Creating and exploiting multimodal annotated corpora

Philippe Blache, Roxane Bertrand & Gaëlle Ferré

Laboratoire Parole et Langage
CNRS & Université de Provence

LREC 2008

Introduction

- Multimodality
    - Information comes from different sources
    - Modalities interaction
        - Each source is partial, incomplete
        - They have to be synchronized
- Multimodal annotation
    - Goals
        - Usually focus on gesture description
        - Mainly in the perspective of communication
    - Conventions and schemes
    - Tools (Praat, Anvil, Elan, etc.)
- Our project
    - Linguistic description
    - Study of interaction: annotation of all domains
    - Unrestricted data (natural situations)

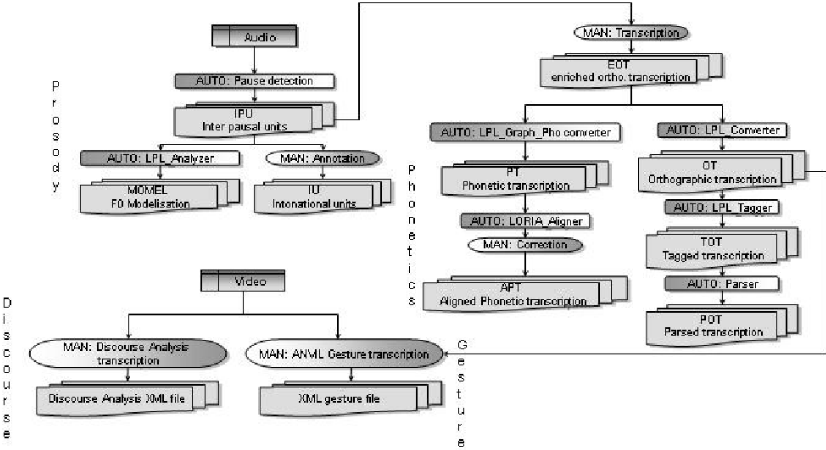
Outline

- The project
  - The CID corpus
  - The annotation process
- Results
  - Backchannels
  - Reinforcing gestures
- Perspectives

The corpus

- *Corpus of Interactional Data*: 8 dialogs, 1 hour each ([Bertrand & al 07])
- Transcribed (orthographic, phonetic)
- Aligned
- Annotated
  - Prosody (intonation, units, contours, etc.)
  - Morphosyntax, syntax,
  - Discourse (markers, speech turns, etc.)
  - Gestures

The annotation architecture



Prosody

Audio

AUTO: Pause detection

IPU
Inter pausal units

AUTO: LPL_Analyzer

MOMEL
F0 Modelisation

MAN: Annotation

IU
Intonational units

MAN: Transcription

EOT
enriched ortho. transcription

Phonetics

AUTO: LPL_Graph_Pho converter

PT
Phonetic transcription

AUTO: LORIA_Aligner

MAN: Correction

APT
Aligned Phonetic transcription

AUTO: LPL_Converter

OT
Orthographic transcription

AUTO: LPL_Tagger

TOT
Tagged transcription

AUTO: Parser

POT
Parsed transcription

Discourse

Video

MAN: Discourse Analysis transcription

Discourse Analysis XML file

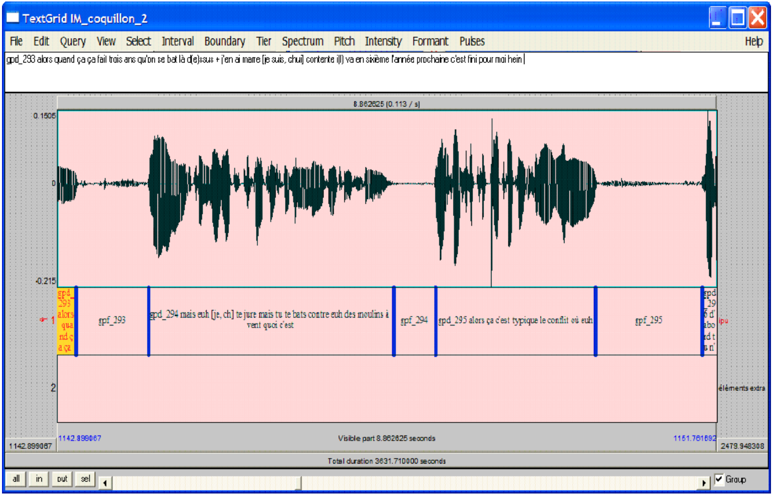
MAN: ANVIL Gesture transcription

XML gesture file

Gesture

Signal segmentation

- Interpausal units segmentation (IPUs)
- Syntactic units detection (pattern method)



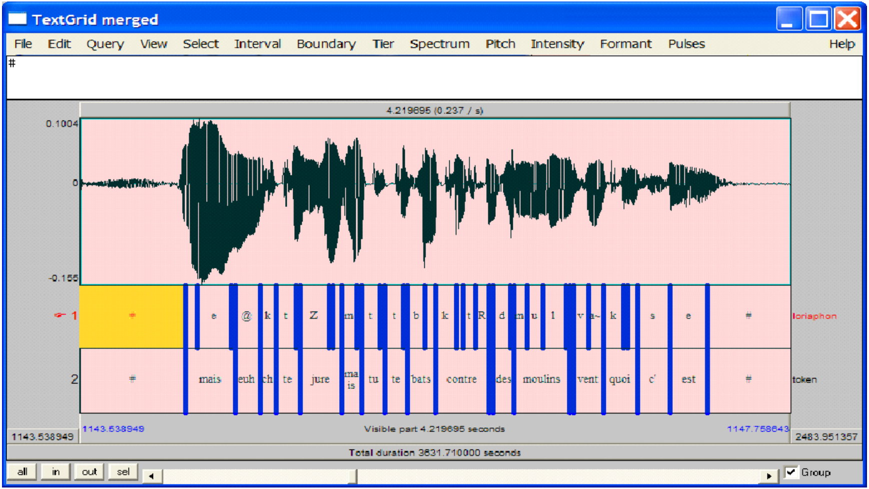
TextGrid IM_coquillon_2

gpd_293 alors quand ça ça fait trois ans qu'on se bat là d(e)ssus + j'en ai marre [je suis, chui] contente i(l) va en sixième l'année prochaine c'est fini pour moi hein

Transcription

- Precise transcription convention
- Transcription by 2 experts
- Enriched orthographic transcription (EOT), needed for different phenomena annotation and alignment (elisions, schwa, etc.)
- Generation of 2 transcription versions
    - Orthographic (for the NLP module)
    - Phonetic (for speech analysis)

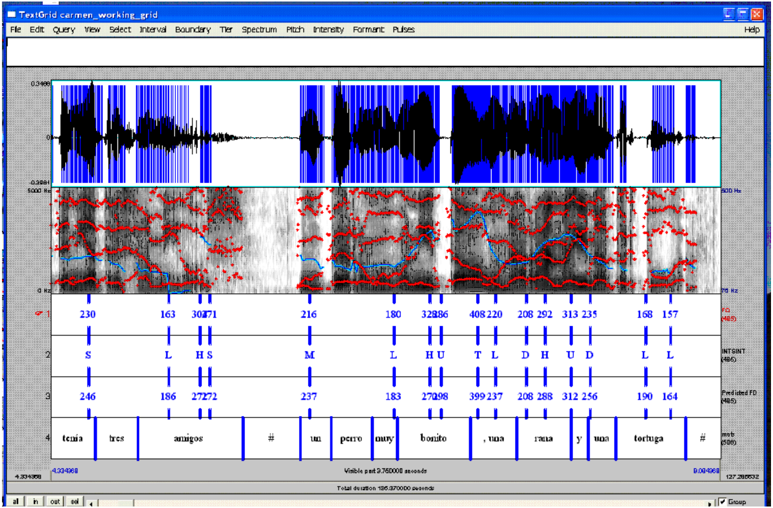
Alignment


TextGrid merged
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
#
4.219695 (0.237 / s)
0.1004
0
-0.155
1
# e @ k t Z m t t b k t R d m u l v a~ k s e #
loriaphon
2
# mais euh ch te jure ma is tu te bats contre des moulins vent quoi c' est #
token
1143.538949
1143.538949
Visible part 4.219695 seconds
1147.758643
2483.951357
Total duration 3831.710000 seconds
all in out sel
Group

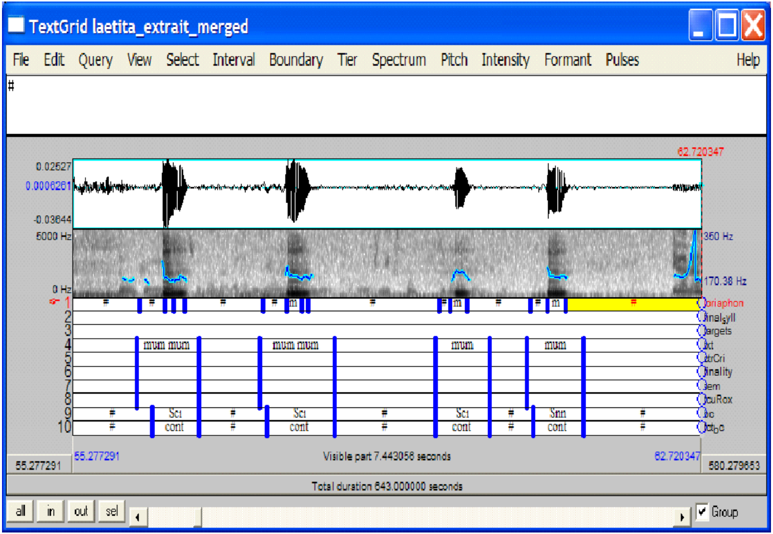
Alignment

- Identifying the phoneme suite
  - Tokenisation
  - Grapheme-phoneme conversion
- Alignment tool
  - Input: list of phonemes + audio signal
  - Temporal localization of the phonemes in the signal
- Manual correction
  - Wrong boundaries
  - Overgeneration (false units)
- Tokens and phonemes are primary levels, used for anchoring other levels

Intonation: INTSINT


TextGrid carmen_working_grid
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
0.3466
0
-0.3981
5000 Hz
0 Hz
500 Hz
70 Hz
1 230 163 302 271 216 180 328 286 408 220 208 292 313 235 168 157
F0 (485)
2 S L H S M L H U T L D H U D L L
INTSINT (486)
3 246 186 272 272 237 183 270 298 399 237 208 288 312 256 190 164
Predicted F0 (485)
4 tenía tres amigos # un perro muy bonito , una rana y una tortuga #
mots (506)
4.334368
4.334968
Visible part 3.760000 seconds
8.094968
127.286632
Total duration 186.370000 seconds
all in out sel
Group

Discourse

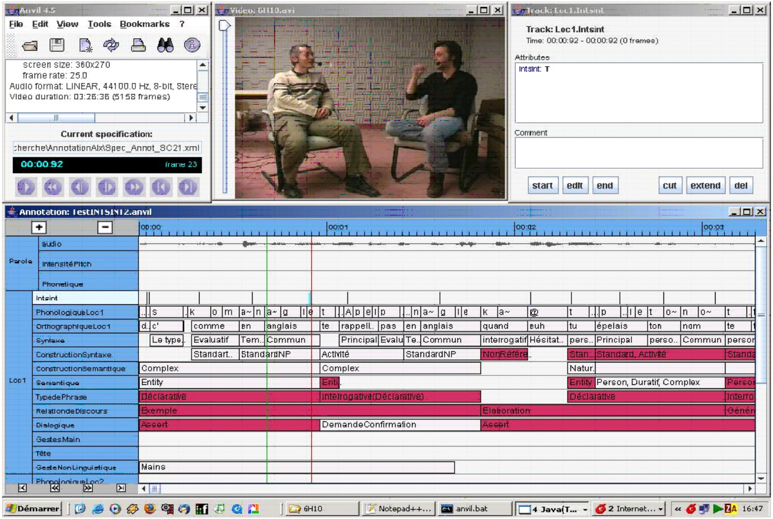


TextGrid laetita_extrait_merged

Visible part 7.443056 seconds

Total duration 643.000000 seconds

Gestures



Anvil 4.5

screen size: 360x270
frame rate: 25.0
Audio format: LINEAR, 44100.0 Hz, 8-bit, Stere
Video duration: 03:26:36 (5158 frames)

Current specification:
:herche\AnnotationAlx\Spec_Annot_SC21.xml

00:00:92 frame 23

Video: 6H10.avi

Track: Loc1.Intsint

Track: Loc1.Intsint
Time: 00:00:92 - 00:00:92 (0 frames)

Attributes
Intsint: T

Comment

Annotation: TestINTSINT2.anvil

00:00 00:01 00:02 00:03

Parole: audio, IntensitéPitch, Phonetique

Loc1:

Intsint

PhonologiqueLoc1: .. s .k o m a~ n a~ g l e t ..A p e l p ..n a~ g l e k a~ @ t ..p ..l e t o~ n o~ t .t

OrthographiqueLoc1: d. c' comme en anglais te rappell.. pas en anglais quand euh tu épelais ton nom te

Syntaxe: Le type. Evaluatif Tem.. Commun Principal Evalu Te. Commun interrogatif Hésitat.. pers.. Principal perso.. Commun persor

ConstructionSyntaxe: Standart.. StandardNP Activité StandardNP NonRéfére.. Stan.. Standard, Activité Standa

ConstructionSemantique: Complex Complex Natur.

Semantique: Entity Enti. Entity Person, Duratif, Complex Persor

TypedePhrase: Déclarative Interrogative(Déclarative) Déclarative Interro

RelationdeDiscours: Exemple Elaboration Génér

Dialogique: Assert DemandeConfirmation Assert

GestesMain

Tête

GesteNonLinguistique: Mains

PhonologiqueLoc2

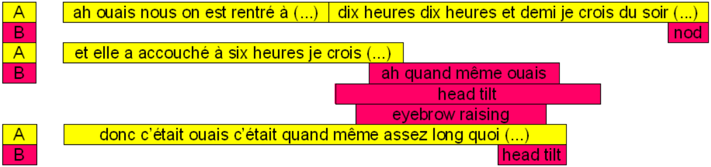
File Edit View Tools Bookmarks ?
start edit end cut extend del
+ −
Démarrer 6H10 Notepad++... anvil.bat 4 Java(T... 2 Internet... 16:47

Summary of the tools

- Fully automatic
  - IPU segmentation
  - Phoneme alignment
  - Intonation
  - POS tagging
- Semi-automatic
  - Intonational units
  - Shallow parsing (still needs a segmentation tool)
- Manual
  - Transcription (we are experimented speech recognition as helping tool)
  - Other annotations
- Tools and resources available from the CRDO (http://crdo.fr/)

First study: Backchannels

- Backchannels: minimal signal produced by the hearer. Vocal and gestural BCs (head movements, smiles and laughter, eyebrow movements, etc.), they have different functions
- *Example*:



A ah ouais nous on est rentré à (...) dix heures dix heures et demi je crois du soir (...)

B nod

A et elle a accouché à six heures je crois (...)

B ah quand même ouais

head tilt

eyebrow raising

A donc c'était ouais c'était quand même assez long quoi (...)

B head tilt

[A] oh yeah we were admitted at 10, 10.30 I think pm
[A] and she had the baby at 6 I think
[B] [oh yeah right?]
[A] so it was yeah it was quite long indeed

- **Question**: Do vocal and gestural BCs behave similarly? In what prosodic and morphological contexts do they appear?

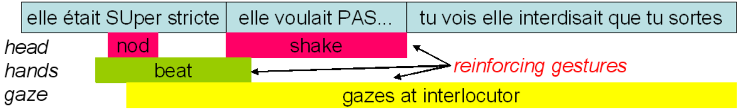
Backchannels

- Vocal and gestural BCs show similar behavior but gestural BCs appear later than vocal ones
- Morphological and discursive context
  - After nouns, verbs and adverbs (words with semantic function)
  - Not after connectors (linking words between conversational units)
- Prosodic context
  - Gestural BCs: after accentual phrases (APs) and intonational phrases (IPs)
  - Vocal BCs: after IPs
  - Encouraged by specific contours (esp. rising), speakers gaze
- *Conclusion*: BCs occur at the end of some units, but not with possible turn change. They also play a role in the elaboration of discourse.

Second study: Reinforcing gestures

- Reinforcing gestures: eyebrow movements, gaze direction, head movements, highlighting discourse elements

- *Example*:



elle était SUper stricte | elle voulait PAS... | tu vois elle interdisait que tu sortes

head nod shake

hands beat

gaze gazes at interlocutor

reinforcing gestures

she [the teacher] was super strict she didn't want... you see she forbade us to leave the room [during lessons]

- **Questions**: What do gestures reinforce? Are they equivalent to known focalization phenomena?

Reinforcing gestures: results

- No correlation with prosodic focalization, no gesture is associated with specific stress or contour
- Correlation with adverbs and connectors at the beginning of speech turns
- Correlation for metaphorics, no correlation for eyebrow movements
- *Conclusion*
  - Reinforcing gestures do not serve to express focus
  - Their role is more discursive than expressive

Conclusion

- CID: large corpus, richly annotated
- Interest of multimodal annotated corpora
  - Study of natural language, in context
  - Study of interaction
- Problems
  - Standardisation: coding schemes
  - Synchronization of the different domains (+/- temporal)
  - Interfacing the different tools
- Perspectives
  - Information structure study
  - Description in terms of constructions (CxG)
  - Multimodal interaction for virtual reality