Multimodal Annotation of Conversational Data

P. Blache[1], **R. Bertrand**[1], **B. Bigi**[1], **E. Bruno**[3], **E. Cela**[6], **R. Espesser**[1], **G. Ferré**[4], **M. Guardiola**[1], **D. Hirst**[1], **E.-P. Magro**[6], **J.-C. Martin**[2], **C. Meunier**[1], **M.-A. Morel**[6], **E. Murisasco**[3], **I Nesterenko**[1], **P. Nocera**[5], **B. Pallaud**[1], **L. Prévot**[1], **B. Priego-Valverde**[1], **J. Seinturier**[3], **N. Tan**[2], **M. Tellier**[1], **S. Rauzy**[1]

(1) LPL-CNRS-Université de Provence (2) LIMSI-CNRS-Université Paris Sud
(3) LSIS-CNRS-Université de Toulon (4) LLING-Université de Nantes
(5) LIA-Université d'Avignon (6) RFC-Université Paris 3
blache@lpl-aix.fr

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

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.

Before entering into more details, let us mention that our data, tools and conventions are described and freely downlodable from our website (http ://www.lpl-aix.fr/ otim/).

The annotation process relies on several tools and conventions, most of them elaborated within the framework of the project. In particular, we propose a generic transcription convention, called *Enriched Orthographic Trancription*, making it possible to annotate all specific pronunciation and speech event, facilitating signal alignment. Different tools have been used in order to prepare or directly annotate the transcription : grapheme-phoneme converter, signal alignment, syllabification, prosodic analysis, morpho-syntactic analysis, chunking, etc. Our ambition is to propose a large corpus, providing rich annotations in all the different linguistic domains, from prosody to gesture. We describe in the following our first results.

2 Annotations

We present in this section some of the annotations of a large conversational corpus, called CID (Corpus of Interactional Data, see (Bertrand08)), consisting in 8 dialogues, with audio and video signal, each lasting 1 hour.

Transcription : The transcription process is done following specific conventions derived from that of the GARS (Blanche-Benveniste87). The result is what we call an *enriched orthographic construction*, from which two derived transcriptions are generated automatically : the standard orthographic transcription (the list of *orthographic tokens*) and a specific transcription from which the *phonetic tokens* are obtained to be used by the grapheme-phoneme converter.

From the phoneme sequence and the audio signal, the aligner outputs for each phoneme its time localization. This aligner (Brun04) is HMM-based, it uses a set of 10 macro-classes of vowel (7 oral and 3 nasal), 2 semi-vowels and 15 consonants. Finally, from the time aligned phoneme sequence plus the EOT, the orthographic tokens is time-aligned.

Syllables : The corpus was automatically segmented in syllables. Sub-syllabic constituents (onset, nucleus and coda) are then identified as well as the syllable structure (V, CV, CCV, etc.). Syllabic position is specified in the case of polysyllabic words.

Prosodic phrasing : Prosodic phrasing refers to the structuring of speech material in terms of boundaries and groupings. Our annotation scheme supposes the distinction between two levels of phrasing : the level of accentual phrases (AP, (Jun, 2002)) and the higher level of intonational phrases

[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.

Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010, pages 186–191,
Uppsala, Sweden, 15-16 July 2010. ©2010 Association for Computational Linguistics

(IP). Mean annotation time for IPs and APs was 30 minutes per minute.

Prominence : The prominence status of a syllable distinguishes between accentuability (the possibility for syllable to be prominent) and prominence (at the perception level). In French the first and last full syllables (not containing a schwa) of a polysyllabic word can be prominent, though this actual realization depends on speakers choices. Accentuability annotation is automatic while prominence annotation is manual and perceptually based.

Tonal layer : 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 : 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). Among the main gesture types, we annotate iconics, metaphoric, deictics, beats, emblems, butterworths or adaptors.

We used the Anvil tool (Kipp, 2004) for the manual annotations. We created a specification files taking into account the different information types and the addition of new values adapted to the CID corpus description (e.g. we added a separate track *Symmetry*). For each hand, the scheme has 10 tracks. 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 to annotate gesture lemmas (Kipp, 2004), the shape and orientation of the hand during the stroke, the gesture space, and contact. We added the three tracks to code the hand trajectory, gesture velocity and gesture amplitude.

Discourse and Interaction : Our discourse annotation scheme relies on multidimensional frameworks 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.

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*. Our rhetorical layer is an adaptation of an existing schema developed for monologic written data in the context of the ANNODIS project.

Disfluencies : Disfluencies are organized around an interruption point, which can occur almost anywhere in the production. Disfluencies can be prosodic (lenghtenings, silent and filled pauses, etc.), or lexicalized. In this case, they appear as a word or a phrase truncation, that can be completed. 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, repairing 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 Quantitative information

We give in this section some indication about the state of development of the CID annotation.

Hand gestures : 75 minutes involving 6 speakers have been annotated, yielding a total number of 1477 gestures. The onset and offset of gestures correspond to the video frames, starting from and going back to a rest position.

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

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

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

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

Disfluencies 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 15mns for 1 minute of the corpus. The following table illustrates the fact that disfluencies are speaker-dependent in terms of quantity and type. These figures also 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

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

Syntax We used the stochastic parser developed at the LPL (Blache&Rauzy, 2008) to automaticaly generate morppho-syntactic and syntactic annotations. The parser has been adapted it in order to account for the specificities of speech analysis. First, the system implements a segmentation technique, identifying large syntactic units that can be considered as the equivalent of sentences in written texts. This technique distinguishes between strong and weak or soft punctuation marks. A second modification concerns the lexical frequencies used by the parser model in order to capture phenomena proper to conversational data.

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

Category	Count	Group	Count
adverb	15123	AP	3634
adjective	4585	NP	13107
auxiliary	3057	PP	7041
determiner	9427	AdvP	15040
conjunction	9390	VPn	22925
interjection	5068	VP	1323
preposition	8693	Total	63070
pronoun	25199		
noun	13419	Soft Pct	9689
verb	20436	Strong Pct	14459
Total	114397	Total	24148

4 Evaluations

Prosodic annotation : Prosodic annotation of 1 dialogue has been done by 2 experts. The annotators worked separately using Praat. Intertranscriber agreement studies were done for the annotation of higher prosodic units. First annotator marked 3,159 and second annotator 2,855

Intonational Phrases. Mean percentage of inter-transcriber agreement was 91.4% and mean kappa-statistics 0.79, which stands for a quite substantial agreement.

Gesture : 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).

Three coders have annotated three minutes for *GestureSpace* including *GestureRegion* and *GestureCoordinates*. The kappa values indicated that the agreement is high for *GestureRegion* of right hand (kappa = 0.649) and left hand (kappa = 0.674). However it is low for *GestureCoordinates* of right hand (k= 0.257) and left hand (k= 0.592). Such low agreement of *GestureCoordinates* might be due to several factors. First, the number of categorical values is important.

Second, three minutes might be limited in terms of data to run a kappa measure. Third, GestureRegion affects GestureCoordinates : if the coders disagree about GestureRegion, they are likely to also annotate GestureCoordinates in a different way. For instance, it was decided that no coordinate would be selected for a gesture in the center-center region, whereas there is a coordinate value for gestures occurring in other parts of the GestureRegion. This means that whenever coders disagree between the center-center or center region, the annotation of the coordinates cannot be congruent.

5 Information representation

5.1 XML encoding

Our approach consists in first precisely define the organization of annotations in terms of typed-feature structures. We obtain an abstract description from which we automatically generate a formal schema in XML. All the annotations are then encoded following this schema.

Our XML schema, 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).

5.2 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 XQuery. From the XML encoding 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 which supports Allen's relations.

6 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 systÃÍme 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.