Language Technology II: Language-Based Interaction

Multimodal Dialogue Systems

Ivana Kruijff-Korbayová
korbay@coli.uni-sb.de
www.coli.uni-saarland.de/courses/late2/

I have reused some slides from presentations of W. Wahlster, M. Johnston and J. Cassell

Outline

- Modes of Interaction
- Embodied Conversational Agents
- Cross-modal Interaction: Fusion and Fission
- Example 1: MATCH
- Example 2: SMARTKO M

Input Modalities

- Natural Language:
  - Text and Speech
- Haptic:
  - Buttons, Joystick, MouseClick
- Graphics:
  - Sketching, Highlighting
- Gesture:
  - Pointing at a region of display, pointing at or manipulating objects in a visual scene (using full visual recognition/data-glove/augmentd reality)
- Mimics:
  - Eye gaze, lip movement

(Wahlster, 2004)

Output Modalities

- Natural Language:
  - Text and Speech
- Menus, tables
- Sounds
- Graphics, Animation
- Pictures, Videos
- Further Modalities (Gesture, Mimics) coming with embodied conversational agents

(Wahlster, 2004)

Multimedia - Multimodal

- Basic distinction between
  - Medium: physical carrier of information
  - Mode: particular sign system
- Examples:
  - Circling objects on a map by visually processed gesture vs. data-glove vs. pen: multimedia + monomodal,
  - Speech plus pointing gesture: multimedia + multimodal
  - Speech vs. Text: mono/multimodal?

(Wahlster, 2004)

Types and Function of Multimodality

- Choice between alternate modalities for (monomodal) turn realisation: Adaptation to the needs of situation
- Simultaneous realisation of (system) turns in parallel modalities, e.g., Speech + Displayed Table: User-friendly redundancy
- Mixed or composite modality in a single (user) turn ("cross-modal dialogue"): User can select best suited mode for certain kind of content
  - Manfred Pinkal's phone number is *3024343* (typed)
  - Zoom in *here* (+ Ink or Gesture)
- Concomitant modalities (mimics, gesture): Support recognition/understanding of spoken utterance

(Wahlster, 2004)

Posture Shifts mark the beginning of new discourse segments (Cassell et al., '01)

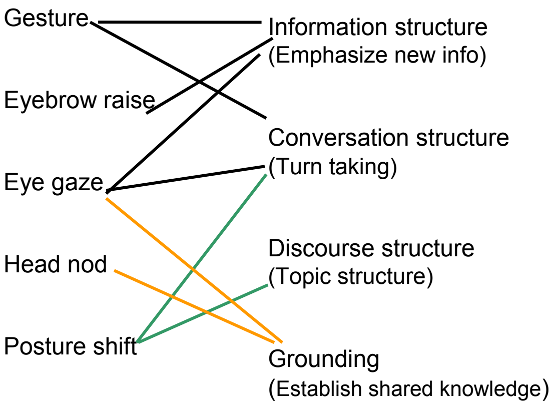
Looks towards the listener indicate that further grounding is needed (Nakano, et al. '02)

Gestures are more likely to occur with rhematic material than thematic material (Cassell et al. '94)

Small talk occurs before face-threatening discourse moves (Bickmore & Cassell, '02)

(Cassell, 2005)

Relationship between Linguistic Structure & Behavioral Cues



Gesture

Eyebrow raise

Eye gaze

Head nod

Posture shift

Information structure
(Emphasize new info)

Conversation structure
(Turn taking)

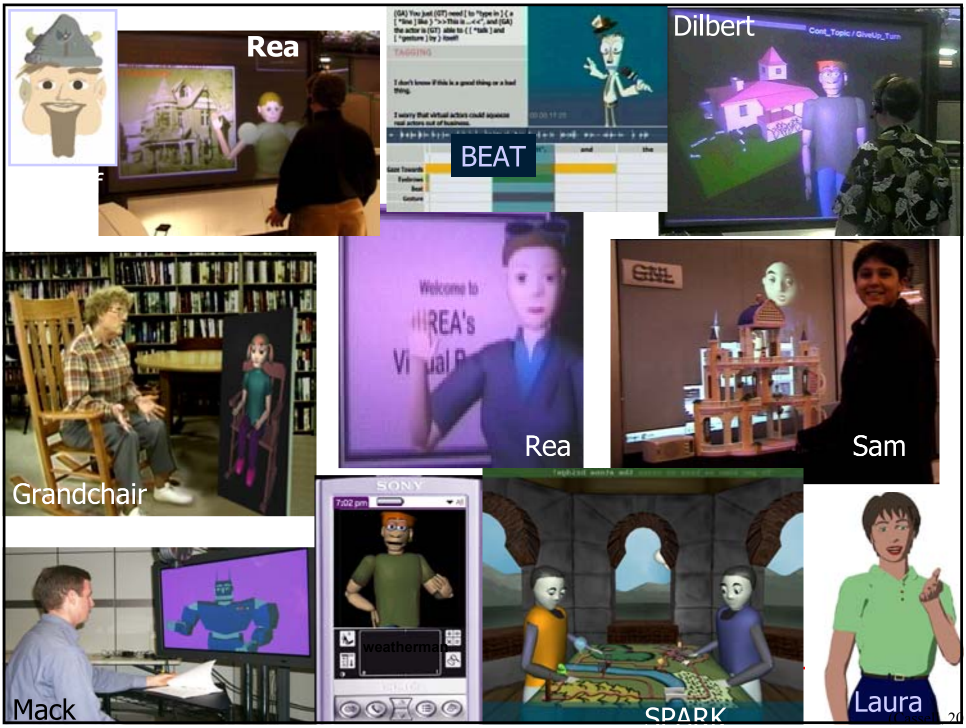
Discourse structure
(Topic structure)

Grounding
(Establish shared knowledge)

(Cassell, 2005)

Anthropomorphic Interfaces

- Interfaces which have a "persona", i.e. at least a face or a whole body
  often also called Embodied Conversational Agents (ECA)
  - Talking heads
  - Virtual animated characters
- Added aspects of social interaction



(Cassell, 20

Composite Multimodality

Coexistence of input and output in different media and modes ≠ ***Effective user interface***

- From alternate modes of interaction to composite multimodality
- Careful coordination of different media and modes in a coherent and cooperative dialogue is required

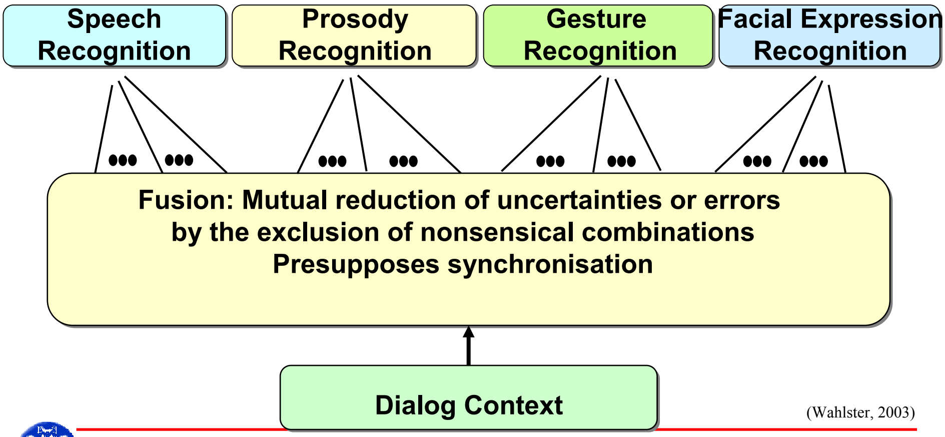
Composite Multimodality: Input

- Composite input:
  - Enabling users to provide a single contribution (turn) which is optimally distributed over the available input modes
    e.g., speech + ink "zoom in here"
- Motivation
  - Naturalness
  - Certain kinds of content within a single communicative act are best suited to particular modes, e.g.,
    - Speech for complex queries or constraints, reference to objects currently not visible or intangible
    - Ink/gesture for selection, indicating complex graphical features

(Johnston, 2004)

Composite Multimodality: Input Fusion

Mutual disambiguation and synergistic combinations: semantic fusion of multiple modalities in dialog context helps to reduce ambiguity and errors



Speech Recognition

Prosody Recognition

Gesture Recognition

Facial Expression Recognition

Fusion: Mutual reduction of uncertainties or errors by the exclusion of nonsensical combinations
Presupposes synchronisation

Dialog Context

(Wahlster, 2003)

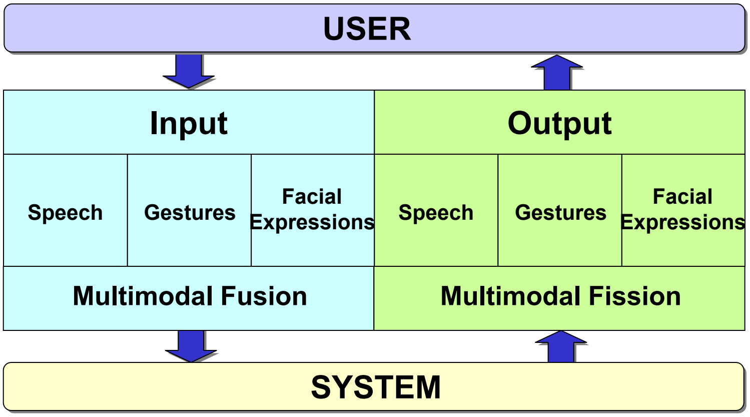
Composite Multimodality: Output

- Composite output:
  - Allowing for system output to be optimally distributed over the available output modes, e.g.,
    - High level summary in speech, details in graphics: “Take this route across town to the Cloister Café”
    - Multimodal help providing examples for the user: “To get the phone number for a restaurant, circle one like *this* and say or write *phone*.”
(Hastie et al. 2002)
  - Output should be dynamically tailored to be maximally effective given the situation and user preferences
- Same motivation as for multimodal input

(Johnston, 2004)

Full Symmetric Multimodality

Symmetric multimodality means that all input modes (speech, gesture, facial expression) are also available for output, and vice versa.



USER

Input: Speech, Gestures, Facial Expressions

Multimodal Fusion

Output: Speech, Gestures, Facial Expressions

Multimodal Fission

SYSTEM

The modality fission component provides the inverse functionality of the modality fusion component.

Challenge: A dialogue system with symmetric multimodality must not only understand and represent the user's multimodal input, but also its own.

(Wahlster, 2003)

Multimodal Understanding

- Associate word sequence + gesture sequence with meaning
  - Early integration: compute meaning of a composite word+gesture sequence: MMFST (Johnston&Bangalore 2002,2004)
  - Late integration: first compute meaning of word sequence and meaning of gesture sequence, then "merge" the meanings, e.g., (Pfleger 2002)

MATCH: Multimodal Access to City Help

- Interactive city guide and navigation for information-rich urban environments
  - Finding restaurants and points of interest, getting info, subway routes for New York and Washington, D.C.
- Composite input and output
  - Speech, ink, graphics
- Mobile (standalone on a PDA or distributed WLAN)
- MATCHKiosk (deployed at AT&T visitor center in DC)
  - Social interaction
  - Also printed output

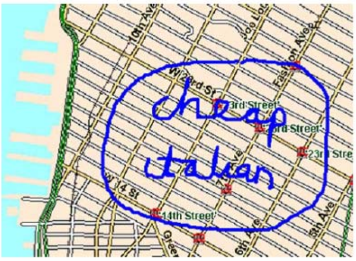
(Johnston, 2004)

MATCH

(Johnston, 2004)

MATCH

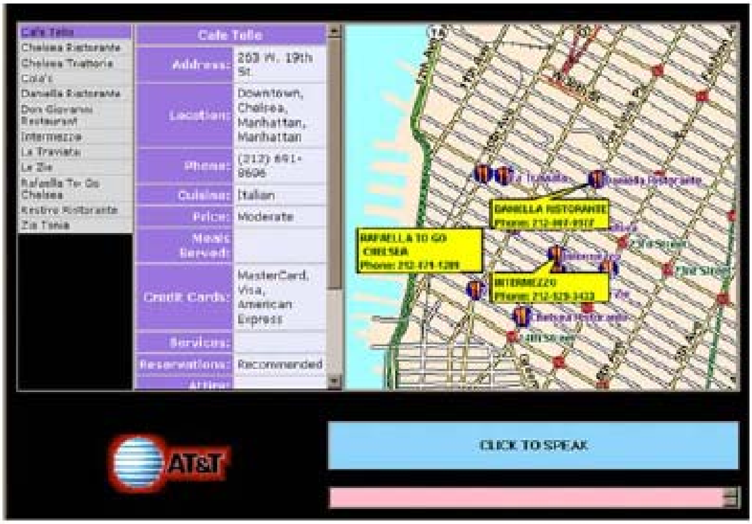
- Finding restaurants
  - Speech: “show inexpensive italian places in chelsea”
  - Multimodal: “cheap italian places in this area”
  - Pen:



  - Getting info: “phone numbers for these”
  - Subway routes: “how do I get here from Broadway and 95th street”
  - Pen/zoom map: “Zoom in here”

(Johnston, 2004)

MATCH



Cafe Tallo
Chelsea Ristorante
Chelsea Trattoria
Cola's
Daniella Ristorante
Don Giovanni Restaurant
Intermezzo
La Traviata
Le Zie
Rafaella To Go Chelsea
Kestre Ristorante
Zio Tonio

Cafe Tallo	
Address:	253 W. 19th St
Location:	Downtown, Chelsea, Manhattan, Manhattan
Phone:	(212) 691-8696
Cuisine:	Italian
Price:	Moderate
Meals Served:	
Credit Cards:	MasterCard, Visa, American Express
Services:	
Reservations:	Recommended
Attire:	

04)

User-Tailored Generation

- User-tailored summaries, comparisons or recommendations can be generated using a model of user preferences

"compare these restaurants"

Compare-A: *Among the selected restaurants, the following offer exceptional overall value. Uguale's price is 33$. It has excellent food quality and good decor. Da Andrea's price is 28$. It has very good food quality an good decor. John's Pizzeria's price is 20$. It has very good food quality and mediocre decor.*

Compare-B: *Among the selected restaurants, the following offer exceptional overall value. Babbo's price is 60$. It has superb food quality. Il Mulino's price is 65$. It has superb food quality. Uguale's price is 33$. It has excellent food.*

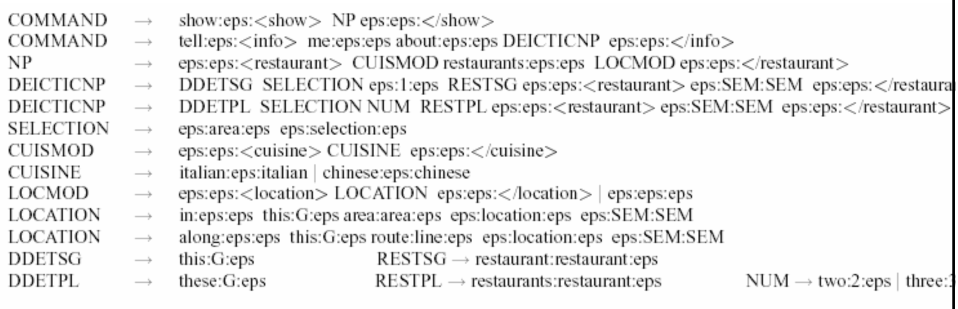
Johnston et al. (2004)

MATCH: Early Multimodal Integration

- Speech and gesture parsing, multimodal integration, and understanding in single MM grammar model
  - (Johnston&Bangalore 2000,2004)
  - Compiled from a declarative multimodal CFG (terminals are triples W:G:M = Words:Gestures:Meaning)
  - Compiled to efficient finite state device
    - G:W transducer aligns speech and ink
    - G_W:M transducer takes a composite alphabet of speech and gesture symbols and outputs meaning
- Robust, efficient
- Enables compensation for errors

(Johnston, 2004)

MATCH MM Grammar Fragment



```
COMMAND      →   show:eps:<show>  NP eps:eps:</show>
COMMAND      →   tell:eps:<info>  me:eps:eps about:eps:eps DEICTICNP  eps:eps:</info>
NP           →   eps:eps:<restaurant>  CUISMOD restaurants:eps:eps  LOCMOD eps:eps:</restaurant>
DEICTICNP    →   DDETSG  SELECTION eps:1:eps  RESTSG eps:eps:<restaurant> eps:SEM:SEM  eps:eps:</restaura
DEICTICNP    →   DDETPL  SELECTION NUM  RESTPL eps:eps:<restaurant> eps:SEM:SEM  eps:eps:</restaurant>
SELECTION    →   eps:area:eps  eps:selection:eps
CUISMOD      →   eps:eps:<cuisine> CUISINE  eps:eps:</cuisine>
CUISINE      →   italian:eps:italian | chinese:eps:chinese
LOCMOD       →   eps:eps:<location> LOCATION  eps:eps:</location> | eps:eps:eps
LOCATION     →   in:eps:eps  this:G:eps area:area:eps  eps:location:eps  eps:SEM:SEM
LOCATION     →   along:eps:eps  this:G:eps route:line:eps  eps:location:eps  eps:SEM:SEM
DDETSG       →   this:G:eps                RESTSG → restaurant:restaurant:eps
DDETPL       →   these:G:eps               RESTPL → restaurants:restaurant:eps          NUM → two:2:eps | three:
```

Fig. 6. Multimodal grammar fragment

A Fragment of the Fragment

COMMAND → tell:ε:<info> me:ε:ε about:ε:ε
DEICTICNP ε:ε:</info>

DEICTICNP → DDETSG SELECTION ε:1:ε RESTSG
ε:ε:<restaurant> ε:SEM:SEM ε:ε:</restaurant>

DDETSG → this:G:ε

SELECTION → ε:area:ε ε:selection:ε

RESTSG → restaurant:restaurant: ε

Semantic Information

COMMAND → tell:ε:<info> me:ε:ε about:ε:ε
DEICTICNP ε:ε:</info>

DEICTICNP → DDETSG SELECTION ε:1:ε RESTSG
ε:ε:<restaurant> ε:SEM:SEM ε:ε: </restaurant>

DDETSG → this:G:ε

SELECTION → ε:area:ε ε:selection:ε

RESTSG → restaurant:restaurant: ε

Semantic Information

COMMAND → tell:ε:<info> me:ε:ε about:ε:ε
DEICTICNP ε:ε:</info>

DEICTICNP → DDETSG SELECTION ε:1:ε RESTSG
ε:ε:<restaurant> ε:SEM:SEM ε:ε: </restaurant>

DDETSG → this:G:ε

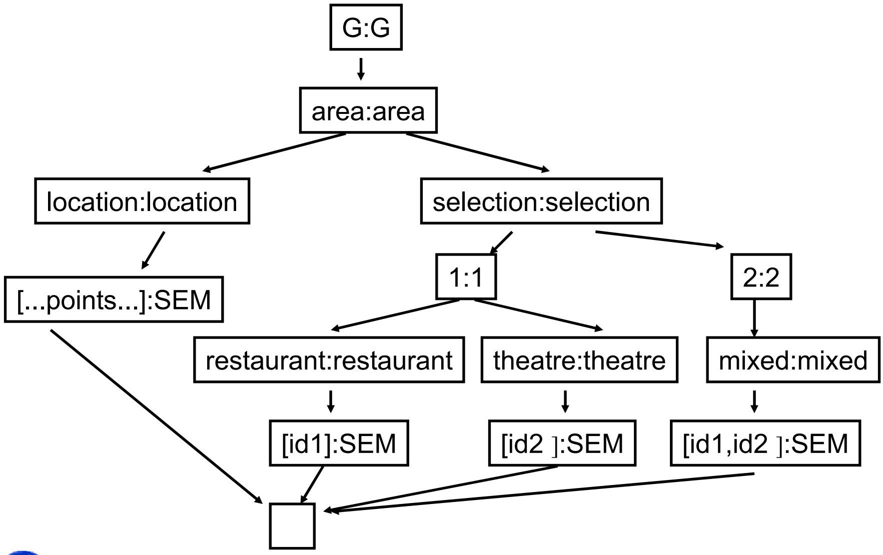
SELECTION → ε:area:ε ε:selection:ε

RESTSG → restaurant:restaurant: ε

Input utterance: "Tell me about this restaurant"

XML Representation read off the semantic slot of the parse-tree terminals:
<info> <restaurant> SEM </restaurant> </info>

Gesture Lattice



G:G

area:area

location:location

selection:selection

[...points...]:SEM

1:1

2:2

restaurant:restaurant

theatre:theatre

mixed:mixed

[id1]:SEM

[id2]:SEM

[id1,id2]:SEM

COMMAND → tell:ε:<info> me:ε:ε about:ε:ε
DEICTICNP ε:ε:</info>

DEICTICNP → DDETSG SELECTION ε:1:ε RESTSG
ε:ε:<restaurant> ε:SEM:SEM ε:ε:</restaurant>

DDETSG → this:G:ε

SELECTION → ε:area:ε ε:selection:ε

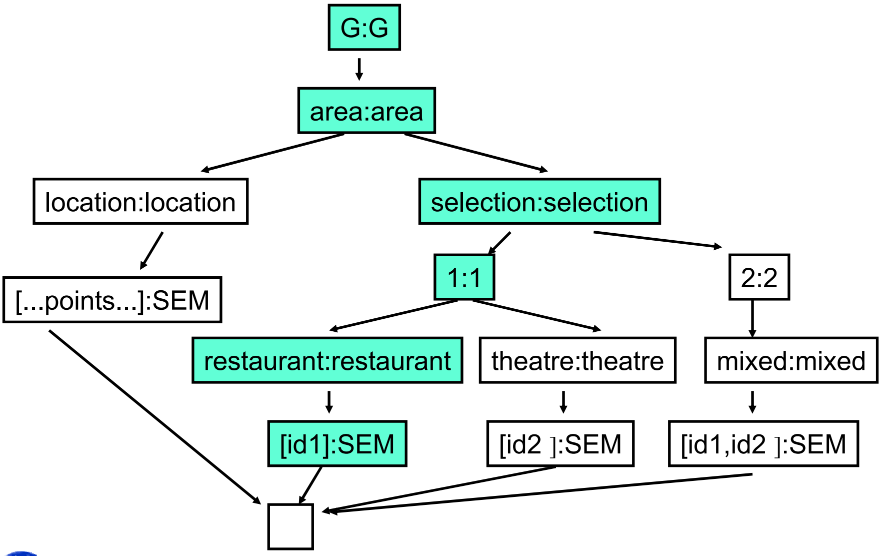
RESTSG → restaurant:restaurant: ε

Cosntraints on Gestural Information

COMMAND → tell:ε:<info> me:ε:ε about:ε:ε
DEICTICNP ε:ε:</info>
DEICTICNP → DDETSG SELECTION ε:1:ε RESTSG
ε:ε:<restaurant> ε:SEM:SEM ε:ε:</restaurant>
DDETSG → this:G:ε
SELECTION → ε:area:ε ε:selection:ε
RESTSG → restaurant:restaurant: ε

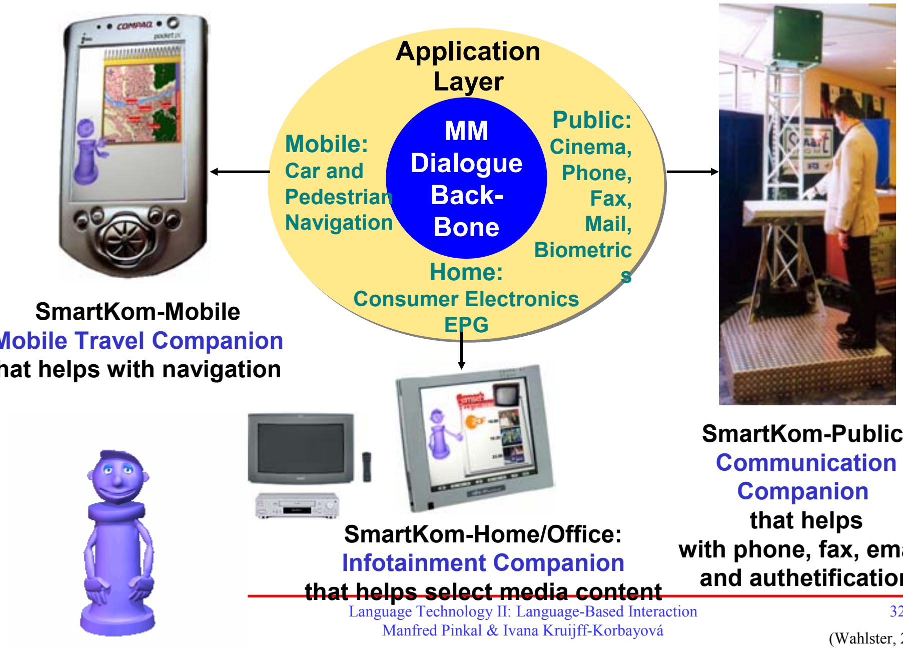
G area selection 1 restaurant SEM

Gesture Lattice



- SEM variable is instantiated by the appropriate reference object from the gesture lattice:
- <info> <restaurant> SEM </restaurant> </info>
- <info> <restaurant> [id1] </restaurant> </info>

SmartKom



Application Layer

MM Dialogue Back-Bone

Mobile: Car and Pedestrian Navigation

Public: Cinema, Phone, Fax, Mail, Biometrics

Home: Consumer Electronics EPG

SmartKom-Mobile
Mobile Travel Companion
that helps with navigation

SmartKom-Home/Office:
Infotainment Companion
that helps select media content

SmartKom-Public:
Communication Companion
that helps
with phone, fax, email, and authetification

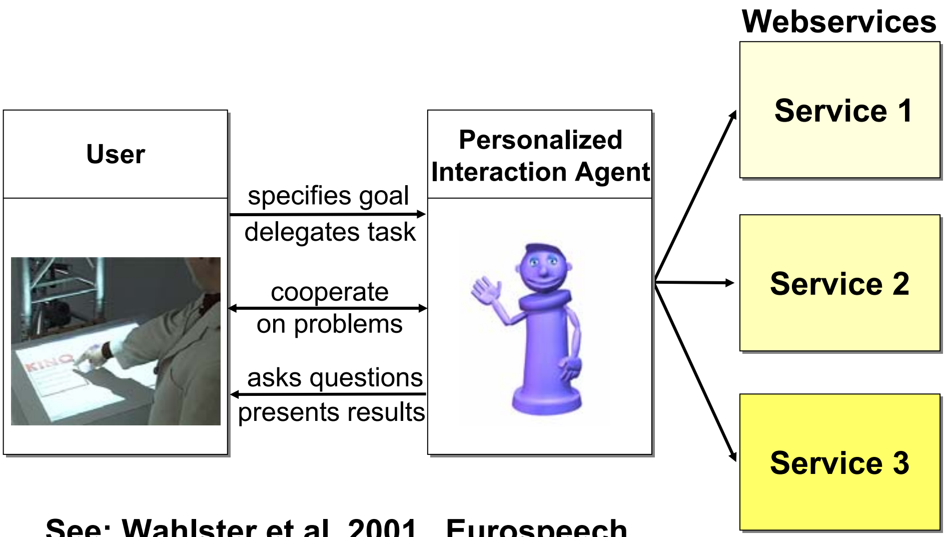
(Wahlster, 2003)

SmartKom

	Input by the User	Output by the Presentation agent
Speech	+	+
Gesture	+	+
Facial Expressions	+	+

(Wahlster, 2003)

SmartKom

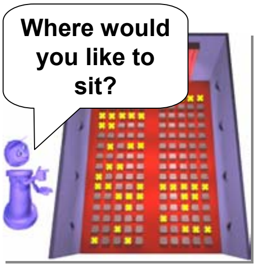


See: Wahlster et al. 2001 , Eurospeech

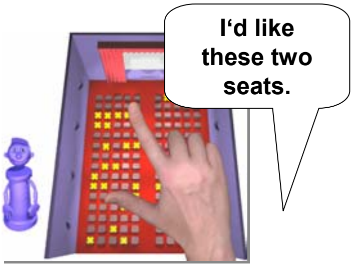
(Wahlster, 2003)

SmartKom – An Example

User Input:
Speech and Gesture



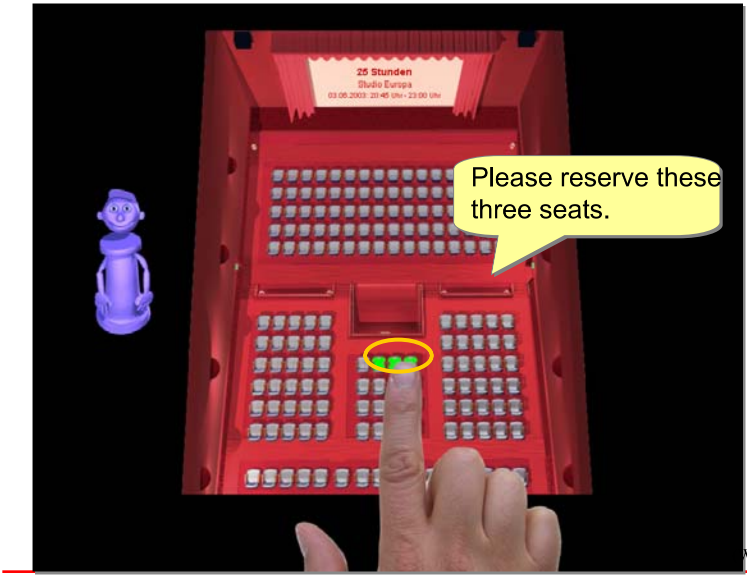
Smartakus Output:
Speech, Gesture and
Facial Expressions



User Input:
Speech and Gesture

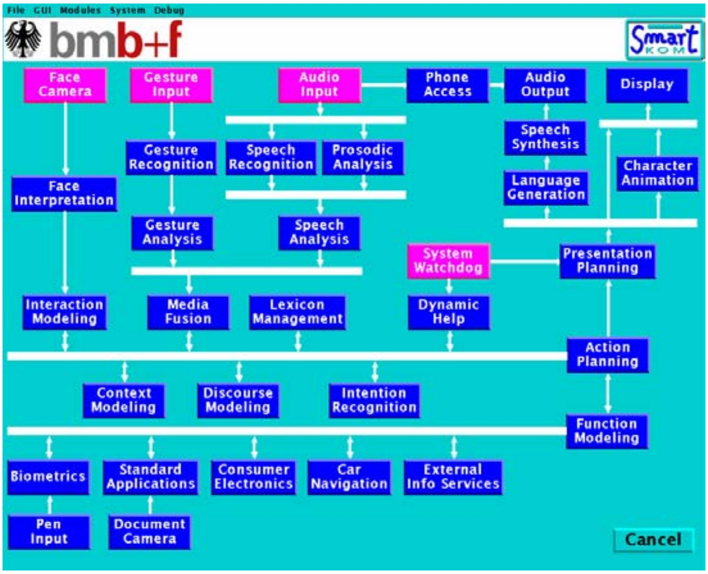
(Wahlster, 2003)

SmartKom – An Example



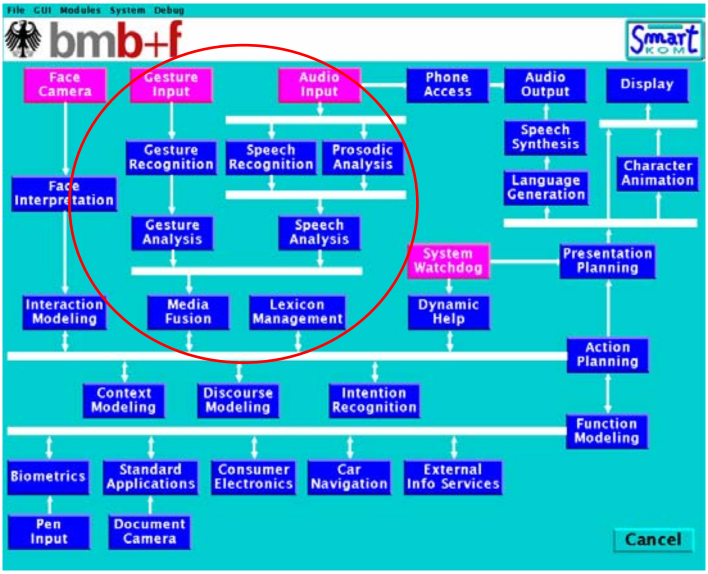
Wahlster, 2003)

The High-Level Control Flow of SmartKom



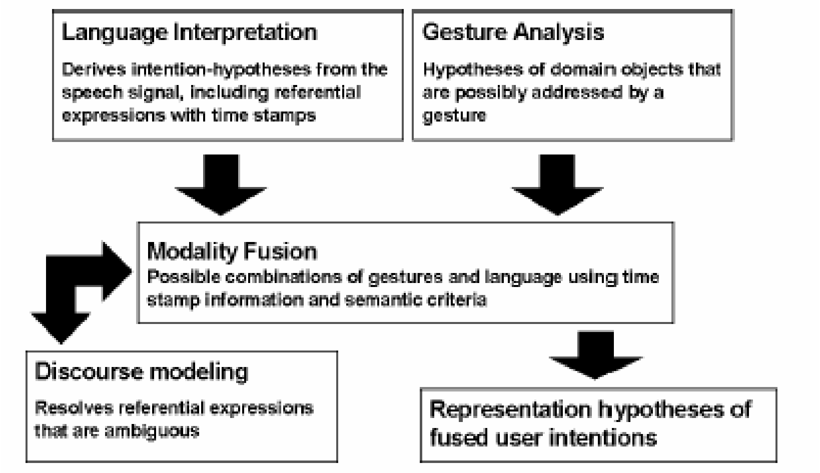
(Wahlster, 2003)

Multimodal Fusion



(Wahlster, 2003)

Late Modality Integration in SmartKom



Language Interpretation

Derives intention-hypotheses from the speech signal, including referential expressions with time stamps

Gesture Analysis

Hypotheses of domain objects that are possibly addressed by a gesture

Modality Fusion

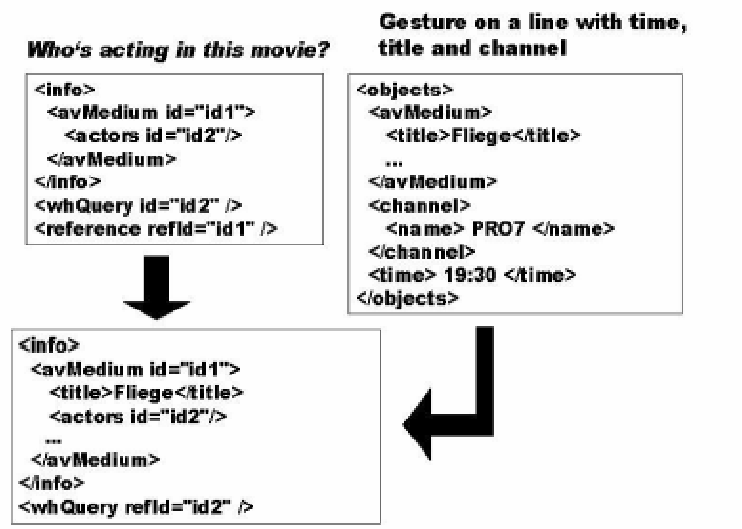
Possible combinations of gestures and language using time stamp information and semantic criteria

Discourse modeling

Resolves referential expressions that are ambiguous

Representation hypotheses of fused user intentions

Late Modality Integration in SmartKom



Who's acting in this movie?

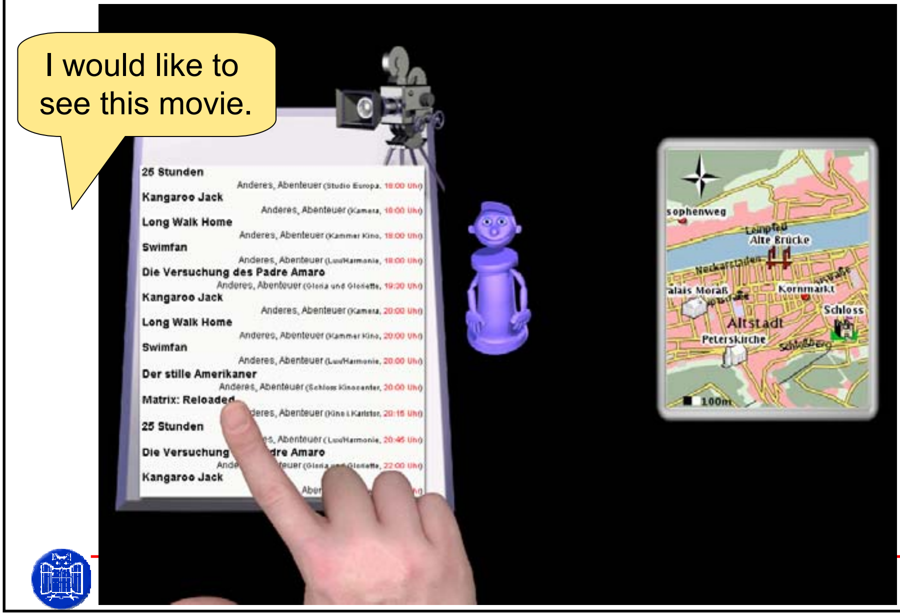
```
<info>
  <avMedium id="id1">
    <actors id="id2"/>
  </avMedium>
</info>
<whQuery id="id2" />
<reference refId="id1" />
```

Gesture on a line with time, title and channel

```
<objects>
  <avMedium>
    <title>Fliege</title>
   ...
  </avMedium>
  <channel>
    <name> PRO7 </name>
  </channel>
  <time> 19:30 </time>
</objects>
```

```
<info>
  <avMedium id="id1">
    <title>Fliege</title>
    <actors id="id2"/>
   ...
  </avMedium>
</info>
<whQuery refId="id2" />
```

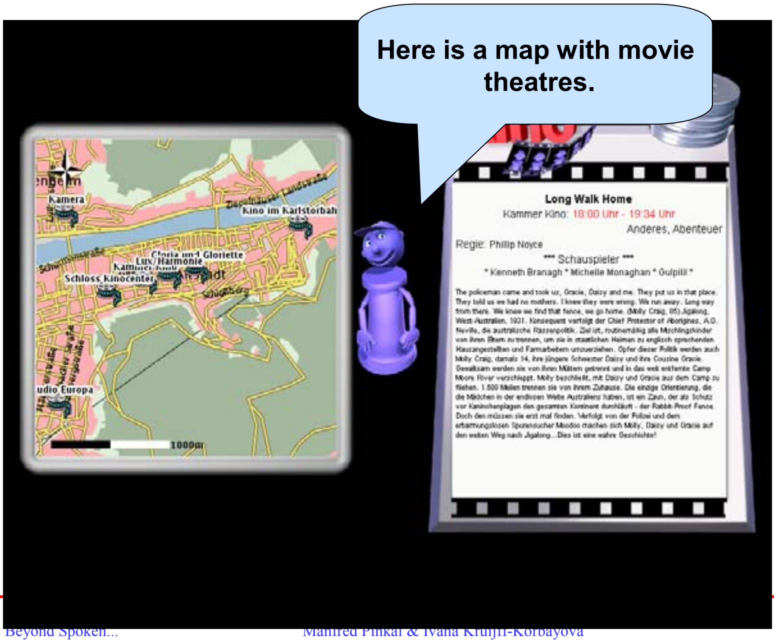
Reference Resolution based on a Symbolic Representation of the Smart Graphics Output



I would like to see this movie.

25 Stunden
Anderes, Abenteuer (Studio Europa, 18:00 Uhr)
Kangaroo Jack
Anderes, Abenteuer (Kamera, 18:00 Uhr)
Long Walk Home
Anderes, Abenteuer (Kammer Kino, 18:00 Uhr)
Swimfan
Anderes, Abenteuer (Lux/Harmonie, 18:00 Uhr)
Die Versuchung des Padre Amaro
Anderes, Abenteuer (Gloria und Gloriette, 19:30 Uhr)
Kangaroo Jack
Anderes, Abenteuer (Kamera, 20:00 Uhr)
Long Walk Home
Anderes, Abenteuer (Kammer Kino, 20:00 Uhr)
Swimfan
Anderes, Abenteuer (Lux/Harmonie, 20:00 Uhr)
Der stille Amerikaner
Anderes, Abenteuer (Schloss Kinocenter, 20:00 Uhr)
Matrix: Reloaded
deres, Abenteuer (Kino i. Karlstor, 20:15 Uhr)
25 Stunden
es, Abenteuer (Lux/Harmonie, 20:45 Uhr)
Die Versuchung dre Amaro
Ande teuer (Gloria Gloriette, 22:00 Uhr)
Kangaroo Jack
Aber

Generating Maps, Animations and Information Displays on the Fly

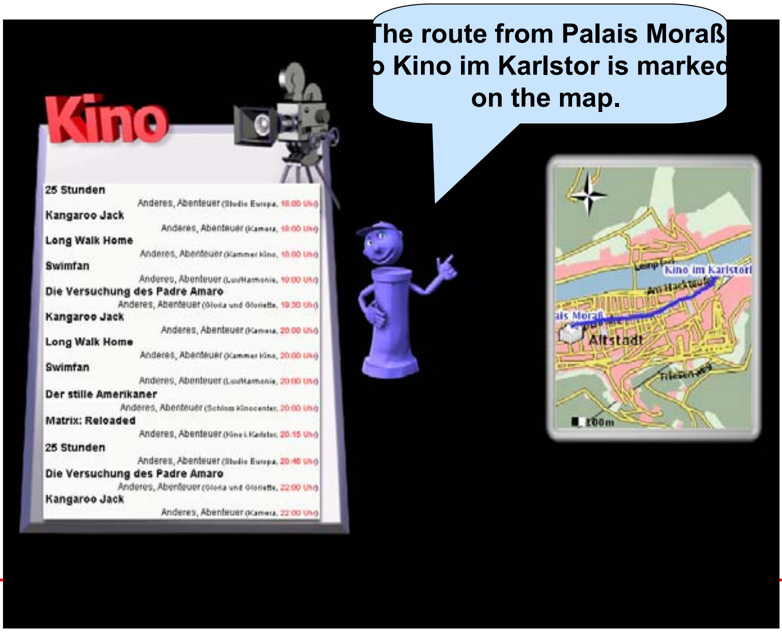


Here is a map with movie theatres.

Long Walk Home
Kammer Kino: 18:00 Uhr - 19:34 Uhr
Anderes, Abenteuer
Regie: Phillip Noyce
*** Schauspieler ***
* Kenneth Branagh * Michelle Monaghan * Gulpilil *

The policeman came and took us, Gracie, Daisy and me. They put us in that place. They told us we had no mothers. I knew they were wrong. We run away. Long way from there. We knew we find that fence, we go home. (Molly Craig, 85) Jigalong, West-Australien, 1931. Konsequent verfolgt der Chief Protector of Aborigines, A.O. Neville, die australische Rassenpolitik. Ziel ist, routinemäßig alle Mischlingskinder von ihren Eltern zu trennen, um sie in staatlichen Heimen zu englisch sprechenden Hausangestellten und Farmarbeitern umzuerziehen. Opfer dieser Politik werden auch Molly Craig, damals 14, ihre jüngere Schwester Daisy und ihre Cousine Gracie. Gewaltsam werden sie von ihren Müttern getrennt und in das weit entfernte Camp Moore River verschleppt. Molly beschließt, mit Daisy und Gracie aus dem Camp zu fliehen. 1.500 Meilen trennen sie von ihrem Zuhause. Die einzige Orientierung, die die Mädchen in der endlosen Weite Australiens haben, ist ein Zaun, der als Schutz vor Kaninchenplagen den gesamten Kontinent durchläuft - der Rabbit-Proof Fence. Doch den müssen sie erst mal finden. Verfolgt von der Polizei und dem erbarmungslosen Spurensucher Moodoo machen sich Molly, Daisy und Gracie auf den weiten Weg nach Jigalong...Dies ist eine wahre Geschichte!

Synchronization of Map Update and Character Behaviour

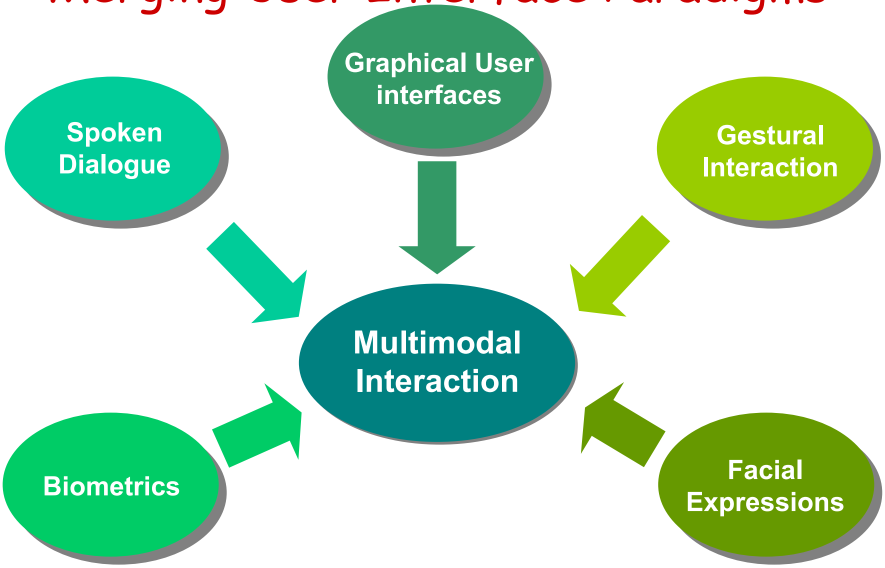


The route from Palais Moraß
o Kino im Karlstor is marked
on the map.

Kino

25 Stunden
Anderes, Abenteuer (Studio Europa, 18:00 Uhr)
Kangaroo Jack
Anderes, Abenteuer (Kamera, 18:00 Uhr)
Long Walk Home
Anderes, Abenteuer (Kammer Kino, 18:00 Uhr)
Swimfan
Anderes, Abenteuer (Lux/Harmonie, 19:00 Uhr)
Die Versuchung des Padre Amaro
Anderes, Abenteuer (Gloria und Gloriette, 19:30 Uhr)
Kangaroo Jack
Anderes, Abenteuer (Kamera, 20:00 Uhr)
Long Walk Home
Anderes, Abenteuer (Kammer Kino, 20:00 Uhr)
Swimfan
Anderes, Abenteuer (Lux/Harmonie, 20:00 Uhr)
Der stille Amerikaner
Anderes, Abenteuer (Schloss Kinocenter, 20:00 Uhr)
Matrix: Reloaded
Anderes, Abenteuer (Kino i. Karlstor, 20:15 Uhr)
25 Stunden
Anderes, Abenteuer (Studio Europa, 20:45 Uhr)
Die Versuchung des Padre Amaro
Anderes, Abenteuer (Gloria und Gloriette, 22:00 Uhr)
Kangaroo Jack
Anderes, Abenteuer (Kamera, 22:00 Uhr)

Merging User Interface Paradigms



Graphical User interfaces

Spoken Dialogue

Gestural Interaction

Multimodal Interaction

Biometrics

Facial Expressions

(Wahlster, 2003)

References

- M. Johnston et al. “MATCH: An architecture for Multimodal Dialogue Systems.” In Proc. Of the 40th Annual Meeting of ACL. pp. 376-383. 2002.
- N. Pfleger et al. “Robust Multimodal Discourse Processing.” In Proc. Of DiaBruck. pp. 107-114. 2003.
  SmartKom website: http://www.smartkom.org/