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

Philippe Blache, Roxane Bertrand, Mathilde Guardiola, Marie-Laure Guénot, Christine Meunier, Irina Nesterenko, Berthille Pallaud, *Laurent Prévot*, Béatrice Priego-Valverde, Stéphane Rauzy

LPL, CNRS & Université de Provence

FirstName.LastName@lpl-aix.fr

LREC, La Valetta
May 21st, 2010

Aims and Scope

- Multilevel analysis of multimodal data
- Broad project aiming at establishing methodologies and best practices for handling large scale data
  - Annotation tools and methodologies
  - Exploitation of the annotated data
- Main corpus studied : Corpus of Interactional Data [Bertrand et al., 2008]
  - Reduce the gap between experimental and field linguistics
  - Project not bound to this corpus

OTIM Project

OTIM : Funded ANR project [2009-2011]
Tools for Processing Multimodal Information (LPL, LSIS, LIMSI, LIA, LLING)

- Examples of studies planned :
  - syntactic / prosodic / discourse boundaries
  - gestures / prosody / conversation structure
  - acoustic properties / turn-taking, ...
- Activities
  - Annotation
  - Identify and complete a set of NLP tools for helping linguistic annotation (syllaber, text/speech aligner, tagger, chunker, parser, segmenters,...)
  - Develop a XML rich querying framework on multi-structure objects (LSIS)
  - Tools for interoperability : format converters, intermediate language for interoperability (LPL, LSIS)

Corpus of Interactional Data (CID)

Goal : study prosody and interactional aspects ⇝ focus on recording quality while preserving spontaneity and "freedom of speech"

Corpus aiming at reducing the gap between experimental and field linguistic studies

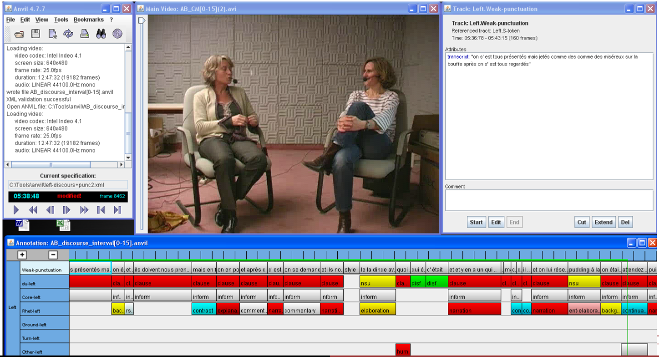
- 8 hours of French conversations
- 2 microphones / anechoic room
- 1 camrecorder facing the speakers

Corpus of Interactional Data (CID)

Goal : study prosody and interactional aspects ⇝ focus on recording quality while preserving spontaneity and "freedom of speech"

Corpus aiming at reducing the gap between experimental and field linguistic studies

- 8 hours of French conversations
- 2 microphones / anechoic room
- 1 camrecorder facing the speakers



```
Loading video:
    video codec: Intel Indeo 4.1
    screen size: 640x480
    frame rate: 25.0fps
    duration: 12:47:32 (19182 frames)
    audio: LINEAR 44100.0Hz mono
wrote file AB_discourse_interval[0-15].anvil
XML validation successful
Open ANVIL file: C:\Tools\anvil\AB_discourse_ir
Loading video:
    video codec: Intel Indeo 4.1
    screen size: 640x480
    frame rate: 25.0fps
    duration: 12:47:32 (19182 frames)
    audio: LINEAR 44100.0Hz mono
```

Current specification:

C:\Tools\anvil\left-discours+punc2.xml

05:38:48 modified! frame 8462

Track: Left.Weak-punctuation

Referenced track: Left.S-token

Time: 05:36:78 - 05:43:15 (160 frames)

Attributes

transcript: "on s' est tous présentés mais jetés comme des comme des miséreux sur la bouffe après on s' est tous regardés"

Comment

Corpus of Interactional Data (CID)

Goal : study prosody and interactional aspects ⇝ focus on recording quality while preserving spontaneity and "freedom of speech"

Corpus aiming at reducing the gap between experimental and field linguistic studies

- 8 hours of French conversations
- 2 microphones / anechoic room
- 1 camrecorder facing the speakers

Protocol :

- “You have 1 hour to talk about things unusual” or “to talk about professional conflicts”
- Participants know each other.

Charasteristics of the corpus

- Highly spontaneous
- Highly interactional (designed for this purpose)
- Alternation of narrative storytelling phases and transition/commenting phases
- Significant amount of overlapping speech

+ high recording quality

Annotations performed

- High quality enriched transcription (including lengthening, mispronunciations...)
- phoneme/sound alignment + syllable grouping (Automatic)
- Prosodic prominences and contours
- Syntactic analysis (chunking and parsing) (Automatic)
- Disfluencies
- Discourse and Interaction
- Gestures (Posture, Face, Hands, Gaze)

Done by different teams in France (LPL, LIMSI, LLING)
Tools used : Praat, ANVIL, ELAN

Enriched transcription

(1) et puis euh je commence à descendre après l(e) premier virage j(e) me casse la gueule me (d)is oh [merde, merdeu] oh quand même @ la saison commence mal et puis euh bon je [rechausse, rechause]

then I start descending / and after the first curve I fall / I tell to myself / Damn it, the season starts bad / and then I put my skis on

Alignment process :

1. Enriched transcription
2. grapheme-phoneme converter
3. Automatic alignment phoneme/sound

The need of a formal model

- Many people from different research traditions
- Several tools (Praat, Anvil, Elan)
- Many levels of analysis must be integrated in one homogeneous “database”

⇝ Not doable if people did not agree on a set of principles for representing the annotated information

⇝ Premilinary to the different annotation schemas

The formal model, basics

Expressed in Typed Feature Structures

- Ingredients : objects, subtype relation, constituence relation, features
- Each object has features
- Each object has a location
  - currently only temporal locations : intervals and points
  - but discontinuous or spatial location are allowed
- Location can be given explicitly by a spatio-temporal feature or coming from constituency structure

The formal model, basics

Expressed in Typed Feature Structures

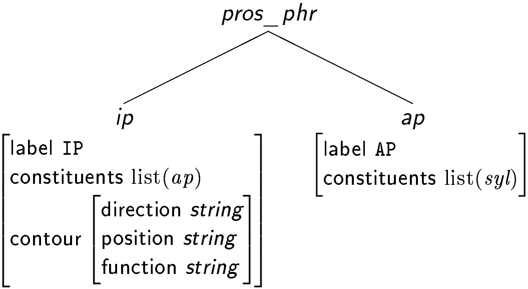
- Ingredients : objects, subtype relation, constituence relation, features
- Each object has features
- Each object has a location
  - currently only temporal locations : intervals and points
  - but discontinuous or spatial location are allowed
- Location can be given explicitly by a spatio-temporal feature or coming from constituency structure

ip : := ap*
ap : := syl$^+$
syl : := const_syl$^+$
const_syl : := phon$^+$
disf : := reprandum break reprans

A formal model, Phoneme

$$
{}_{\textit{phon}}\begin{bmatrix}
\text{sampa\_label} & \textit{sampa\_unit} \\
\text{cat} & \left\{ \textit{vowel, consonant} \right\} \\
\text{type} & \left\{ \textit{occlusive, fricative, nasal, etc.} \right\} \\
\text{artic\_gest} & \begin{bmatrix}
\text{lip} & \begin{bmatrix} \text{protusion} & \textit{string} \\ \text{aperture} & \textit{aperture} \end{bmatrix} \\
\text{tongue} & \begin{bmatrix}
\text{tip} & \begin{bmatrix} \text{location} & \textit{string} \\ \text{degree} & \textit{string} \end{bmatrix} \\
\text{body} & \begin{bmatrix} \text{location} & \textit{string} \\ \text{degree} & \textit{string} \end{bmatrix}
\end{bmatrix} \\
\text{velum} & \textit{aperture} \\
\text{glottis} & \textit{aperture}
\end{bmatrix} \\
\text{role} & \begin{bmatrix} \text{epenthetic} & \textit{boolean} \\ \text{liaison} & \textit{boolean} \end{bmatrix}
\end{bmatrix}
$$

Prosody, Type hierarchy



pros_phr

ip

label IP

constituents $\mathrm{list}(ap)$

contour

direction *string*

position *string*

function *string*

ap

label AP

constituents $\mathrm{list}(syl)$

Prosody, an annotated IP

$${}_{ip}\begin{bmatrix} \text{label} & \texttt{IP} \\ \text{index} & 18 \\ \text{location} & \begin{bmatrix} \text{start} & 83.11 \\ \text{end} & 204.21 \end{bmatrix} \\ \text{constituents} & \left[\left\{ \text{ap} \begin{bmatrix} \text{label} & \texttt{AP} \\ \text{index} & 25 \\ \text{location} & \begin{bmatrix} \text{start} & 192.28 \\ \text{end} & 204.21 \end{bmatrix} \end{bmatrix} \right\}\right] \\ \text{contour} & \begin{bmatrix} \text{direction} & \text{falling} \\ \text{position} & \text{final} \\ \text{function} & \text{conclusive} \end{bmatrix} \end{bmatrix}$$

Discourse units

$${}_{du}\left[\begin{array}{ll}
\text{index} & \textit{integer} \\
\text{constituents} & \textit{set(token)} \\
\text{form} & \textit{du\_form} \\
\text{functions} & \textit{set(}\left[\begin{array}{ll} \text{type} & \textit{communicative\_function} \\ \text{target} & \textit{set(du)} \end{array}\right]\textit{)} \\
\text{producer} & \left[\begin{array}{ll} \text{role} & \left\{\textit{hearer, speaker}\right\} \\ \text{identity} & \textit{string} \end{array}\right] \\
\text{voice} & \left[\begin{array}{ll} \text{reality} & \left\{\textit{real, fictitious}\right\} \\ \text{type} & \left\{\textit{speaker, hearer, other, generic}\right\} \end{array}\right]
\end{array}\right]$$

Relation to existing efforts

- Formal tools (Typed Feature Structures) and data format (XML) are compatible with standards
- Try to remain compatible or reuse emerging standards with regard to Annotation Schemas
- DiaML (ISO TC 37/4) (Dialogue Act Mark-up language) [ISOTC37/4, 2009]
    - Identify an interesting standard for building our Annotation Schema
    - Extend it with optional information fitting with the overall structure of the schema (Discourse Relations, Reported Speech, Humor) [Prévot et al., 2010]

Future work

Current :

- More annotations
- Annotation Guidelines development
- Deeper integration with the ISO standards
- Querying system and multi-level analysis ($\leadsto$ systematic studies cross-modalities studies)

Future :

- Tools development (discourse unit segmenter)
- OWL version of the schema

Thanks for your attention

OTIM

http ://aune.lpl.univ-aix.fr/∼otim

CRDO (Spoken Language Description Resource Center)

http ://crdo.up.univ-aix.fr/

References I

Bertrand, R., Blache, P., and Espesser, R. (2008).
Le cid - corpus of interactional data - annotation et exploitation multimodale de parole conversationnelle.
TALN, 49(3).

ISOTC37/4 (2009).
Language resource management - semantic annotation framework – part 2 : Dialogue acts.
Technical Report N442 rev5, ISO.
Working Draft.

Prévot, L., Bertrand, R., Priego-Valverde, B., and Blache, P. (2010).
Discourse and interaction in french conversations, a case study for interoperable semantic annotation.
In *Proceedings of Interoperable Semantic Annotation Workshop*.

Why Enriched transcription ?

Enriched transcription vs. orthographic transcription ?

- More costly for transcribing (between 25 to 45 minutes / 1 minute of speech)
- But can be directly processed for statistics on phonetic variations
- Current evaluation for determining which method has the best ratio cost/quality