The OTIM formal annotation model: a preliminary step before annotation scheme

P. Blache, R. Bertrand, M. Guardiola, M.-L. Guénot, C. Meunier, I. Nesterenko, B. Pallaud, L. Prévot, B. Priego-Valverde, S. Rauzy

LPL, CNRS & Université de Provence
5 Avenue Pasteur, 13604 Aix en Provence - France
FirstName.LastName@lpl-aix.fr

Abstract

Large annotation projects, typically those addressing the question of multimodal annotation in which many different kinds of information have to be encoded, have to elaborate precise and high level annotation schemes. Doing this requires first to define the structure of the information: the different objects and their organization. This stage has to be as much independent as possible from the coding language constraints. This is the reason why we propose a preliminary formal annotation model, represented with *typed feature structures*. This representation requires a precise definition of the different objects, their properties (or features) and their relations, represented in terms of type hierarchies. This approach has been used to specify the annotation scheme of a large multimodal annotation project (OTIM) and experimented in the annotation of a multimodal corpus (CID, *Corpus of Interactional Data*). This project aims at collecting, annotating and exploiting a dialogue video corpus in a multimodal perspective (including speech and gesture modalities). The corpus itself, is made of 8 hours of dialogues, fully transcribed and richly annotated (phonetics, syntax, pragmatics, gestures, etc.).

1. Introduction

Linguistic annotation requires, especially when annotating many different domains, a very precise description of the information to be annotated before doing any encoding. Such a preliminary step is of great importance for many reasons. First, knowledge representation has to be as much independent as possible from the coding language. In other words, it is necessary to define first linguistic information and its structure before entering into the specification of a coding scheme and a fortiori before doing any encoding.

We propose in this paper an approach relying on such preliminary formal specification by means of *typed feature structures*. This approach has been used to specify the annotation scheme of a large multimodal annotation project (OTIM[1], see (Blache, 2009)) and experimented in the annotation of a multimodal corpus (*Corpus of Interactional Data*, also called CID (Bertrand, 2008). This project aims at collecting, annotating and exploiting a dialogue video corpus in a multimodal perspective (taking into account the needs of analysis at every linguistic level, from phonetics and prosody to syntax or discourse analysis). The corpus itself, is made of 8 hours of dialogues, fully transcribed and partly annotated (phonetics, syntax, pragmatics, gestures, etc.). Because of the specificity of the data (spontaneous dialogue), we chose an enriched orthographic transcription. Our corpus was segmented in IPU (*Inter Pausal Units*). Inside these temporally aligned units, we use manual orthographic transcription, with more precisions about some particular pronunciations, and about some phenomena we wanted to study. This corpus is under permanent evolution, adding new information (e.g. specific constructions such as detachments, specific phenomena as disfluences, etc.), many corrections and new annotation have been done since the beginning of the project. In particular, the orthographic transcription has been done by three experts (one for the first transcription, the others for the correction). A new version of the corpus is now stable as for transcription, prosodic segmentation and phonetic alignment (in particular thanks to a precise annotation of particular pronunciations and laughs). Part of it is freely available through the CRDO[2].

We propose in this paper a concrete presentation of our formal model illustrating the interest of our approach for some of the domains encoded in the OTIM project: phonetics, prosody, discourse and disfluencies. This presentation underlines the interests of a homogeneous formal representation in the perspective of the development of a large annotation scheme, covering all domains and modalities. Such a scheme is an imperative pre-requisite before studying precisely multimodality.

2. The formal model

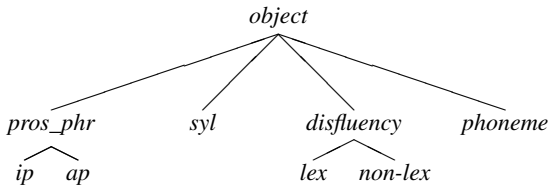
Many different annotation schemes exists, almost one per project. Only few are directly interoperable due to lack of associated semantics. In most of the cases, the annotation scheme and the annotation guidelines are the same document, as for (Dipper, 2007). We think necessary to clearly distinguish the formal description of the information from the way it is encoded. We propose for this to use *typed feature structures* as description language for this formal model.

This formal model basically proposes two kinds of information: the description of the objects (by means of feature structures) and the relations between the different objects (in terms of type and constituent hierarchies). The most general type, *object* (types are noted in italics), dominates many different subtypes. For example, prosodic phrases, syllables, phonemes, etc. are subtypes of object. To their

[1] *OTIM: Outils pour le Traitement de l'Information Multimodale* / Tools for multimodal annotation. See http://aune.lpl.univ-aix.fr/ otim. OTIM is an ANR project (ANR BLAN08-2-349062)

[2] CRDO (Resource centre for oral description), http://crdo.up.univ-aix.fr/

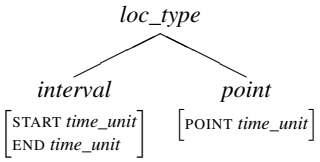
turn, each subtype can also dominate sub-subtypes, etc. The following tree represent a part of the type hierarchy used in our formal model:



Each object, whatever its level, is associated to a set of features specific to its level (also called *appropriate* features). For example, each object (which basically corresponds to any type of annotation, has an index and a location specification. These two information are represented by the following feature structure:

$$_{object}\begin{bmatrix}\text{INDEX } integer\\ \text{LOCATION } loc\_type\end{bmatrix}$$

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



It is important to distinguish type hierarchy from constituency hierarchy. It is clear for example that a *word fragment* is a kind of *lexicalized disfluency*, the difference between them being the level of precision of the object, both of them belonging to type hierarchy rooted by *disfluency*). It is also clear that a *phoneme* constitute *syllables*, but a phoneme is not a specific type of syllable. In this case, a phoneme is a constituent of a syllable. More generally, type hierarchy can be represented as a relation *is-a*, where constituent hierarchy corresponds to a relation *belongs-to*. A constituent is then an object with the particularity that it has to be aligned with an upper-level one. Concretely, when using Anvil for example, an object and its constituents will be represented respectively as primary and secondary tracks.

A subset of the constituent hierarchy (used in this paper) is presented by the following rules (note that the grammar is not complete in the sense that not all non terminals correspond to a left-hand side of a rule):

IP ::= AP^*
AP ::= SYL^+
SYL ::= $CONST\_SYL^+$
$CONST\_SYL ::= PHON^+$
DISF ::= REPRANDUM BREAK REPRANS

We will present in the remaining of the paper different parts of the formal model, corresponding to different linguistic domains or phenomena. In each case, we propose a presentation in terms of types feature structures and detail the information they make it possible to encode.

3. Phonetic description

Phonetic annotation has been done automatically. The phonetizer takes as input the orthographic transcription enriched with specific pronunciations. The output is the list of phonetized tokens. From this list, an automatic aligner localizes each phoneme in the signal. Each phoneme (SAMPA unit) can be then automatically associated to a set of characteristics indicating their category (vowels, consonants), their type (plosives, fricatives, nasals, etc.) as well as articulatory gestures (lip aperture, tongue tip constriction location, velum, etc.) and their role (epenthetic, liaison).
The following figure presents the complete phoneme feature structure:

$$_{phon}\begin{bmatrix}\text{SAMPA\_LABEL } sampa\_unit\\ \text{CAT}\left\{vowel,\ consonant\right\}\\ \text{TYPE}\left\{occlusive,\ fricative,\ nasal,\ etc.\right\}\\ \text{ARTIC\_GEST}\begin{bmatrix}\text{LIP}\begin{bmatrix}\text{PROTUSION } string\\ \text{APERTURE } aperture\end{bmatrix}\\ \text{TONGUE}\begin{bmatrix}\text{TIP}\begin{bmatrix}\text{LOCATION } string\\ \text{DEGREE } string\end{bmatrix}\\ \text{BODY}\begin{bmatrix}\text{LOCATION } string\\ \text{DEGREE } string\end{bmatrix}\end{bmatrix}\\ \text{VELUM } aperture\\ \text{GLOTTIS } aperture\end{bmatrix}\\ \text{ROLE}\begin{bmatrix}\text{EPENTHETIC } boolean\\ \text{LIAISON } boolean\end{bmatrix}\end{bmatrix}$$

Figure 1: Phoneme object

The CAT and TYPE features classifies phoneme into families. Articulation gestures (feature ARTIC_GEST) describe the state and position of the different articulators for each phonetic unit. This description is based on the *Articulatory Phonology Theory* (Browman, 1989) in which gestures are used to characterized real articulation within speech production. In the phonetic annotation of the CID corpus, these gestures are used as articulatory features associated to phonemes. Obviously, this is not a description of the real position of articulators during speech production, but the canonical articulatory target associated to each phoneme. We use gesture labels rather than binary acoustic or phonetic features because we are interested in change and stability of articulator positions in long time-domain. Furthermore, acoustic/phonetic features would be heavier to analyze (two or three acoustic/phonetic features are sometimes needed to describe one gesture).
Examples of features associated to phonemes in the French word *épi* (the label `none` is used when the feature is not relevant for the characterization of the phonetic unit):

$$\begin{bmatrix}\text{SAMPA\_LABEL } /\texttt{e}/ \\ \text{ARTIC\_GEST} \begin{bmatrix}\text{TONGUE}\begin{bmatrix}\text{BODY}\begin{bmatrix}\text{LOCATION } \textit{palatal}\\ \text{DEGREE } \textit{mid}\end{bmatrix}\end{bmatrix} \\ \text{VELUM } \textit{closed} \\ \text{GLOTTIS } \textit{closed}\end{bmatrix}\end{bmatrix}$$

$$\begin{bmatrix}\text{SAMPA\_LABEL } /\texttt{p}/ \\ \text{ARTIC\_GEST} \begin{bmatrix}\text{LIP}\begin{bmatrix}\text{APERTURE } \textit{closed}\end{bmatrix} \\ \text{VELUM } \textit{closed} \\ \text{GLOTTIS } \textit{wide}\end{bmatrix}\end{bmatrix}$$

$$\begin{bmatrix}\text{SAMPA\_LABEL } /\texttt{i}/ \\ \text{ARTIC\_GEST} \begin{bmatrix}\text{TONGUE}\begin{bmatrix}\text{BODY}\begin{bmatrix}\text{LOCATION } \textit{palatal}\\ \text{DEGREE } \textit{narrow}\end{bmatrix}\end{bmatrix} \\ \text{VELUM } \textit{closed} \\ \text{GLOTTIS } \textit{closed}\end{bmatrix}\end{bmatrix}$$

Consequently the sequence of gestures for the production of /epi/ is, for each feature:

LA :	none, closed, none
TBCL :	palatal, none, palatal
TBCD :	mid, none, narrow
VEL :	closed, closed, closed
GLO :	wide, closed, closed

Our aim is to provide statistical inventory of gestures which are potentially present in conversational speech. Furthermore, this description provides time-domain information of articulatory gestures within the speech flow. For instance, phonemes are characterized by several gestures, but gestures may sometimes exceed the limits of a single phoneme (in the French sequence "*un nom*" (transl. a name, pronounced /9ñõ/) the velum position is wide open during the production of the three phonemes). An analysis relying on statistical and sequential informations of the articulatory characteristics have never been conducted before on spontaneous speech.

4. Prosody

One new annotation of the CID is syllables, whose boundaries were automatically detected thanks to a rule-based system we have developed (Bigi, 2010). The notion of syllable is of deep importance because of its role in phonology or phonetics as well as in the description of phonotactics constraints or the analysis of synchronization phenomena between tonal events and segmental strings. In terms of annotation, each syllable bears structural and metrical properties. The following structures describes the different features for syllable description:

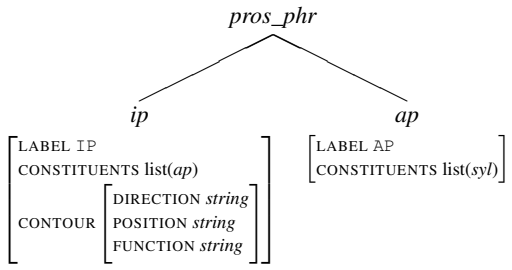
$$_{syl}\begin{bmatrix}\text{CONSTITUENTS list}(\textit{const\_syl}) \\ \text{STRUCT } \textit{syl\_struct} \\ \text{POSITION} \begin{bmatrix}\text{RANK} \left\{ \textit{integer} \right\} \\ \text{SYL\_NUMBER} \left\{ \textit{integer} \right\}\end{bmatrix} \\ \text{ACCENTUABLE } \textit{boolean} \\ \text{PROMINENCE } \textit{boolean}\end{bmatrix}$$

Syllable constituents are classically segmented in three parts: onset, nucleus and coda. Such information is interesting for many reasons, for example because their acoustical properties seems to be different and can have impact on duration (Hawkins, 2003).A second feature describes the syllable structure (V, CV, CCV, etc.), which distribution is important in spontaneous speech description. Syllabic position in polysyllabic words is also specified for duration and tonal alignment studies. As for syllable metrical status, we distinguish between accentuability (a possibility depending on the position) and prominence (a syllable is perceived as prominent if more salient than others).

- *Example*: The following structure describes the syllable /ful/ of the word /meaningful/:

$$_{syl}\begin{bmatrix}\text{LABEL} & \texttt{syl} \\ \text{INDEX} & 42 \\ \text{LOCATION} & \begin{bmatrix}\text{START } 195.12 \\ \text{END } 204.21\end{bmatrix} \\ \text{CONSTITUENTS} & \left\{ \begin{bmatrix}\text{CONST\_TYPE onset} \\ \text{PHON } /\texttt{f}/\end{bmatrix}, \begin{bmatrix}\text{CONST\_TYPE nucleus} \\ \text{PHON } /\texttt{u}/\end{bmatrix}, \begin{bmatrix}\text{CONST\_TYPE coda} \\ \text{PHON } /\texttt{l}/\end{bmatrix} \right\} \\ \text{STRUCT} & \text{CVC} \\ \text{POSITION} & 3/3 \\ \text{ACCENTUABLE} & \text{false} \\ \text{PROMINENCE} & \text{false}\end{bmatrix}$$

At the prosodic phrasing level, we propose a hierarchical organization: following prosodic models for French (Jun & Fougeron, 2002) we make a distinction between *accentual phrases* (minor prosodic units) and *intonational phrases* (major prosodic units). Besides the consensus in French prosodic studies concerning this distinction, there is however a lack of empirical evidence concerning a potential third unit called *intermediate phrase* (Jun & Fougeron 2002, Di Cristo & Hirst 1996). More investigations are still necessary to show its existence. It is in particular of great importance to compare different types of data, such as controlled versus conversational speech (the CID data).
Finally, we propose an extra unit in our prosodic annotation, annotated as "?" (Nesterenko et al. 2010). This unit is used to cope with the specificity of conversational speech, particularly, with the presence of interruptions, false starts and abandoned stretches of spoken material.
The following figure recaps the prosodical type hierarchy:



Intonational phrase description bears an associated melodic contour. Its annotation both contains functional (i.e. conclusive/non conclusive tune) and formal aspects (tune direction and alignment). At the tonal level, f0 curve is semi-automatically annotated with INTSINT coding scheme (Hirst et al. 2000) which encode turning points of f0 curve. To work on such a large database enables to acquire the probabilistic information on relationship between contours and sequences of INTSINT labels.

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

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

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

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

5. Discourse

The first version of our discourse annotation scheme was starting from HCRC dialogue annotation scheme, used for the Edimburg MAPTASK. However, we needed a broader coverage of the communicative functions to account for the phenomena found in truly conversational data (as opposed to task-oriented). We decided then to build a new scheme, relying on the multidimensional functional frameworks such as DIT++ (Bunt, 2009) and compatible with the guidelines defined by the Semantic Annotation Framework (Dialogue Act) working group of ISO TC37/4. This body of work has taken advantage of the experience of several dialogue annotation project and their related schema (in particular DAMSL, SWBD-DAMSL (Jurafsky, 1997)) as well as a thorough empirical and theoretical concerning the multifonctionality and multidimensionality of communicative behavior (Bunt, 2009; Petukhova, 2009). Only few additions have been made to this scheme, in order to allow for the study of discourse relations, humor and reported speech. These new aspects are due to specific discourse interests as well as to the necessity to deal with some new phenomena (in particular, humor and reported speech) that are less represented in task-oriented dialogues.

$$_{du}\begin{bmatrix}\text{INDEX } \textit{integer}\\ \text{LOCATION } \textit{location\_type}\\ \text{PRODUCER}\begin{bmatrix}\text{ROLE}\left\{\textit{hearer, speaker, mixed}\right\}\\ \text{IDENTITY } \textit{string}\end{bmatrix}\\ \text{FORM}\left\{\text{CLAUSE, NSU, DISF, NONVERB}\right\}\\ \text{CONTENT } \textit{logical\_form}\\ \text{FUNCTIONS } \textit{list}(\begin{bmatrix}\text{FUNCTION } \textit{cf}\\ \text{DIRECT } \textit{boolean}\\ \text{PARTIAL } \textit{boolean}\\ \text{VOICE } \textit{voice\_type}\end{bmatrix})\\ \text{TARGET } \textit{du}\end{bmatrix}$$

This discourse and interaction layer is grounded on discourse units coming from our syntactic layer. These units features information about their producer, have a form, 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 multifonctionnality that cannot be avoided (Bunt, 2009).

Form includes clauses, *non-sentential utterances* (nsu) as described in (Fernandez, 2002), *disfluency* that also constitute another annotation level of the project but that need to be minimally characterized at the discourse level as well, and *non-verbal* that covers laughter but that could be extended to gesture if gesture analysis at the discourse level was performed.

Function, although the model allows for a very fine grained segmentation that can avoid all the linear multifunctionality,[3]) they cannot completely avoid multifunctionality (specially at the level of the Core communicative functions). Therefore, we allowed several functions for a unit. With a fine-grained segmentation the case should not occur frequently however.

Compared to standard dialogue act annotation frameworks, three main additions are proposed: *rhetorical function*, *reported speech* and *humor*. The additions are due mostly to the nature of conversational data. Monologic sequences that can only be described in detail by taking into account the rhetorical relations that holds between each element of the narrative. Moreover, the storytelling nature of the data results in a considerable amount of reported speech. This phenomena is quite difficult to deal with in standard dialogue frameworks that have been usually designed for handling task-oriented dialogues. Our rhetorical layer is an adaptation of an existing schema developed for monologic written data in the context of the ANNODIS project (Pery, 2009). Finally, humor cannot be overlooked since the same sequence or discourse unit may have very different functions in humorous and non-humorous sequence.

The features elaborating the function include the function properties (ISO-DA2009, 2009) describing orthogonal values that are not properly speaking functions and can apply be applied to several discourse dimensions. We propose to include reported speech properties at this level of description.

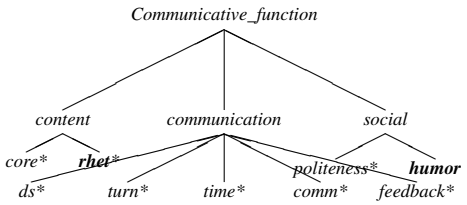
The communicative acts in a conversation can be directed toward following (*forward-looking*) or previous (*backward-looking*) elements or play simply plays in a role in the conversation that cannot be related to linguistic context (e.g *warnings*). Therefore, we took a middle position between framework requiring all elements to be somehow related together and those that gave up providing a relational analysis. For each of our discourse units, the annotator can decide whether or not a target (either forward or backward looking) should be specified.

We also added the notion of sequences of discourse units, in order to deal with thematic grouping, dialogue games,

[3]The multifunctionality due to a too coarse granularity of elementary units.

etc. Despite their significance, the lack of agreement on their definition in the literature renders their annotation difficult. They constitute however a useful level of analysis for providing a description of higher level functions such as humor. For example, we have analysed humorous sequences by annotating both utterances produced by the speaker and the hearer's reactions to reveal the co-construction activity by the two participants.

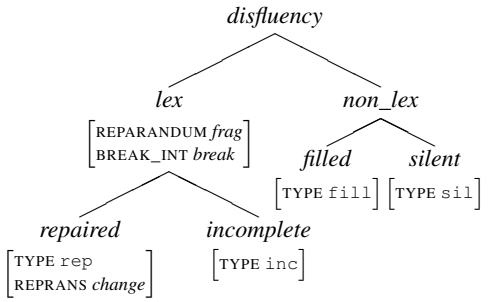
The communicative functions are organized in a taxonomy as illustrated in the next figure. In this figure, bold font signals our proposal while stars (*) signals that the existing taxonomy is further developed. The rest is inspired by (Bunt, 2009; ISO-DA2009, 2009). The taxonomy of *communicative functions* is fairly elaborated and is therefore presented as several smaller taxonomies that are taking place under the upper part introduced below.



core	question, inform, promise...
rhet	Rhetorical relation: narration, explanation, contrast,...
ds	Discourse Structuring : closure, opening, shift
turn	Turn management: grab, yield, ...
comm	Com. management: completion, self-repair ...
fdback	Feedback: acknowledgment, understanding,...

6. Disfluencies

Disfluences are ruptures in the syntagmatic flow. We define and encode them as follows:



Disfluencies are organized around an interruption point (the break), which can occur almost anywhere in the production. They can involve lexical material or not. The latter are part of the prosodic domain (it consists typically in lenghtenings, silent and filled pauses: TYPE <*filled, silent*>); the former share properties from different linguistic domains and are characterized by a word or a phrase truncation, that can either be completed (TYPE *repaired*) or left incomplete (TYPE *incomplete*) after the interruption.

In a lexicalized disfluency, some more information must be precised. We separate linguistic material preceding the interruption point (the REPARANDUM, according to (Shriberg, 1994) typology) and those following it, which is an accumulation over the reparandum's paradigm (Blanche-Benveniste, 1987). In the latter, we distinguish between the content of the the final utterance of the disfluency (REPARANS) and the elements that can take place between the interruption point and the reparans (BREAK_INTERVAL). While the reparandum is mandatory in these constructions, the break interval is optional, and the reparans is forbidden in incomplete disfluencies.

- In the reparandum, we can indicate the nature of the interrupted unit (*word* or *phrase*), and the type of the truncated word (*lexical* or *grammatical*);
- In the break interval, we can indicate a list of the filling elements that appear, among which: silent or filled pause, discursive connector, truncation repetition, parenthetic statement;
- In the reparans, we can indicate the position of the repair (no restart, word restart, determiner restart, phrase restart or other), and its FUNCTIONING (simple continuation of the item, repair without change, continuing through repeating, repair with change in the truncated word, or repair with multiple changes).

7. Conclusion

Large annotation projects, typically those addressing the question of multimodal annotation in which many different kinds of information have to be encoded, have to elaborate precise and high level annotation schemes. Doing this requires first to define the structure of the information: the different objects and their organization. It is not realistic to do this directly into a markup language, necessarily close and even dependent from the data. We think preferable, as presented in this paper, an approach proposing a preliminary definition of knowledge representation by means of typed feature structure. This step offers a very precise description (a formal model) starting from which the coding scheme is automatically produced.

8. References

Bertrand, R., Blache, P., Espesser, R., et al. 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., 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. 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

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

Di Cristo A. and D. Hirst 1996. "Vers une typologie des unitï£¡s intonatives du français.", in *actes des XXIï£¡mes Journï£¡es d'Etude sur la Parole*

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

Fernández R. and J. Ginzburg (2002) “Non-Sentencial Utterances: A Corpus Study”, in *Traitement Automatique des Langues* vol. 43

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.

ISO-DA2009 (2009) Language Resource Management - Semantic Annotation Framework – Part 2: Dialogue acts ISOTC37/4, Working Draft

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

Jurafsky D, E. Shriberg and D. Bisaca (1997) Switchboard SWBD-DAMSL shallow discourse-function annotation (coders manual, draft 13)

Nesterenko I., Rauzy S. Bertrand R. 2010. “Prosody in a corpus of French spontaneous speech: perception, annotation and prosody syntax interaction?”, in proceedings of *Speech Prosody 2010*

M.-P. Péry-Woodley and N. Asher and P. Enjalbert and F. Benamara and M. Bras and C. Fabre and S. Ferrari and L.-M. Ho-Dac and A. Le Draoulec and Y. Mathet and P. Muller and L. Prévot and J. Rebeyrolle and L. Tanguy and M. Vergez-Couret and L. Vieu and A. Widlocher (2009) “ANNODIS : une approche outillée de l’annotation de structures discursives”, in proceedings of *TALN 2009*

Petukhova V. and H. Bunt (2009) *Dimensions in communication*, TiCC Technical Report 2009"

Petukhova V. and H. Bunt (2009) “The independence of dimensions in multidimensional dialogue act annotation”, in proceedings of *Human Language Technologies*

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