasm: Boolean
  Mockery: Boolean
  Deadpan_type: Boolean
  Joke: Boolean
  Target: {Speaker, Hearer, Other person, Situation, Gesture, Object, Language}
  Face_work: {Threatening, Saving, Double Bind}
```

It is easy from these description to generate an XML enconding for specific annotation. We propose in the following example an illustration of the XML result for encoding morphosyntactic information[5]:

```
     <Token orth=“mange”>
         <Lex  id=“625”  cat=“Verb”  lemma=“manger”  rank=“1”  prob=“0.6”  freq=“78636”
phon=“ma~Z”>
             <Msd number=“singular” mood=“indicative”
               tense=“present”
               conjugation_table=“23”
               verbal_type=“main”
               subcat=“transitive_direct”
               pronominalization=“optional”
               personal=“true”/>
         </Lex>
     </Token>
```

We elaborated the coding scheme while working on the CID video corpus (*Corpus of Interactional Data*, a corpus of spoken French in face-to-face interactions, described in Bertrand and al., 2007), yet it can be applied to the annotation of any other video corpus of conversation.

[5] Some definitions necessary to understand the example: tag id= position of token in the sample; tag rank=rank in the phrase; tag prob= probability in context of this type of category; form id= raw spelling; form lemma= lemma from which the token is derived; form freq= usage frequency in the language; form phon= phonemic form of the token; form ref= reference of entry in the lexicon.

4. Examples of multimodal studies

In order to illustrate how we make use of the corpus and the different annotations made, we will briefly describe two studies. Both are studies under development although the first results will be presented at conferences and published in the proceedings. Both studies have also been derived from the same annotation file (i.e. the annotation is large enough to cover several linguistic phenomena). The first work presented will be on backchannels, see Bertrand et al. (forthcoming, 2007), that is minimal answers from the interlocutor and which show the collaborative work in the interaction process and the second study concerns reinforcing gestures (also called intensive gestures), see Ferré et al. (forthcoming, 2007).

4.1 Backchannels

For a study on backchannels, we put into relationship verbal content, morphological categories, prosodic units and gestures within two samples of 15 minutes each, involving four speakers (2 males and 2 females). The corpus was first integrally transcribed in enriched orthography (a transcription that takes into account specific pronunciations). It was then transcribed in phonemes and the phonemes aligned with the speech signal which allows a finer annotation of the prosodic units and contours. These were annotated with Praat (Boersma and Weenink, 2005). Two prosodic units were retained: APs (accentual phrases, the smaller intonational units) and IPs (intonational phrases, higher in the hierarchy). In addition, each unit type was attributed a contour (minor pitch movements for APs and major ones for IPs). We also noted the flat pitch which is not considered as a contour in itself but plays a role in story-telling and is quite frequent in our samples.

From the orthographic transcription, we also annotated in Praat all the simple vocal BCs (leaving aside for the time being complex BCs such as repetitions, reformulations, etc). Each BC was attributed a function in the interaction (acknowledgement, assessment, understanding, etc.). We also noted discourse markers such as connectors (linking words between Turn Constructional Units), punctuators (which are produced at the end of the TCU), phatic markers...

For gestures, we annotated all the speakers' head and eyebrow movements, facial expressions such as smiles, laughters and gaze direction using Anvil (Kipp, 2004). In a second time, we considered the role in the interaction of each movement/gesture which could be a BC, but also a reinforcing gesture, a direct answer to a question, etc. In a third step, we also attributed a function to the gestural BCs (the functions of gestural BCs were the same as the functions of vocal ones).

An example of vocal and gestural BCs:

A:[il était quatre heures]$_{AP}$ [de l'après-midi]$_{IP}$ [on en pouvait plus]$_{AP}$ [d'attendre]$_{IP}$
minor rise *major rise* *minor rise* *major rise*

B: **ah ouais**
head nod
laughter

[Translation*: it was four in the afternoon (oh yeah) we couldn't wait hem any longer*][6]

The hypothesis was that BCs provide information on speaker's discourse elaboration processes (Fox Tree, 1999). In fact BCs mark some important steps in discourse which can be signalled by various cues at different linguistic levels, such as prosodic units, pitch contours, morphological categories, discourse markers or gaze direction. We therefore examined the contexts in which BCs are produced in order to find out the different steps in the discourse elaboration of the speaker.

The results of the study showed that vocal and gestural BCs occur in the same kinds of environment (see Bertrand and al., forthcoming 2007, for details of the differences between the kinds of BCs) but gestural BCs seem to be delayed as compared to vocal ones (they occur some time after the end of the intonational unit). Gestural BCs are also encouraged when the speaker is gazing at the interlocutor.

As far as the morphological context is concerned, gestural BCs occur preferentially after nouns, verbs and adverbs. These categories correspond to words with important semantic functions: predicate, referential objects and predicate modifier. They correspond to categories playing a central role in the argument structure, explaining the fact that specifiers or modifiers are not connected to BCs.

BCs are not favored by accentual or intonational phrases. Yet they occur preferentially after rising or flat contours. By producing a BC after a rising contour, the listener shows that he understands that the speaker has not finished yet. By producing a flat contour the speaker signals an event called "aside" which is defined as a parenthetic element inserted in a story projecting a later end of the story, and this projection is acknowledged by the BC produced by the listener.

Lastly, none of the discourse markers tested encouraged the production of a BC. We deduced from this that the phatic function is rather assumed by gaze, and we showed that when the speaker is gazing at the interlocutor, the latter produces a BC.

Our preliminary results confirm that backchannel signals do not only play a role in the listening and understanding processes but they also play a role in the elaboration of discourse, in marking different steps in the conversation. These steps have to do with the information discourse properties as well as the relationships between the participants (common-ground shared by the participants for instance).

4.2 Reinforcing gestures

During the annotation process of backchannels, which has been presented in the previous subsection we annotated all the head and eyebrow movements, hand gestures and gaze direction of the speakers on part of the corpus. Some of these movements were either identified as backchannels or as phatic movements which also play a role

[6] Transcription conventions are given at the end of the paper.

in the feedback process. However, other gestures/movements had different functions one of these being the reinforcing function which was noted as well. For example, a head nod may be produced by the speaker (in which case it is not a backchannel) and may not have any phatic function: a phatic marker invites a backchannel from the interlocutor. It has been shown that when the speaker is expecting some collaboration from the interlocutor (either in the case of questions or in the case of phatics) he/she gazes at the interlocutor (Bertrand and al., forthcoming 2007). Reinforcing gestures are not produced in contexts of mutual gaze and are not followed by any backchannel from the interlocutor. This means that they may be distinguished from phatics and play a different role in the communication process.

After having noted what was perceived by the annotators as reinforcing gestures, we asked ourselves two questions. The first question was: “What do reinforcing gestures reinforce? The second question was whether reinforcing gestures could be assimilated to prosodic focalization processes which serve as emphasis of some parts of dicourse. The issue at stake was to determine whether these gestures were redundant with other reinforcement processes or whether they played a distinctive role in the communication process. In the case of redundancy, then there is no particular need to speak of the gestures at all since an audio analysis of speech would be sufficient to catch the linguistic phenomenon under study. In case the gestures are not redundant with speech however, it becomes important and even essential to take them into consideration since they add information to the vocal message. In the following example about a school teacher, the speaker makes two reinforcing head gestures and a hand beat:

```
A: elle était SUper stricte elle voulait PAS tu vois elle interdisait que tu
sortes
               head nod      head shake
               hand beat
```

[Translation*: she was super strict she didn’t want (...) you see she forbade us to go out (during class*)]

In this example, both head movements have been interpreted as reinforcing gestures by the annotators. Yet, in such an example, one may wonder whether the head movements are in some way correlated with the two focalization accents or not, and whether they are correlated with the co-occurring hand movement, here a beat.

The results of the statistical analysis showed that first of all, reinforcing gestures are not produced together with focalization accents. Consequently, the two emphasizing processes are not redundant: they may appear conjointly or separately and they play different roles. The role played by reinforcing gestures is very much determined by the type of element they accompany.

Reinforcing gestures significantly co-occur with adverbs (either lexical degree adverbs or the negation particle) at the morphological level and with connectors (linking conjunctions or interjections between Turn Constructional Units) at the level of discourse analysis. As regards gestures, they significantly co-occur with metaphorics in McNeill’s gesture typology (1992). Metaphorics are gestures which illustrate abstract ideas. This body of evidence shows that reinforcing gestures’ major role concerns discourse planning and organizing rather than emphasizing some discourse element for the interactant. Even if they are seen by the interactant who necessarily takes them into account, they are not made to be acknowledged by the interlocutor but rather play a syntactic role.

What can be concluded from this example as well as from the preceding example on backchannels is that communication is a complex process and that a study of language based on only one linguistic field is not sufficient to account for this complexity. Verbal content, prosody and gestures are not redundant and a multimodal study adds to the description of what's going on in an interaction.

Conclusion and perspectives

We have provided a coding scheme for the annotation of a multimodal corpus of face-to-face interactions. This coding scheme is extremely precise in every field of research, yet it may be used with other corpora: since the fields (such as prosody, syntax, gesture studies) are encoded independently from each other, part of the coding scheme may be used for the description of audio files only for instance, or in one linguistic area. It has been developed for the description of French and English but may easily be adapted to other languages when need be. Some parts of the scheme (such as the annotation of gestures) may even be used for any other language. The coding scheme is available online on a website designed for opensource corpora and tools: the CRDO (*Centre de Ressources de Données Orales*, http://crdo.up.univ-aix.fr/roa.php?langue=fr).

The scheme would need some adaptations for other types of interactions: it has especially been conceived for the description of face-to-face interactions and would need some additional tracks and labels for the description of actions involving objects and occurring in other types of setting.

What is particularly interesting in this coding scheme however is its XML structure and the complex anchorage of labels. Both properties render possible the exportation of the annotations into other annotation tools. This has appeared as a major issue at the last ISGS conference[7], where a whole panel session[8] was devoted to the exchange of different annotation files between tools. Tools developers used annotation graphs to recover the information in the XML files and export the annotations into the format of their own tool. This is actually what we have been doing with our annotations since they were created with different tools adapted to the needs of every linguistic field and then related to one another using ANVIL. This approach is particularly interesting for multimodal analyses since a high degree of precision may be reached in each linguistic area, without any loss of information, and the data may then be put into relation with other types of units and labels.

Now we have settled on a coding standard, our next step will be the adaptation of existing analysis tools, and more specifically the adaptation of the tools developed by our research partner teams (signal editors, POS taggers, parsers, etc.) to output productions matching those standards. We plan to use or develop data manipulating and processing tools matching our specific needs. There again, we will base our work on existing tools and adapt them to our needs.

[7] Conference of the International Society for Gesture Studies: *Integrating Gestures*, Evanston, IL., 18-21 June 2007.

[8] Panel animated by D. Loehr , S. Duncan and T. Rose, *Annotation interchange among multimodal annotation tools*, ISGS, 2007.

Transcription conventions

CAPITALS	focalization accent
Underline	Part of speech during which a gesture is produced
(...)	Silent pause
A/B	Speakers
IP	Intonational phrase
AP	Accentual phrase

References

Allwood J., L. Cerrato, L. Dybkjaer, and al. (2005). The MUMIN Multimodal Coding Scheme, NorFA yearbook 2005, http://www.ling.gu.se/~jens/publications/B%20files/B70.pdf

Bertrand R.; Blache P.; Espesser R.; Ferré G.; Meunier C.; Priego-Valverde B.; Rauzy S. (2007). Le CID - Corpus of Interactional Data -: protocoles, conventions, annotations. Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence (TIPA), vol. 25: 25–55.

Bertrand R., Ferré G., Blache P., Espesser R. and Rauzy S. (Forthcoming, 2007). Backchannels revisited from a multimodal perspective. In Proceedings of the Auditory-Visual Speech Processing Conference. Hilvarenbeek, The Netherlands, 31 August-3 September.

Bird S. and Liberman M. (1999). A Formal Framework For Linguistic Annotation. Rapport Interne Ms-Cis-99-01, Department Of Computer And Information Science, University Of Pennsylvania.

Blache P. (2003), Meta-level constraints for linguistic domain interaction. In proceedings of *International Workshop on Parsing Technologies (IWPT-03).*

Boersma P. and D. Weenink (2005). Praat: doing phonetics by computer (release 4.3.14), http://www.praat.org/

Carletta J. and Isard A. (1999), The MATE Annotation Workbench: User Requirements, in Proceedings of the ACL Workshop: Towards Standards and Tools for Discourse Tagging, pages 11–17, University of Maryland, June 1999.

Efron D. (1941). *Gesture And Environment*. New York, King's Crown Press.

Ekman P. (1999). Emotional and conversational nonverbal signals. In L.S. Messing and R. Campbell, *Gesture, Speech and Sign*, New York, Oxford University Press.

Ekman P. and Friesen W. V. (1978). Manual For The Facial Action Coding System. Palo Alto, Ca, Consulting Psychology Press.

Ferré G., Bertrand R., Blache P., Espesser R. and Rauzy S. (Forthcoming, 2007). Intensive gestures in French and their multimodal correlates. In proceedings of Interspeech, Antwerp, Belgium, 27–31 August.

Fox Tree J.E. (1999). Listening in on Monologues and Dialogues. *Discourse Processes* 27(1): 35–53.

Humaine (2004-2008). Network Of Excellence Humaine (Human-Machine Interaction Network On Emotions). Http://Emotion-Research.Net/

Isard A. (2001). An Xml Architecture For The Hcrc Map Task Corpus. In P. Kuehnlein, H. Rieser, H. Zeevat, Eds, Bi-Dialog 2001.

Kendon A. (1967). Some functions of gaze-direction in social interaction, *Acta Psychologica*, Vol. 26, pp. 22–63.

Kendon A. (1980). Gesticulation and Speech: Two Aspects of the Process of Utterance. In M. Ritchie Key (Ed.), The Relationship of Verbal and Nonverbal Communication, The Hague: Mouton, pp. 207–227.

Kendon A. (2004). *Gesture : Visible Action As Utterance*. Cambridge: CUP.

Kipp M. (2004). Gesture Generation By Imitation. From Human Behavior To Computer Character Animation. Florida, Boca Raton, Dissertation.Com. 1581122551. Http://Www.Dfki.De/~Kipp/Dissertation.Html

Knudsen M. W., Martin J.-C., Dybkjær L., Berman S., Bernsen N. O., Choukri K., Heid U., Kita S., Mapelli V., Pelachaud C., Poggi I., Van Elswijk G. and Wittenburg P. (2002). Isle Natural Interactivity And Multimodality Working Group Deliverable D8.1. Rapport Interne, Isle Project.

Kogure M. (2007). Nodding and smiling in silence during the loop sequence of backchannels in Japanese conversation, Journal of Pragmatics 39, pp. 1275–89.

Krenn B. and Pirker H. (2004). Defining The Gesticon: Language And Gesture Coordination For Interacting Embodied Agents. Aisb-2004 Symposium On Language, Speech And Gesture For Expressive Characters, University Of Leeds, UK.

Martin J.-C. (2006). Multimodal Human-Computer Interfaces And Individual Differences. Annotation, Perception, Representation And Generation Of Situated Multimodal Behaviors. Habilitation A Diriger Des Recherches En Informatique. Université Paris XI.

Martin J.-C., Abrilian S. and Devillers L. (2005). Annotating Multimodal Behaviors Occurring During Non Basic Emotions. 1st International Conference On Affective Computing and Intelligent Interaction (Acii'2005), Beijing, China, October 22–24 Spinger-Verlag Berlin. 550–557.

Martin J.-C., Caridakis G., Devillers L., Karpouzis K. and Abrilian S. (2006). Manual Annotation And Automatic Image Processing Of Multimodal Emotional Behaviours: Validating The Annotation Of Tv Interviews. Language Ressources And Evaluation Conference (Lrec'2006), Genoa, Italy, 24–27may

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

McNeill D. (2005) *Gesture and Thought*. Chicago: University of Chicago Press.

Mumin (2006). A Nordic Network For Multimodal Interfaces. http://Www.Cst.Dk/Mumin/

Norris S. and Jones R. H. (2005). *Discourse in Action. Introducing Mediated Discourse Analysis*. New York, Routledge.

Pelachaud C. (2005). Multimodal Expressive Embodied Conversational Agent. Acm Multimedia, Brave New Topics Session, Singapore, Acm.

Tusnelda (2005). Tübingen collection of reusable, empirical, linguistic data structures. http://www.sfb441.uni-tuebingen.de/tusnelda-engl.html

Wallbott H. G. (1998). "Bodily Expression Of Emotion." European Journal Of Social Psychology 28: 879–96. Http://Www3.Interscience.Wiley.Com/Cgi-Bin/Abstract/1863/Abstract