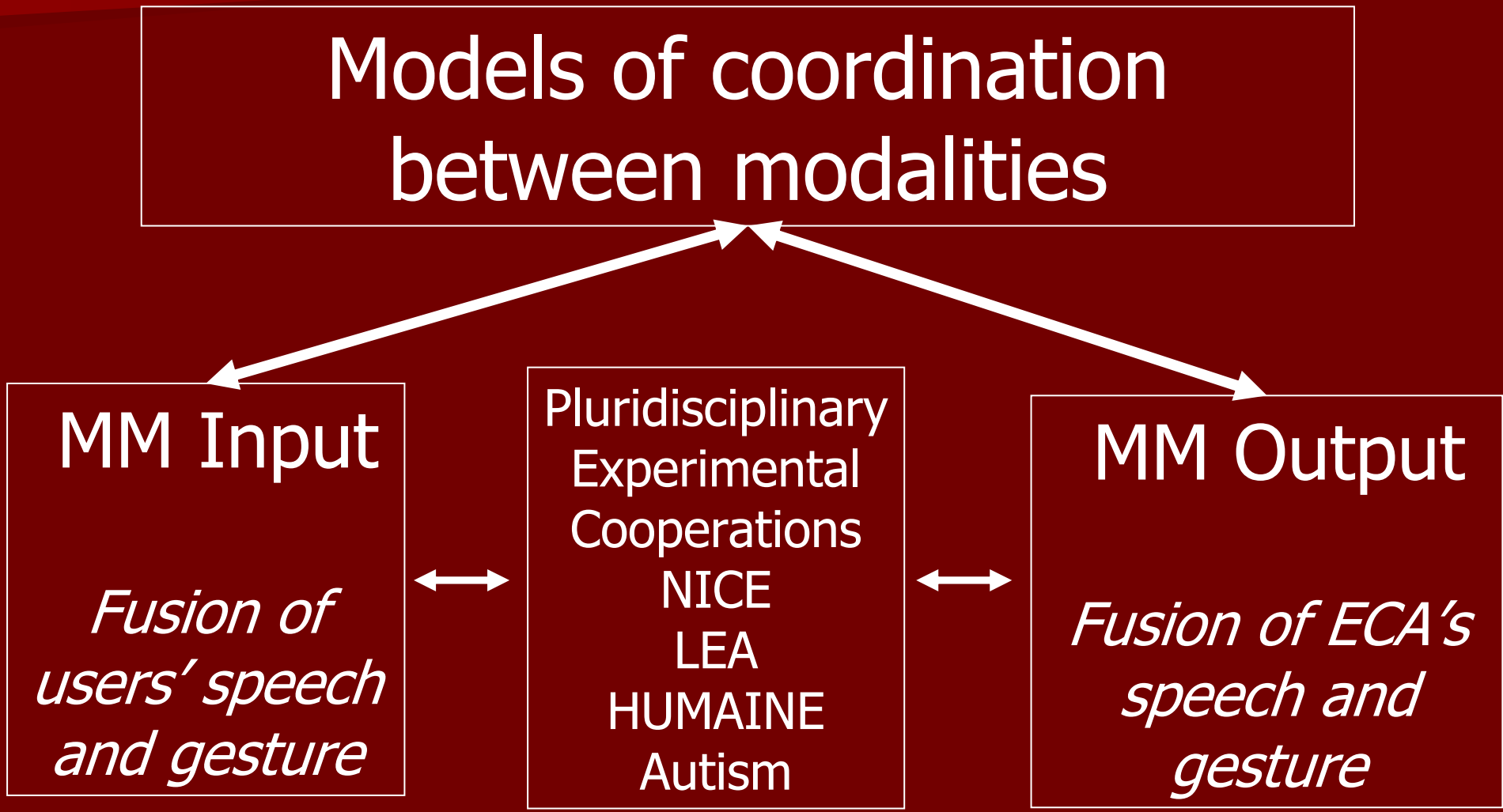
Cooperation between Modalities

Jean-Claude MARTIN
LIMSI-CNRS, Orsay, France
martin@limsi.fr
www.limsi.fr/Individu/martin

Research framework



Outline

- Introduction
- Part I : Models of Multimodal Behavior
- Part II : Input Fusion
- Part III : Embodied Conversational Agents
- Future Directions

Research context
Limitations of current HCI

- Limited configurations
  - Input:
    - keyboard, mouse
  - Output:
    - screen, speakers
- Limited dynamics
  - Understanding
  - Generation
  - Bidirectionnal

Research goals

- Study multimodal human-human communication
- For the design of more intuitive human-computer interfaces

Outline

- Introduction
- Part I : Models of Multimodal Behavior
- Part II : Input Fusion
- Part III : Embodied Conversational Agents
- Future Directions

Part I : Models of Multimodal Behavior

- Manual annotation
- HCI
  - LEA 2D
  - NICE 3D
- Human-Human
  - EmoTV Humaine

Larson, G. (1984).
The far side, Futura publications.

Deictic in multimodal input HCI

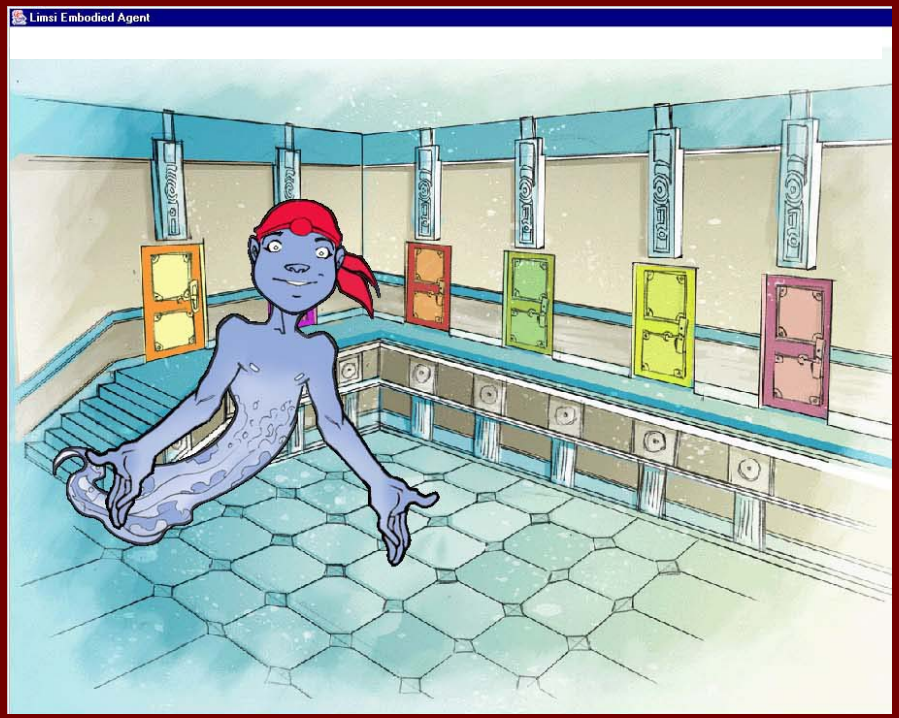
References	Application	Modalities	Some results
(Huls, Claassen & Bos, 1995)	File manager	Typed text Mouse pointing gestures	Variety in use of referring expressions
(Oviatt, De Angeli & Kuhn, 1997)	Map task	Speech Pen gestures 2D graphics	60% of multimodal constructions do not contain any spoken deictic
(Oviatt & Kuhn, 1998)	Map task	Speech Pen gestures 2D graphics	Most common deictic terms are “here”, “there”, “this”, “that”. Explicit linguistic specification of definite and indefinite reference is less common compared to speech only.
(Kranstedt, Kühnlein & Wachsmuth, 2003)	Assembly task	Speech 3D gestures 3D graphics	Temporal synchronization between speech and pointing Influence of spatio-temporal restrictions of the environment (density of objects) on deictic behavior

Requirement on behavior analysis

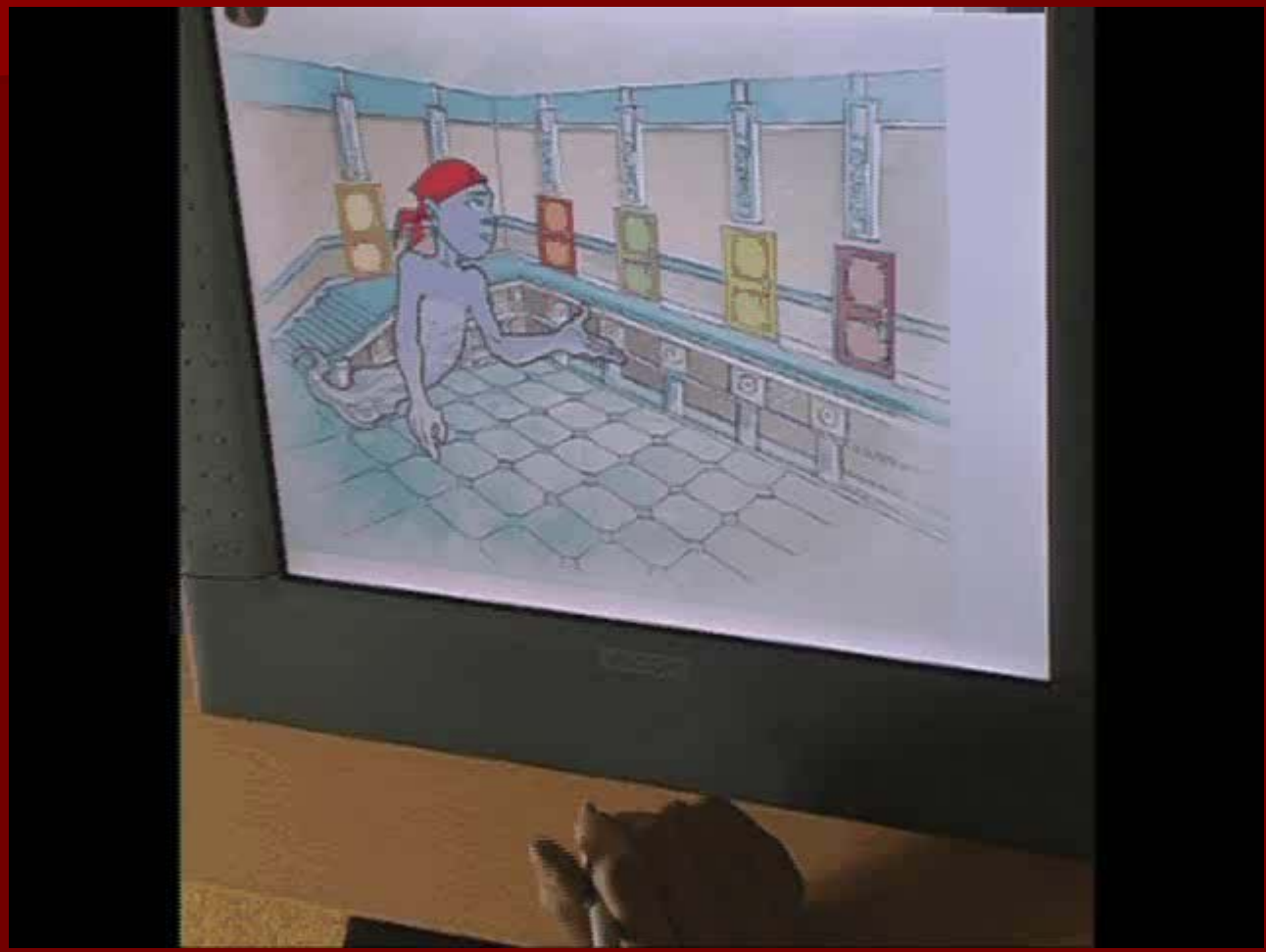
- In order to design a multimodal input system, one needs to have a model of
  - Application
  - User's spoken behavior
  - User's gestures
  - User's combination of modalities
    - TYCOON typology (Martin 95)
      - Equivalence
      - Specialisation
      - Transfer
      - Redundancy – Complementarity (Compatibility vs. Conflict)
    - Temporal patterns
    - Semantic relations
    - Relations graphics / speech / gesture (Landragin 2004)

A conversational game

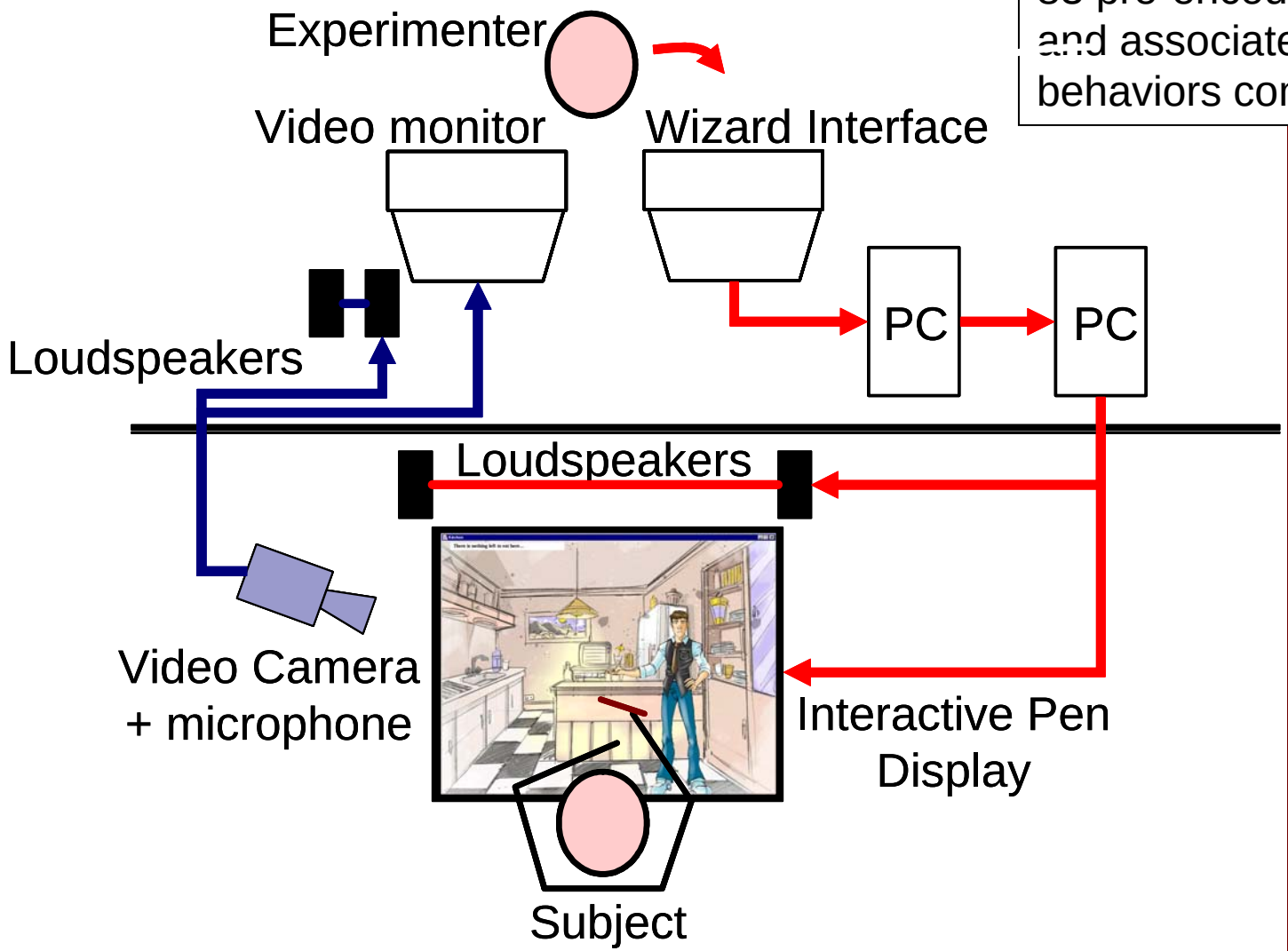
- A djiin asks for help
- Characters / rooms / objects



Video



Wizard of Oz device



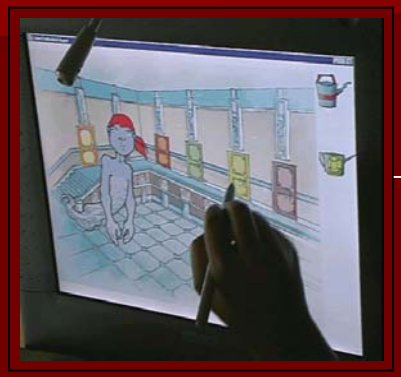
83 pre-encoded utterances and associated nonverbal behaviors combinations

- speech
- pen & speech

- 10 children
- 7 adults

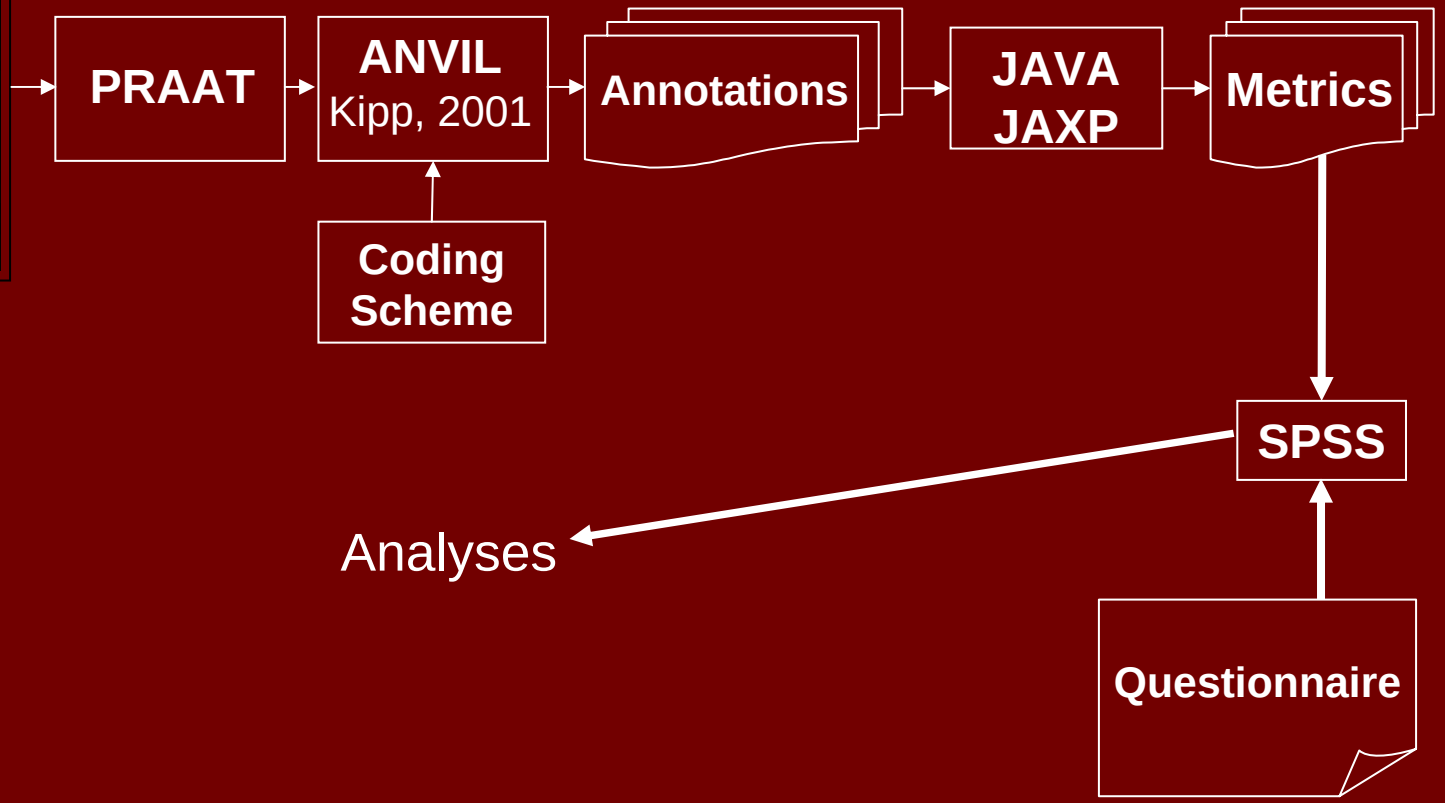
Buisine, S. and Martin, J.-C. (2005).
Children's and Adults' Multimodal Interaction with 2D Conversational Agents. CHI'2005 Portland, Oregon, 2-7 April.

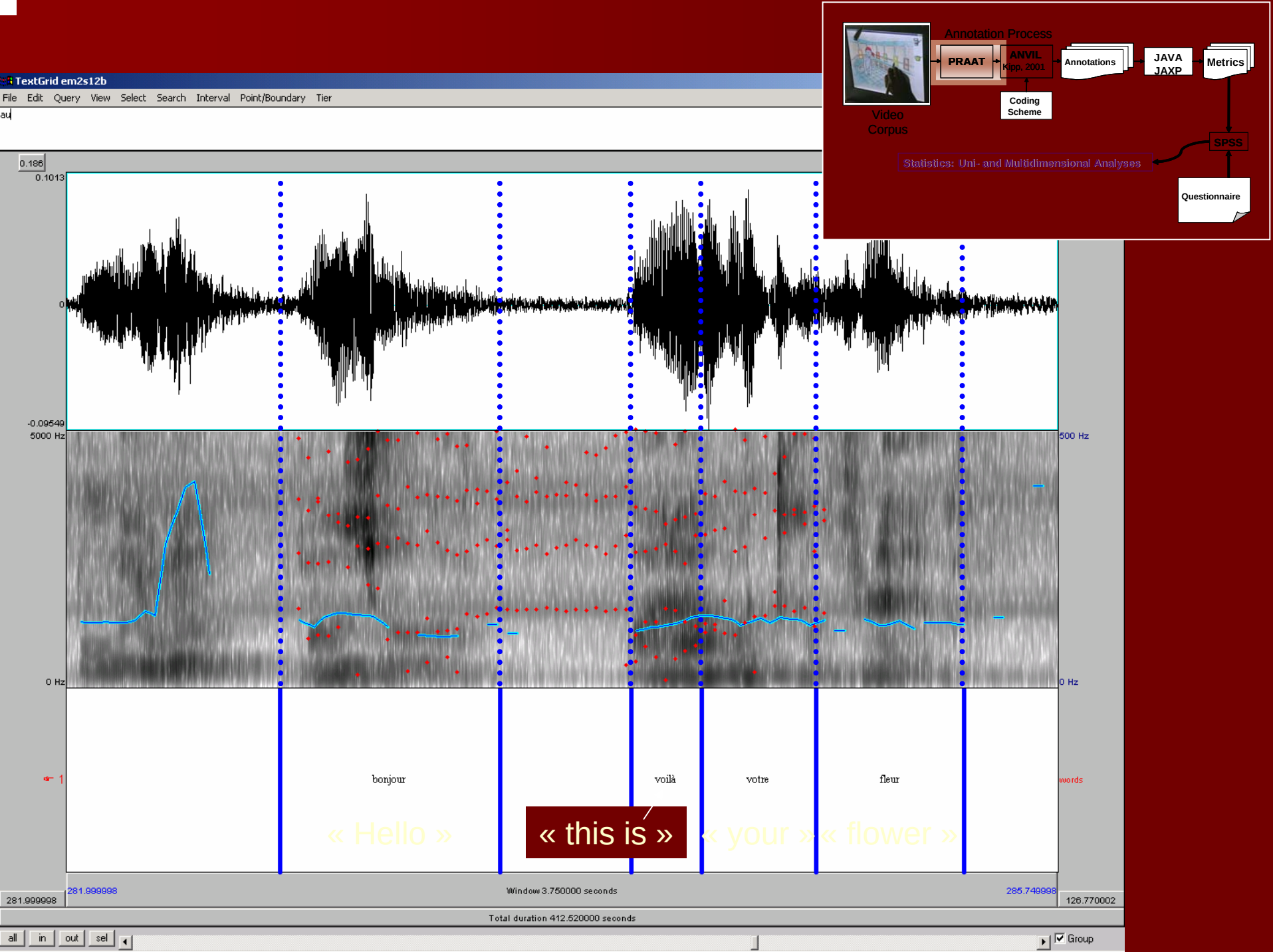
Annotation process



Video Corpus

Annotation Process

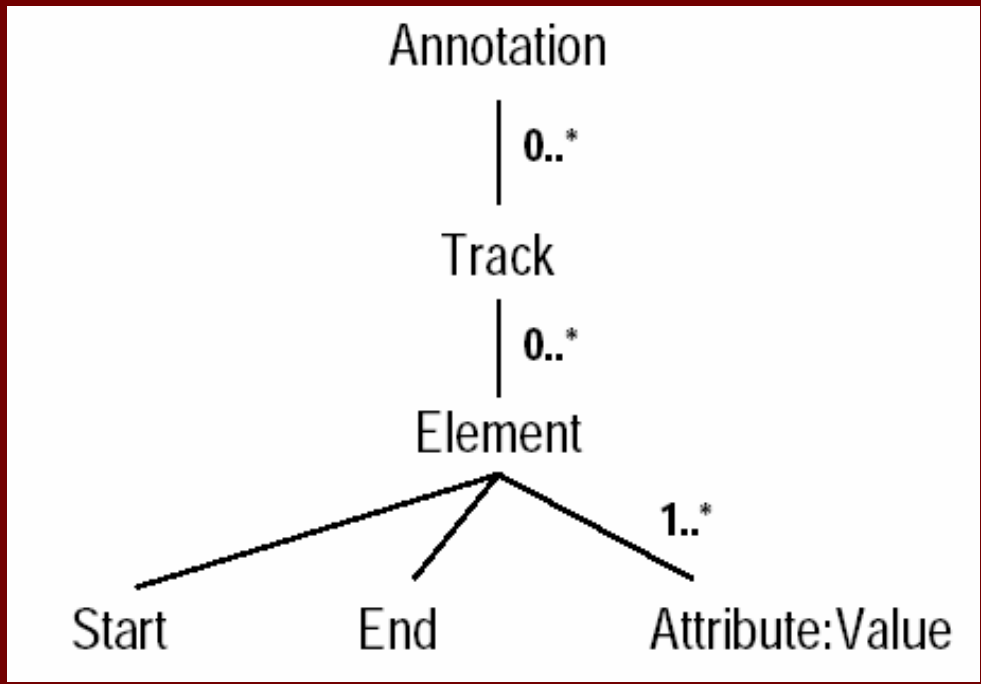



Annotation Process
PRAAT
ANVIL
Kipp, 2001
Annotations
JAVA
JAXP
Metrics
Coding
Scheme
Video
Corpus
SPSS
Statistics: Uni- and Multidimensional Analyses
Questionnaire
TextGrid em2s12b
File Edit Query View Select Search Interval Point/Boundary Tier
0.186
0.1013
0
-0.09549
5000 Hz
0 Hz
500 Hz
0 Hz
1
bonjour
voilà
votre
fleur
words
« Hello »
« this is »
« your »
« flower »
281.999998
Window 3.750000 seconds
285.749998
281.999998
126.770002
Total duration 412.520000 seconds
all
in
out
sel
Group

Anvil (Kipp 2004)

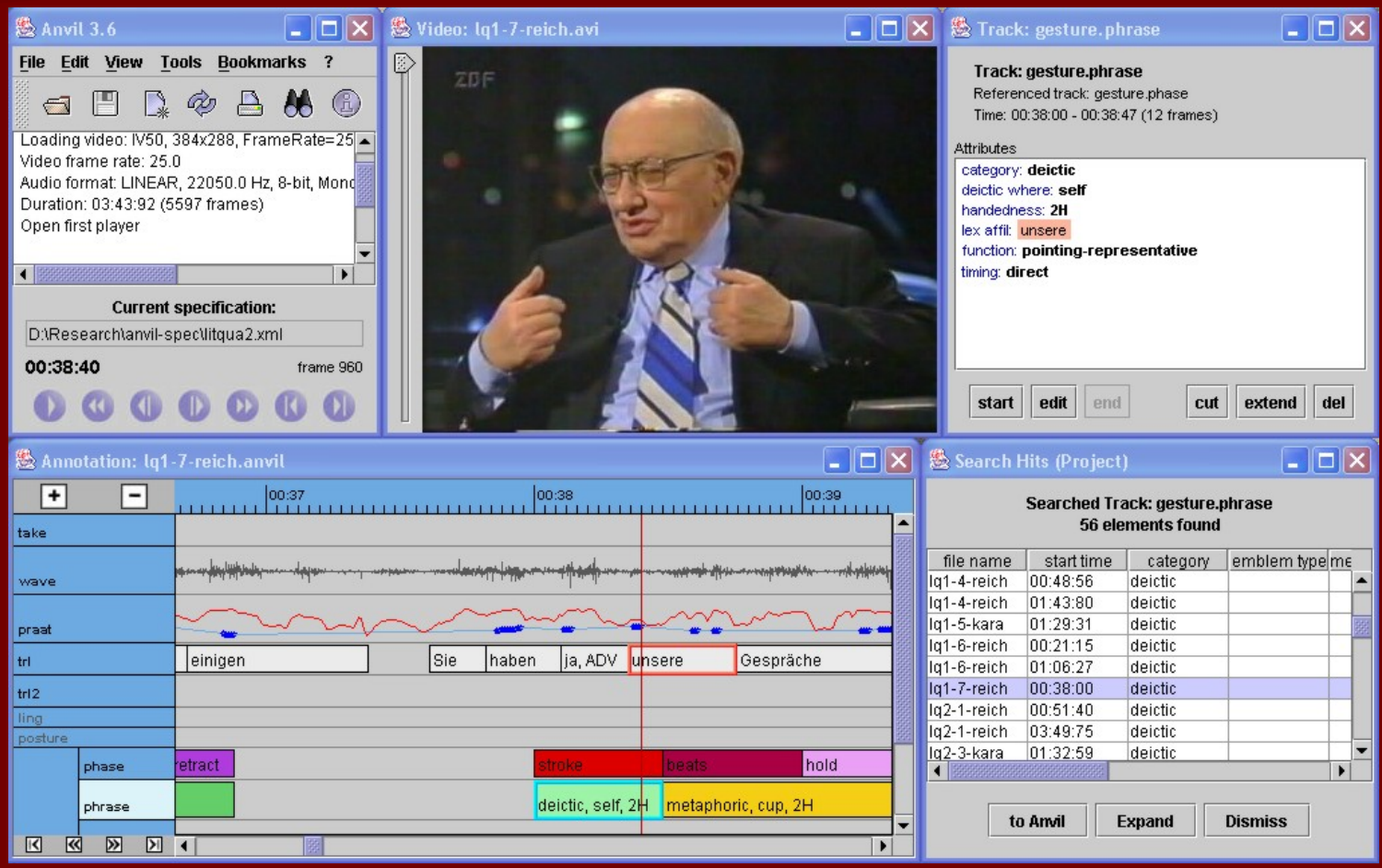
http://www.dfki.uni-sb.de/~kipp/research/index.html

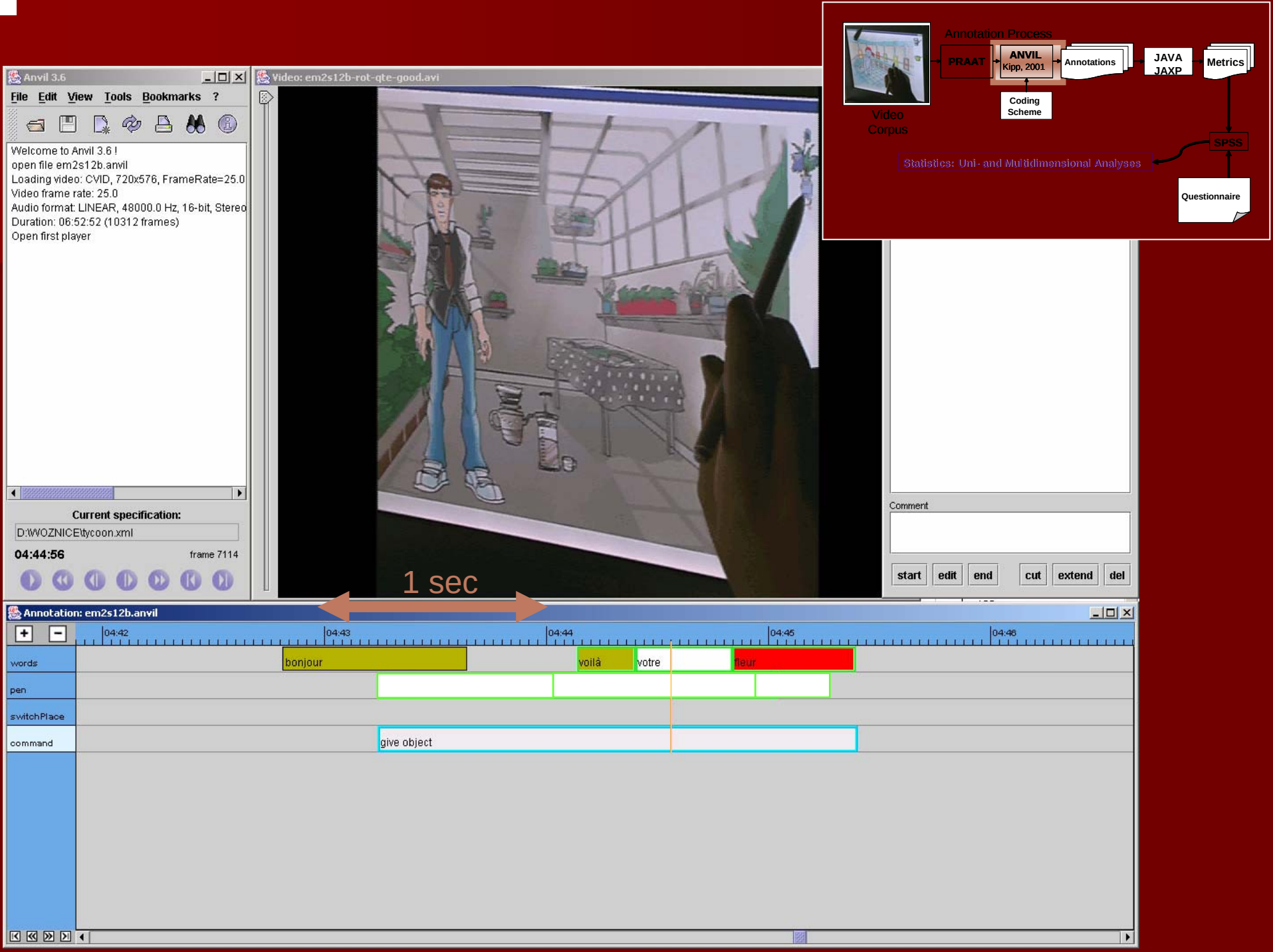
Anvil Annotation Model



Anvil (Kipp 2004)

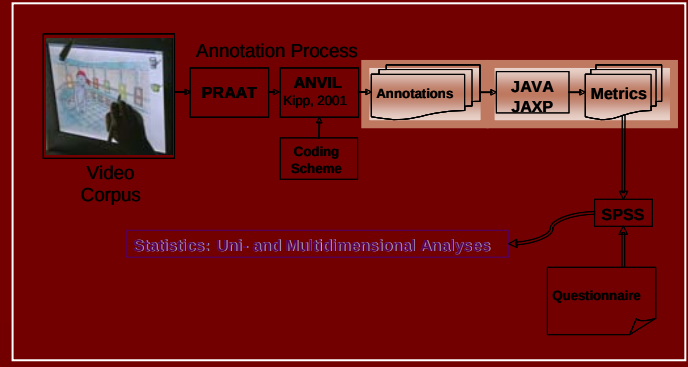
http://www.dfki.uni-sb.de/~kipp/research/index.html




Annotation Process
PRAAT
ANVIL
Kipp, 2001
Annotations
JAVA
JAXP
Metrics
Coding
Scheme
Video
Corpus
SPSS
Statistics: Uni- and Multidimensional Analyses
Questionnaire
Anvil 3.6
File Edit View Tools Bookmarks ?
Welcome to Anvil 3.6 !
open file em2s12b.anvil
Loading video: CVID, 720x576, FrameRate=25.0
Video frame rate: 25.0
Audio format: LINEAR, 48000.0 Hz, 16-bit, Stereo
Duration: 06:52:52 (10312 frames)
Open first player
Current specification:
D:\WOZNICE\tycoon.xml
04:44:56
frame 7114
Video: em2s12b-rot-qte-good.avi
1 sec
Comment
start
edit
end
cut
extend
del
Annotation: em2s12b.anvil
04:42
04:43
04:44
04:45
04:46
words
pen
switchPlace
command
bonjour
voilà
votre
fleur
give object

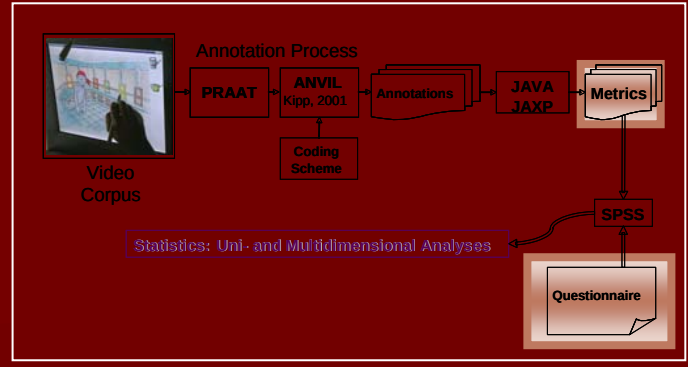
```
<track name="words" type="primary">
  ...
  <el index="93" start="282.81033" end="283.64114">
    <attribute name="token">bonjour</attribute>
    <attribute name="category">locution</attribute>
  </el>
  <el index="94" start="284.13635" end="284.40149">
    <attribute name="commandGroup">
      <value-link ref-track="command" ref-index="34" />
    </attribute>
    <attribute name="token">voilà</attribute>
    <attribute name="category">locution</attribute>
  </el>
  <el index="95" start="284.40149" end="284.83569">
    <attribute name="commandGroup">
      <value-link ref-track="command" ref-index="34" />
    </attribute>
    <attribute name="token">votre</attribute>
    <attribute name="category">adjective</attribute>
  </el>
  <el index="96" start="284.83569" end="285.39285">
    <attribute name="commandGroup">
      <value-link ref-track="command" ref-index="34" />
    </attribute>
    <attribute name="token">fleur</attribute>
    <attribute name="category">substantive</attribute>
  </el>
  ...
</track>

<track name="pen" type="primary">
  ...
  <el index="75" start="283.16" end="284.04">
    <attribute name="phase">preparation</attribute>
  </el>
  <el index="76" start="284.04" end="284.96">
    <attribute name="phase">stroke</attribute>
    <attribute name="shape">pointing</attribute>
  </el>
```



- Output file from ANVIL:
  - Results displayed track by track
  - Need of a program to extract metrics and connections between them

Variables collected



- Behavioral metrics:
  - Duration of use for each modality
  - Characteristics of use (syntactic categories of words, movement's shape …)
- Subjective data (questionnaire):
  - Easiness, pleasantness…

% of multimodal combinations ?

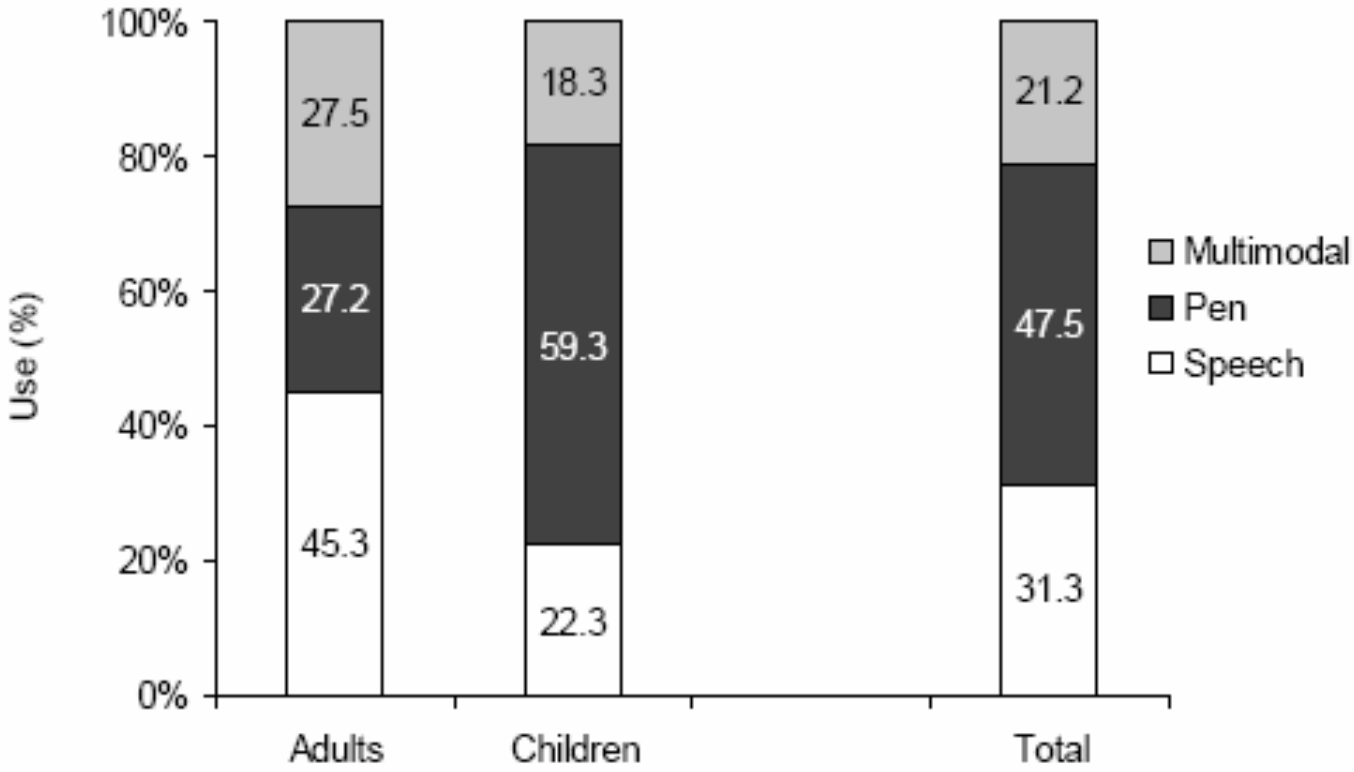
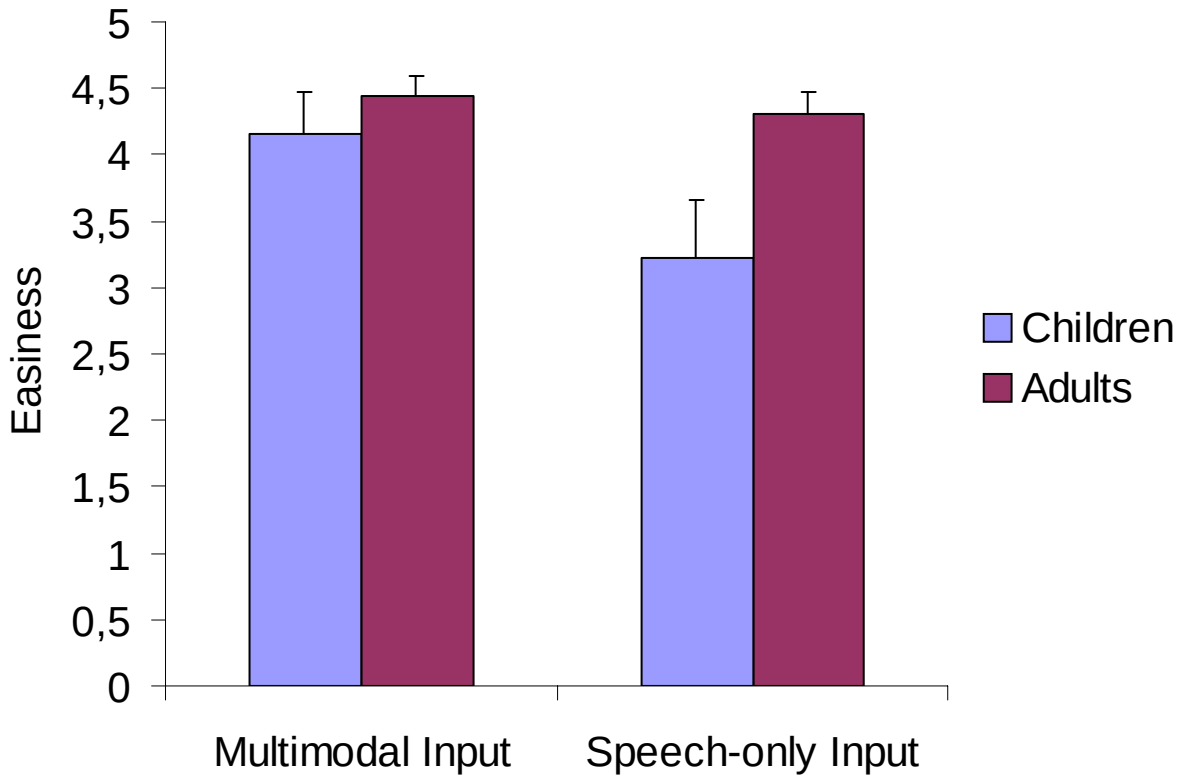


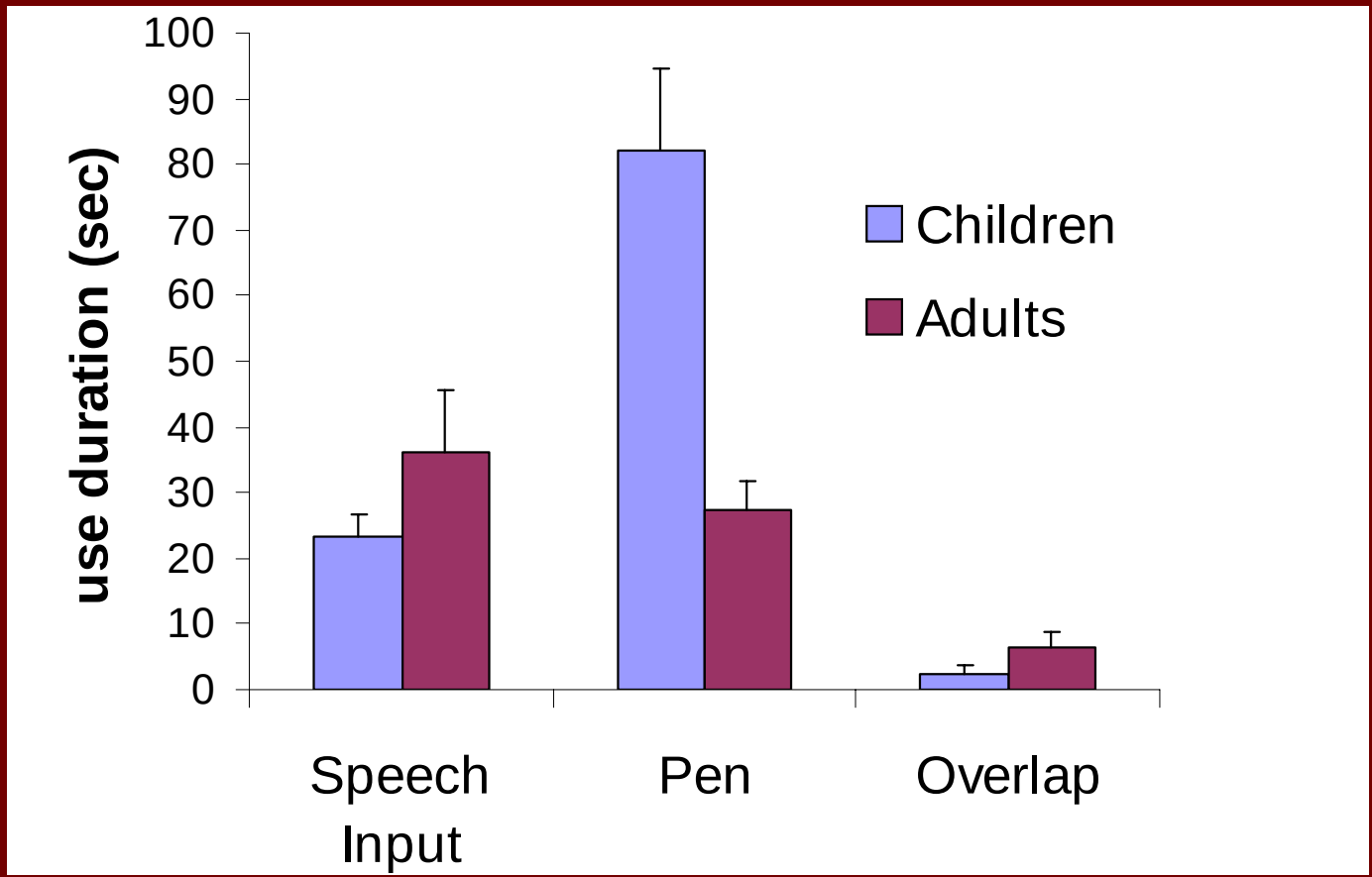
Figure 2. Use of modalities by adults and children, and for the whole sample.

Differences speech and multimodal ?

- Multimodal scenarios significantly shorter
  + yield higher and more homogeneous ratings of easiness

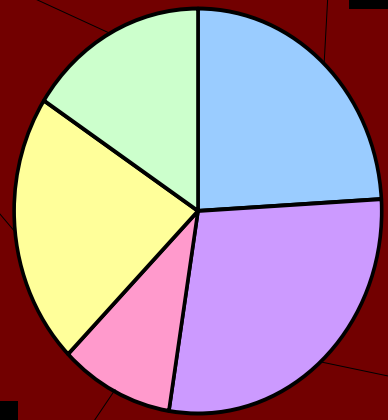


Differences children / adults ?



Most frequent combination of modalities ?

« Hello » + cake



« I want to go to the pink room »
+ pink door

« I take the red book »
- 3-sec. lag
+ red book

« Take this » + book

« Can I take a cake? »
- « Help yourself »
- + cake

- 91% of Human-Human syntax. « May I take the red book, please? »
- 2 social cues /min: Politeness, feedback from user.« Please », « thanks »…

Part I : Models of Multimodal Behavior

- Human-Computer Interaction
  - LEA 2D
  - NICE 3D
- Human-Human Interaction
  - EmoTV Humaine

The NICE project www.niceproject.com

- Multimodal interaction (speech + 2D gesture) with H.C. Andersen
  - Conversation about HCA's life and fairytales
  - Exploration of HCA's study, selection of objects
- 9 to 18 year-old users

Buisine, S., J.-C. Martin and N. O. Bernsen (2005). Children's Gesture and Speech in Conversation with 3D Characters. HCI International 2005, Las Vegas, USA.

Existing multimodal prototypes

- Task oriented / adults
  - Spatial applications (Oviatt 2003)
  - Srisis management (Sharma et al. 2003)
  - Bathroom design (Catizone et al. 2003),
  - Logistic planning (Johnston et al. 1997; Johnston 1998),
  - Tourist maps (Almeida et al. 2002; Johnston and Bangalore 2004)
  - Real estate (Oviatt 1997)
  - Graphic design (Milota 2004),
  - Intelligent rooms (Gieselmann and Denecke 2003; Juster and Roy 2004)
  - 3D interaction (Kaiser et al. 2003)
- Experimental studies
  - temporal patterns (Oviatt et al. 2003)

Existing multimodal prototypes

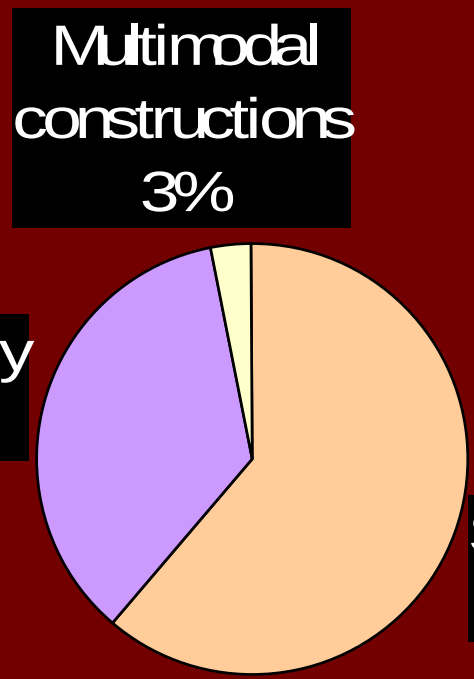
- Conversational adults
  - Mission Rehearsal system (Traum and Rickel 2002)
  - MAX agent for assembly task (Sowa et al. 2001)
- Conversational children
  - Pedagogical agent : speech reco simulated (Oviatt et al. 2004)
  - CHIMP (Narayanan et al. 1999)

NICE Prototype 2

User tests of Prototype 1

- 9 boys, 9 girls – 10 to 18 years
- Input modalities:
  - Speech (microphone headset, recognition simulated)
  - 2D Gesture (tactile screen or mouse)
  - User-controlled HCA navigation (arrow keys)
- ≈ 35-min conversation
  - 15 min free-style conversation
  - 20 min set of written scenario

Use of modalities observed in videos



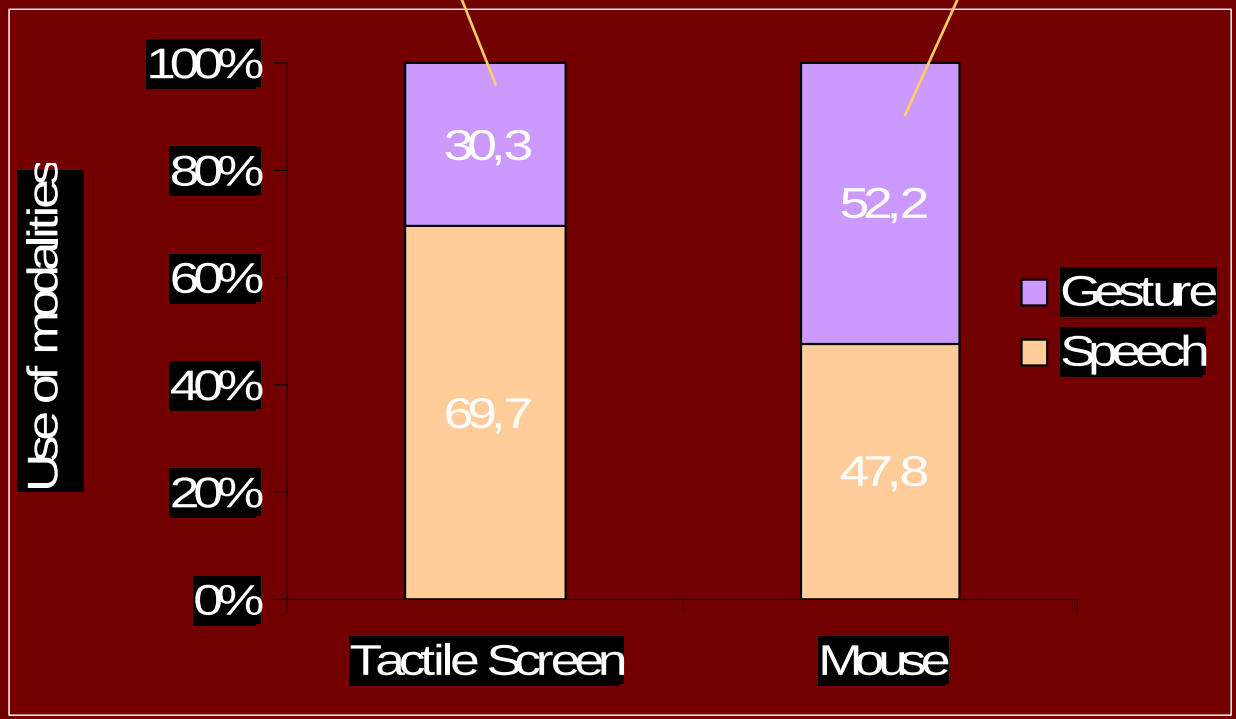
High rate due to the conversational nature of the game

Important for young users (transfer of behavioral patterns from computer games)
[Buisine & Martin 03]

Effect of gestural device from logs

Tactile screen:
More relevant for a conversational scenario

Mouse:
Used like a gaming device



Multimodal constructions from videos (PT1)

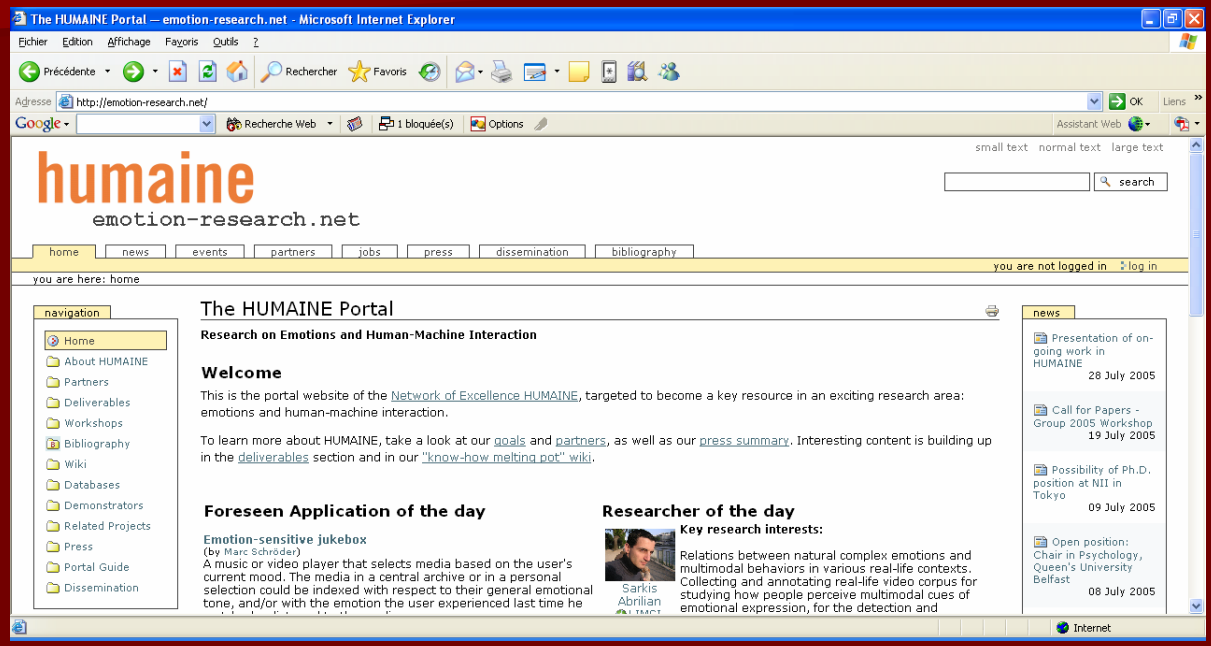
- **Temporal** integration of modalities
  - **Related**
    - 50% simultaneous (overlap between modalities)
    - 50% sequential (no overlap, gesture before speech)
  - **Concurrent (not related)**
    - « Oh I remember that one now » + gesture on picture
    - « How old are you? » + gesture on a vase
- **Semantic** integration of modalities
  - **Redundancy**
    - « I want to know something about your hat » + gesture on hat
  - **Complementarity**
    - Compatibility : « what is this? » + gesture on a picture
    - Conflict : « tell me about these two » + gesture on a statue

Part I : Models of Multimodal Behavior

- Human-Computer Interaction
  - LEA 2D
  - NICE 3D
- Human-Human Interaction
  - EmoTV Humaine (Martin, Devillers, Abrilian)
    - Annotation of emotion
    - Annotation of expressive multimodal behavior
    - Copy-synthesis approach

Humaine Network of Excellence

http://emotion-research.net



- Emotion and HCI
- 1st January 04 / duration : 48 months
- coordinator: Queen's University Belfast
- 33 partners from 11 countries

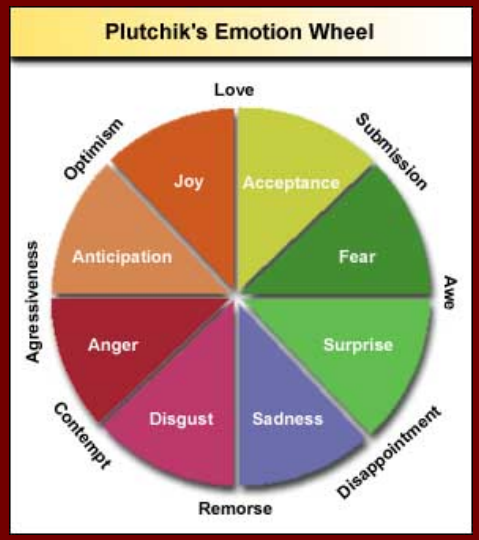
Goals

- Building a community working jointly on emotion-oriented systems
  - phase 1: establish common language and research directions
  - phase 2: exemplars to show "how to do things in a principled way"
  - not an IP! We are not building systems!

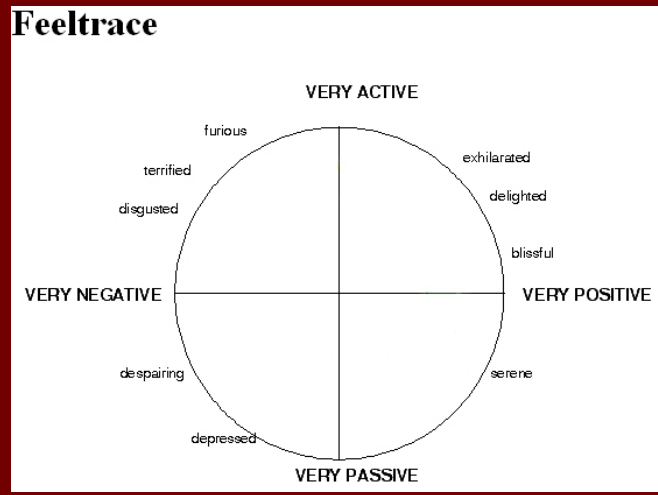
Humaine thematic areas / WPs

- theories and models of emotion
- signals to signs of emotion
- data and databases
- emotion in interaction
- emotion in cognition & action
- emotion in communication & persuasion
- usability of emotion-oriented systems
- ethics and good practice

Studying Emotions



- Emotional State
  - Verbal Categories (Ekman, Plutchick)
  - Abstract Dimensions (Osgood, Cowie)
- Processes
  - Appraisal Dimensions (Scherer)
  - OCC event/agent/object (Ortony)
  - Evaluation + coping (Gratch & Marsella)
- Mostly
  - Acted or induced
  - Monomodal



➡ Need for real life multimodal data

EmoTV

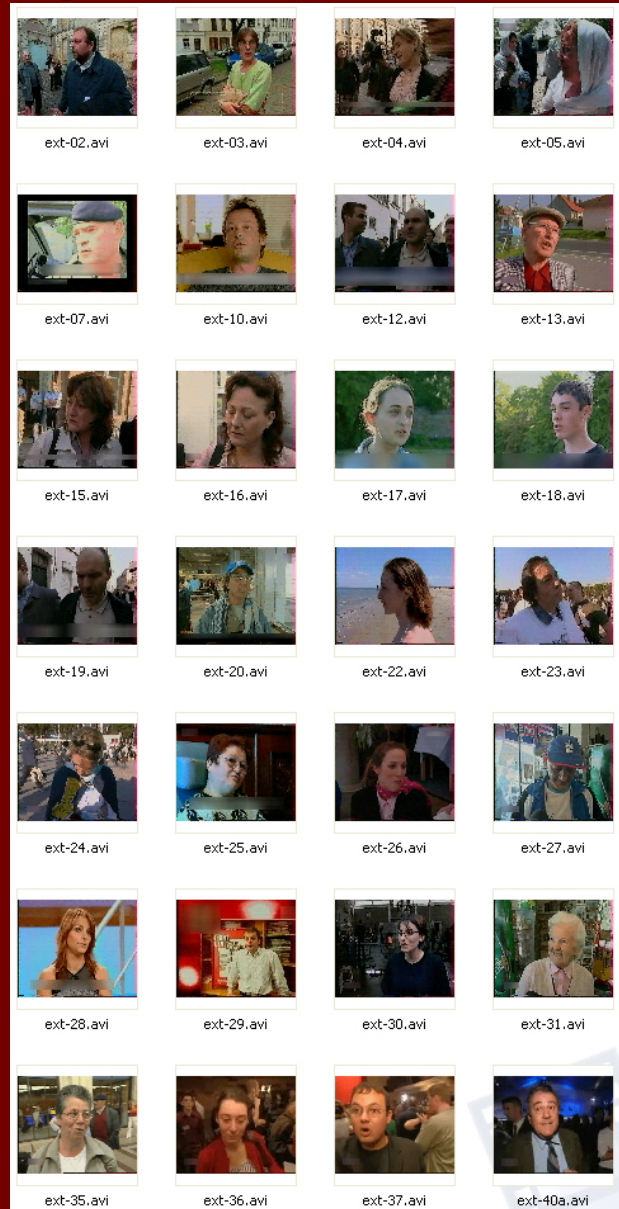
- Collection and annotation of Real-life Emotions

- Modeling of Multimodal Emotional Behaviors for the Specification of ECA (e.g. storytelling)

Corpus
Example

The EmoTV Corpus

- Selection criteria :
  - Emotion
  - Multimodal
  - Spontaneous
  - only one interviewed
  - « ordinary people » …
- 51 video clips from French TV Interviews (mostly news)
- 48 different people
- 24 topics : politics, sport, law…
- Length: 12 min (4 - 43 sec. / clip)
- Lexicon: 2500 (800 distinct words)
- Wide range of pos. & neg. emotions



Abrilian et al. 2005a & b.
HCII'2005.

Corpus
Advantages & Drawbacks

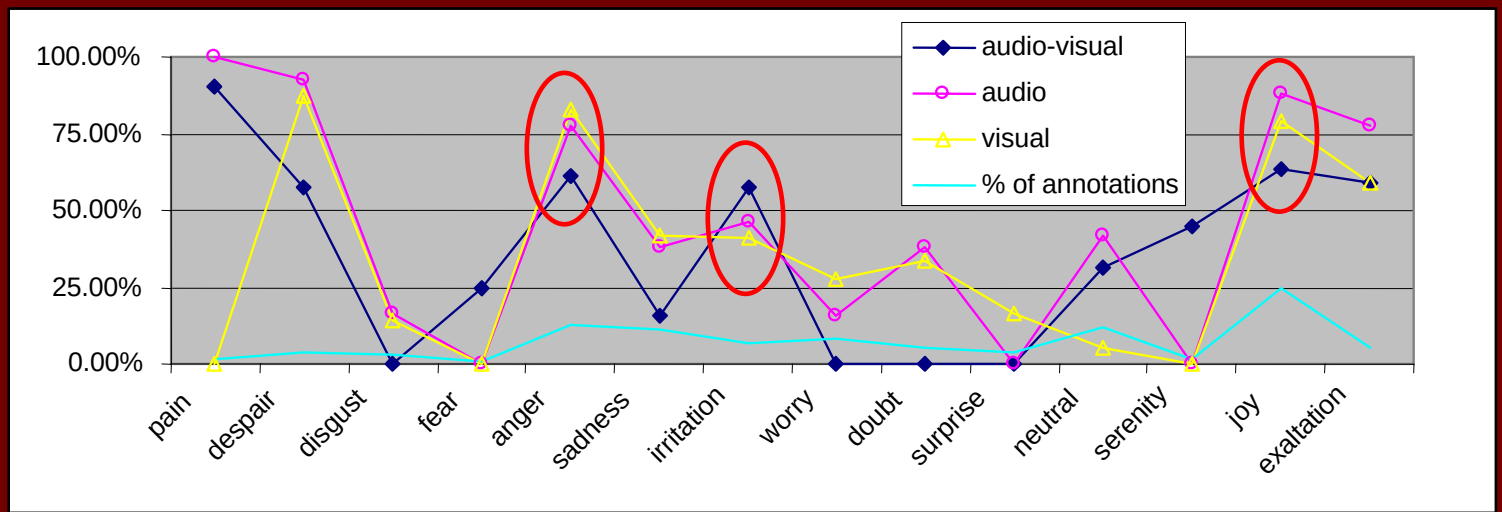
- Advantages
  - Spontaneous : non acted, not induced
  - Variety of contexts
  - Illustration of requirements on annotation schemes at several levels
- Drawbacks
  - Visibility of
    - Gestures
    - Facial Expressions: glasses / hairs / beard
  - Audio & Video quality

Exploratory corpus

Emotion Annotation Results

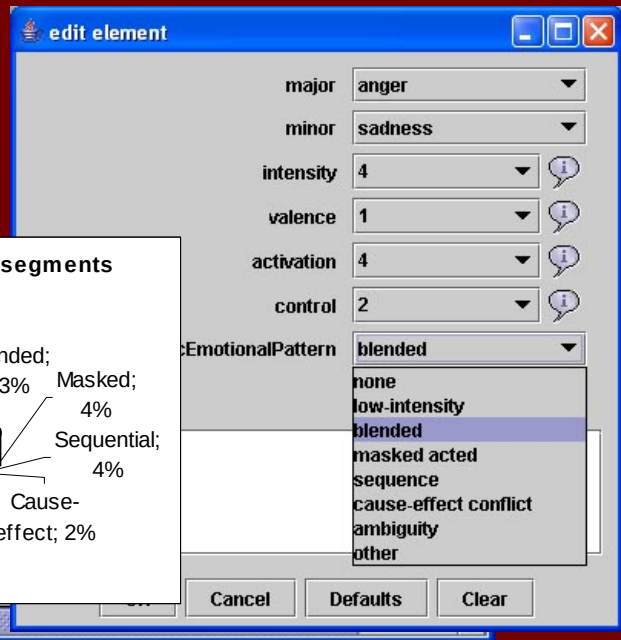
- 3 conditions, 2 coders
- Inter-Coder Agreement Measures
  - Douglas, Devillers, Martin, Cowie, Abrilian et al. Interspeech 2005

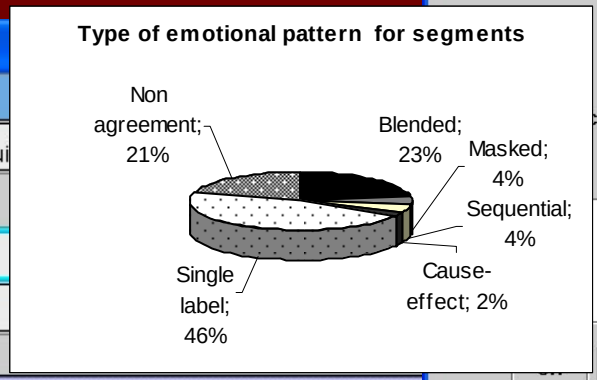
Condition	Kappa on Emotion Labels	Alpha on Emotion Intensity	Alpha on Emotion Valence
Audio only	0.540	0.865	0.915
Video only	0.430	0.510	0.709
Audio & Video	0.370	0.254	0.574



Emotion Annotation with 2 labels Ongoing work

- Modified coding scheme for the annotation of emotion
  - Permit to annotate combination of labels (major, minor)
  - Type of non basic emotional pattern
    - Masked, Blended, Ambiguity, Cause-effect conflict
- From 2 to 4 abstract dimensions
  - Activation: value between 1 (passive) & 5 (active)
  - Valence: between 1 (negative) & 5 (positive)
  - Intensity: between 1 (low) & 5 (high)
  - Control: between 1 (uncontrolled) & (controlled)





Studies on Multimodality and emotions

Expressivity Parameters

Acted

Basic Emotions

(Montepare et al. 1999)	- Hand positions - Gait - Fluidity - Stiffness - Strength - Speed - spatial expansion - Activity	Acted	No	All bodily activity except facial expression	Portray emotional situations	82 younger and older adults decoding emotion in brief video extracts	- Happy - Sad - Angry - Neutral
(Wallbott 1998)	- Upper body - Shoulders (up, backward, forward) - Head (downward, backward, turned sideways, bent sideways) - Arms - Hands - Movement quality (activity, spatial expansion, movement dynamics, energy, power) - Symmetry	Acted	Yes	All bodily activity	Non vocal movement and postural activity for each emotion	12 drama students and trained observers	from Scherer's list (1986) - Joy - Sadness - Pride - Shame - Fear/terror/horror - Anger/rage - Disgust - Contempt
(Boone and Cunningham 1996; Boone and Cunningham 1998)	- Changes in tempo - Directional changes - Frequency - Muscle tension - Duration	Acted	Yes	Face, Torso, Facial Muscles, Body Leaning.	Dance performances on for basic emotions	103 subjects: 79 children (4 to 8 year old), 24 adults.	- Anger - Happiness - Fear - Sadness
(DeMeijer 1991)	- Trunk (stretching, bowing) - Arm (opening, closing) - Vertical direction (upward, downward) - Sagittal direction (forward, backward) - Force (strong, light) - Velocity (fast, slow) - Directness	Acted	No	Trunk, Arms.	Expressiveness of aggression and grief in 16 distinct movement expressions. 7 dichotomous dimensions describing movements.	42 adults (21 male and 21 female).	- Aggression - Grief

Studies on Multimodality and emotions

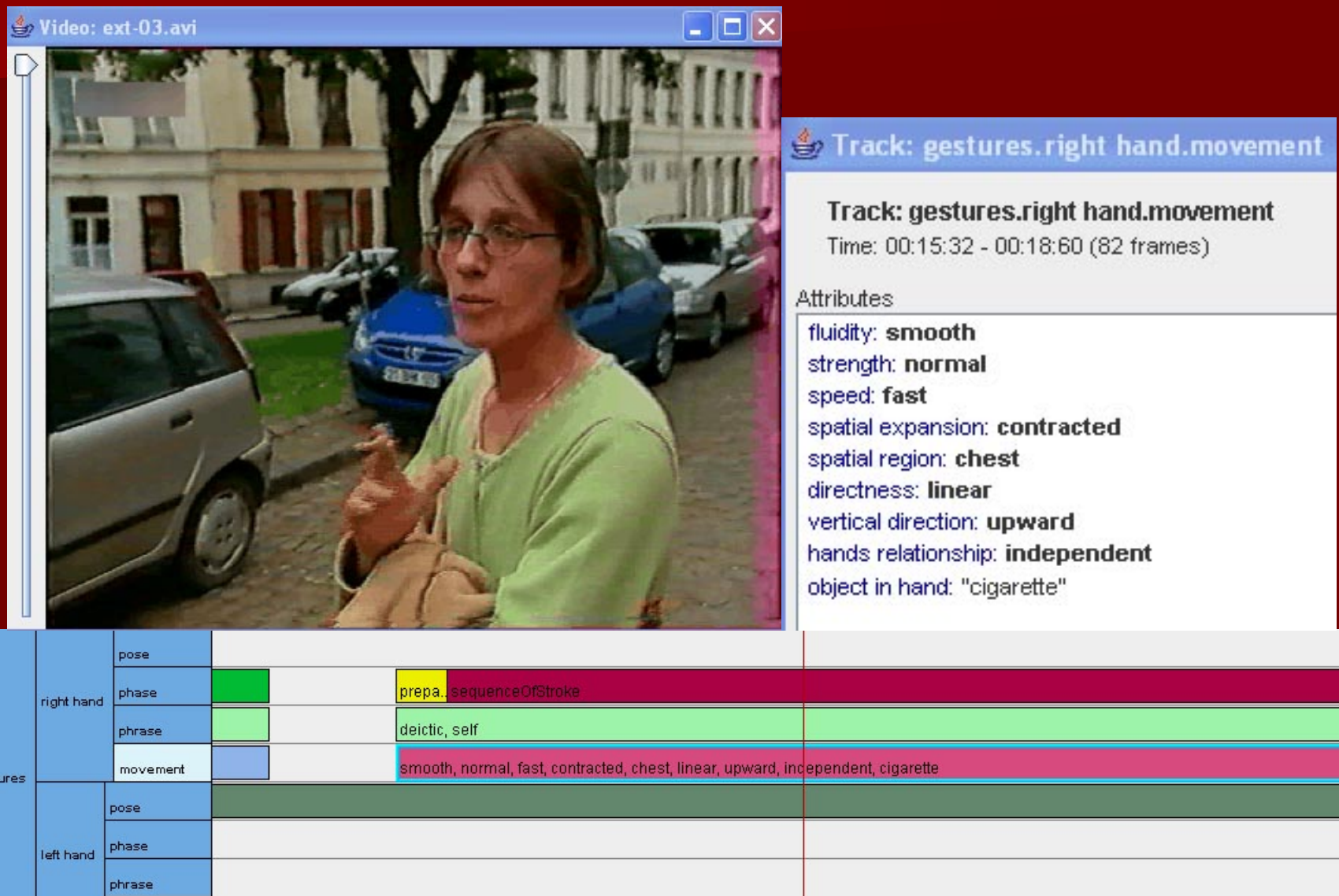
- Annotation of multimodal behavior in TV videos
  - (Allwood et al. 2004 ; Kipp 2004)
  - Focus on communicative functions rather than emotion

- Expressive ECAs (Pelachaud et al. 2004)
  - Overall activation
  - Spatial extent
  - Temporal extent
  - Fluidity
  - Power
  - Repetitivity

Hartmann, B., M. Mancini and C. Pelachaud (2005). Implementing Expressive Gesture Synthesis for Embodied Conversational Agents. Gesture Workshop, Vannes.

Example of Multimodal Annotation

Gesture: repetitive expressive deictic



Conclusions

- Similarity of most frequently annotated behaviors (% of total number of annotations)
- Observed disagreement between coders
  - Coder1 : 1839 - coder 2 : 914
  - Correlation speed and energy
  - Naturalistic Data ≠ Acted Data
    - Contradictory channels
    - Subtle / controlled / masked
  - Parts of coding scheme inapropriate
    - Face : use AU instead of FAPS
    - Number of values for speed : 5 -> 3
  - Improvement of annotation protocole
    - Expressivity annotated relatively to the video
    - Graphical representation in annotation guide

Future work

- 2nd annotation phase and scheme: emotion + mm
- Validation / reliable information
  - Trade-off quantity information / time / agreement
  - Study redundancy / complementarity of different annotations (build links between dimensions and labels)
  - Copy-synthesis with an ECA
- Representations
  - Emotion Representation Language
  - Expressivity profile

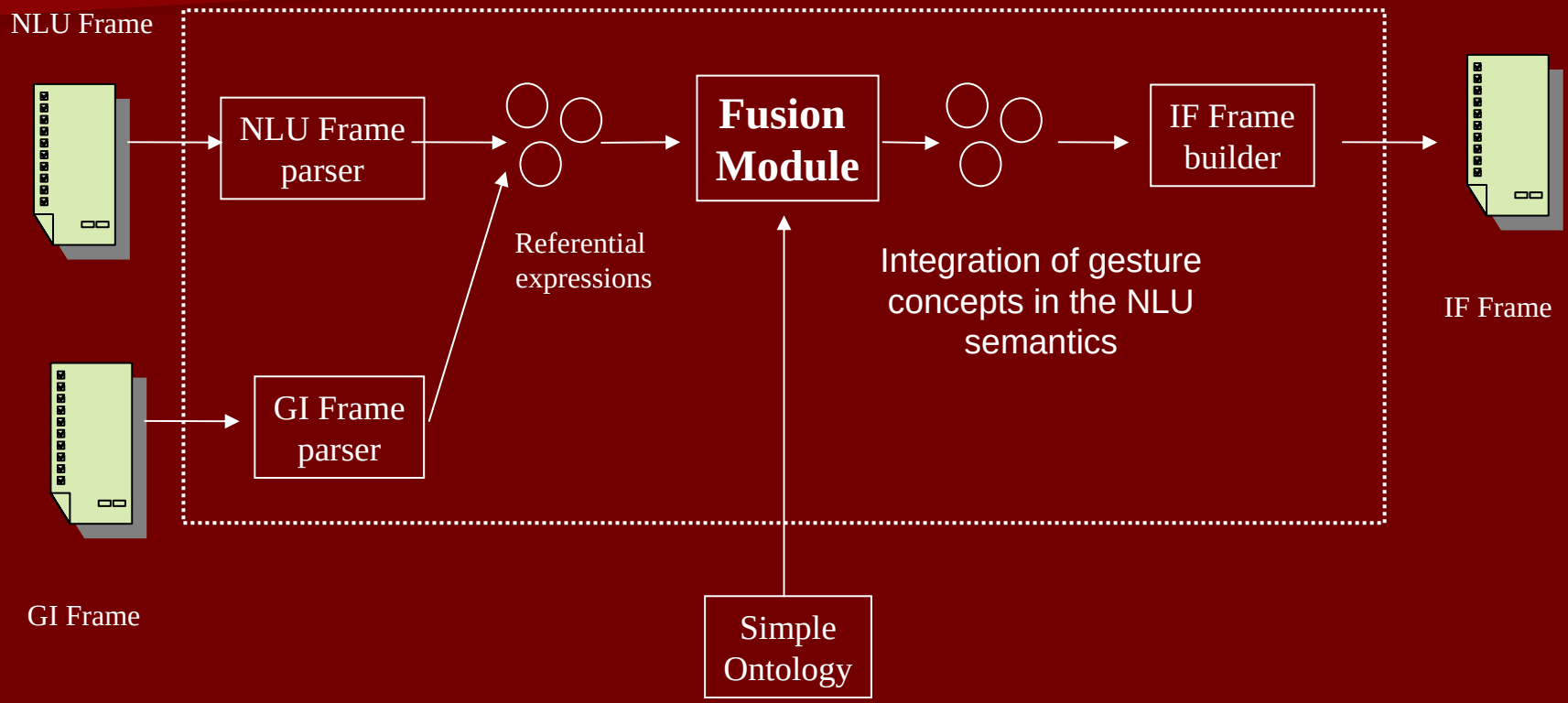
Expressivity profiles of three videos

Video#	#3	#36	#30
Duration	37s	7s	10s
Emotion labels	Anger (66%), despair (25%), sadness (12%)	Anger (55%), despair (44%)	Exaltation (50%), Joy (25%), Pride (25%)
Average intensity (1: min – 5: max)	5	4.6	4
Average valence (1: negative, 5: positive)	1	1.6	4.3
% head movement	56	60	72
% torso movement	28	20	27
% hand movement	16	20	0
% fast vs. slow	47 vs. 3	33 vs. 13	83 vs. 0
% hard vs. soft	17 vs. 17	20 vs. 0	0 vs. 27
% jerky vs. smooth	19 vs. 8	6 vs. 0	5 vs. 50
% expanded vs. contracted	0 vs. 38	13 vs. 20	0 vs. 33

Outline

- Introduction
- Part I : Models of Multimodal Behavior
- Part II : Input Fusion
- Part III : Embodied Conversational Agents
- Future Directions

NICE
Input Fusion Internal Structure

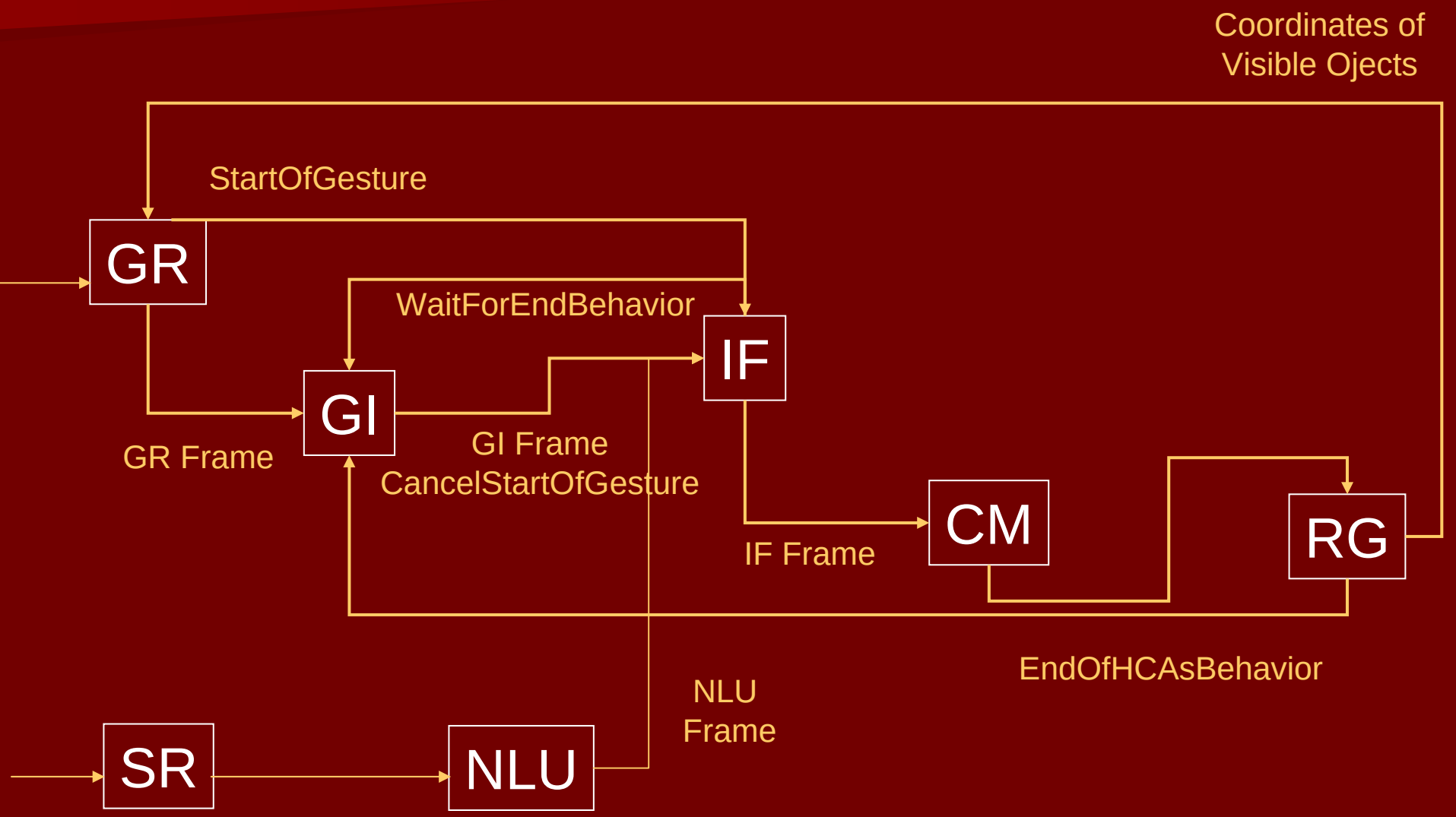


Fusion via 3 dimensions

Input Fusion Algorithm
1) Temporal Dimension

- Timers
  - Speech-waiting-for-gesture
    - short-delay
    - long-delay (following a StartOfGesture)
  - Gesture-waiting-for-speech
    - short-delay
    - long-delay (following a StartOfSpeech)
- Events
  - a new NLU / GI frame is received by the IF
  - a StartOfSpeech / StartOfGesture is received by the IF
  - a Speech-waiting-for-gesture / Gesture-waiting-for-speech is over

Input Fusion Algorithm
1) Temporal Dimension



Input Fusion Algorithm
2) Singular / Plural Dimension

GI NLU	No message from GI	1 message from GI but "noObject"	1 object detected by GI "select"	Several objects detected by GI "referenceAmbiguity"
No message from NLU	1	2	3	4
1 message from NLU but no explicit reference in NLU frame	5	6	7	8
1 message from NLU with 1 singular reference	9	10	**11**	**12**
1 message from NLU with 1 plural reference	13	14	**15**	**16**

Input Fusion Algorithm
3) Semantic / Perceptual dimension

- Task analysis test bed (233 multimodal combinations)
- Verbal reference
  - Deictic: has higher priority
  - Object property: « Jenny Lind ? »
  - Implicit reference:
    - Do you like travelling + <travelBag>
- Simple ontology
  - NLU concepts: singular / plural / both
  - GI concepts: affordance: singular / plural / both
  - Relations
- Parsing NLU semantics
  - Concepts : objectinstudy, fairytale, fairytalecharacter, family, other_works, friends, country, location
  - Property: travel, hobby, mary
  - Number

Input Fusion – HCA Study Examples

```
<TRACE text=" startOfGesture" date="2005-02-15 15:20:04.312"/>
<TRACE text=" SENT BY GI TO IF" date="2005-02-15 15:20:05.828"/>
<?xml version="1.0" encoding="UTF-8"?>
<giFrame          endOfDetectionPeriod="2005-02-15 15:20:04.312"
                  startOfDetectionPeriod="2005-02-15 15:20:04.312">
   <select>
      <name>pictureJonasCollin</name>
   </select>
</giFrame>
```

```
<TRACE text=" SENT BY NLU TO IF" date="2005-02-15 15:20:06.843"/>
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE nluframe SYSTEM "nlu_nice.dtd">
<nluframe>
   <date>2005-2-15 15:20:6</date>
   <speech_input> how was this woman </speech_input>
   <semantic cs="368">
      <number_of_domain val="1">
         <domain val="no_value">
            <number_of_concept val="1">
               <concept val="no_value">
                  <number_of_subconcept val="1">
                     <subconcept val="no_value"/>
                  </number_of_subconcept>
               </concept>
            </number_of_concept>
         </domain>
      </number_of_domain>
...
```

```
 …
<number_of_property val="2">
        <property val="diectic">
          <number_of_property_type val="1">
            <property_type val=“ this "/>
          </number_of_property_type>
        </property>
        <property val="gender">
          <number_of_property_type val="1">
            <property_type val="woman"/>
          </number_of_property_type>
        </property>
      </number_of_property>
      <number_of_dialogue_act val="1">
        <dialogue_act val="question">
          <number_of_dialogue_act_type val="1">
            <dialogue_act_type val="general"/>
          </number_of_dialogue_act_type>
        </dialogue_act>
      </number_of_dialogue_act>
      <number_of_dialogue_act_subject val="1">
        <dialogue_act_subject val="no_value"/>
      </number_of_dialogue_act_subject>
    </semantic>
</nluframe>
```

```
<ifFrame fusionStatus="ok">
   <nluframe>
      <date>2005-2-15 15:20:6</date>
      <speech_input> how was this woman </speech_input>
      <semantic cs="368">
         <number_of_domain val="1">
            <domain val="physicalpresence">
               <number_of_concept val="1">
                  <concept val="objectinstudy">
                     <number_of_subconcept val="1">
                        <subconcept val=“ pictureJonasCollin "/>
                     </number_of_subconcept>
                  </concept>
               </number_of_concept>
            </domain>
         </number_of_domain>
         <number_of_property val="2">
            <property val="diectic">
               <number_of_property_type val="1">
                  <property_type val="this"/>
               </number_of_property_type>
            </property>
            …
   </nluframe>
</ifFrame>
```

Reasons of failure for the processing of multimodal behaviors

- Estimation of 7% of multimodal turns
- 60% of interaction success for mm turns
- IF sucessful fusion in 25% of the mm cases
- Reasons of failure for the processing of multimodal behaviors :

	%
Timer Too Small	43
Speech Recognition Error	18
Input Inhibited	12
Not A Referenceable Object	8
Gesture Not Detected	8
System Crash	4
Unexplained Reason	4
Gestured Object Not Detected	2
TOTAL	100

Outline

- Introduction
- Part I : Models of Multimodal Behavior
- Part II : Input Fusion
- Part III : Embodied Conversational Agents
- Future Directions

Part III : Embodied Conversational Agents

- Emo-TV / Greta
- The LEA 2D tool
- Evaluation of ECAs' multimodal strategies

Embodied Conversational Agent A definition

- HCI
  - With which the user can interact in an intuitive way
  - Which uses verbal and non verbal signs of human communication for various communicative functions

Cassell, J., J. Sullivan, S. Prevost and E. Churchill (2000). Embodied Conversational Agents, MIT Press.

Goals

- Endow HCI with the
  - richness
  - subtleties
  - intuitiveness

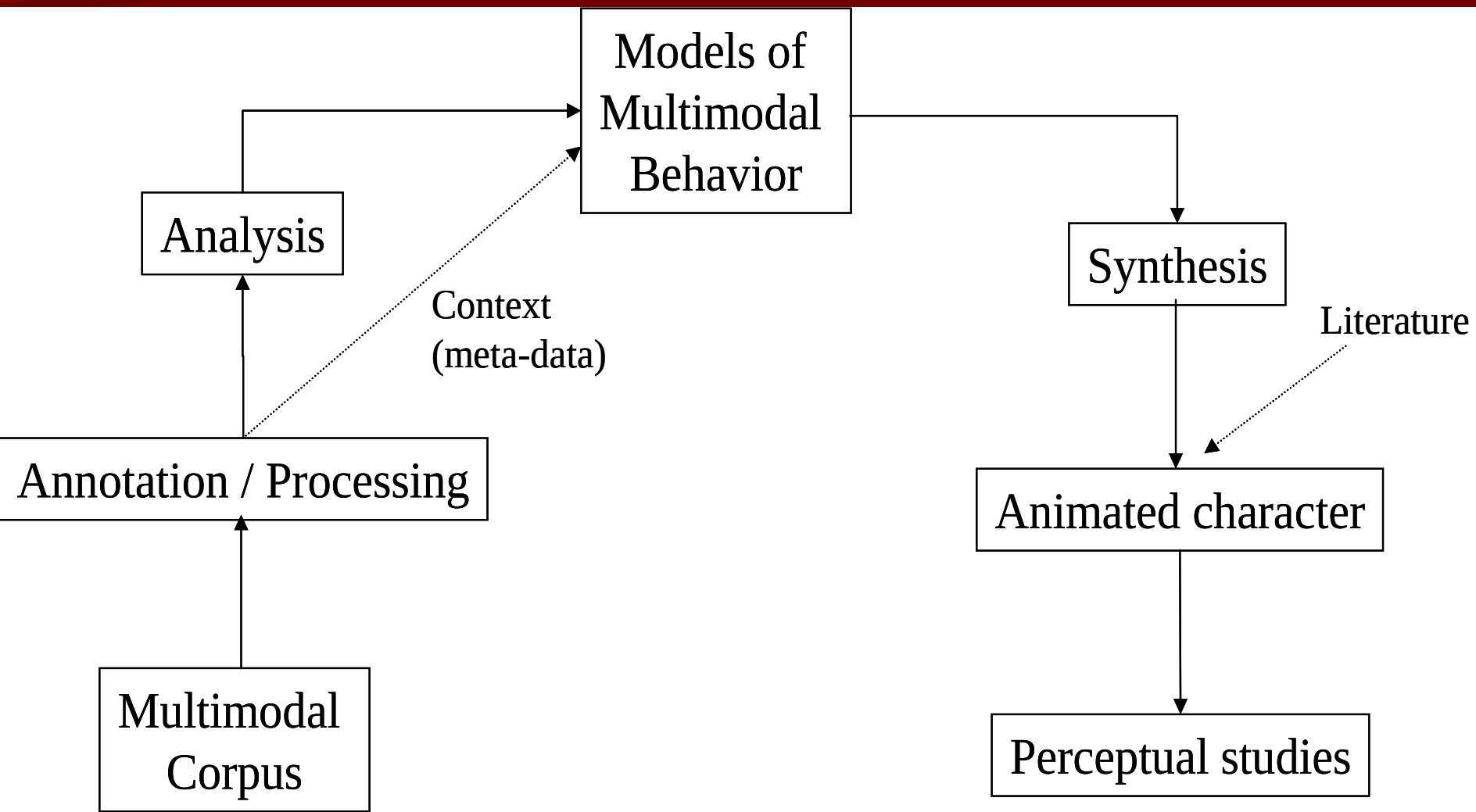
of human communication

- Evaluation

Dehn, D. M. and S. van Mulken (2000)

Ruttkay, Z. and C. Pelachaud (2004)

Emotional behavior Annotation – synthesis



EmoTV - Greta

- Separate specifications
  - Facial expressions
  - Expressivity of gesture
  - Real speech
- Basic emotions
  - Anger, despair
  - From literature
- Mixed emotion
  - From EmoTV annotation
- Goal of experiment
  - Do subject perceive a combination of emotion

From annotation to specification basic vs. combined emotion

Outline

- Introduction
- Part I : Models of Multimodal Behavior
- Part II : Input Fusion
- Part III : Embodied Conversational Agents
- Future Directions

On-going work and future directions

- Mutiple sources for models of multimodal behavior
  - Literature, acting theories
  - Corpus
  - Motion capture
  - Image processing
  - Manual annotation
  - Other copora : weather forecast

On-going work and future directions

- Bidirectional multimodality
  - Standards : W3C EMMA
  - Storytelling application
  - MM HCI for autistic people
  - Corpus of 3D bidirectionnal interactions
  - Training / intuitive
  - Plural / singular dimension

Acknowledgement

- Humaine NoE IST-2002-2.3.1.6
- NICE project IST-2001-35293
- PhD students and post-docs
  - S. Buisine, S. Abrilian, G. Pitel, O. Grynszpan
- Colleagues
  - L. Devillers, C. Pelachaud
  - NICE and Humaine partners