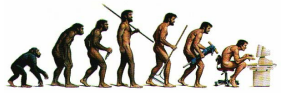
Human-Computer Interaction

Session 12
Multimodal Interfaces

Evolution of HCI



Year	Paradigm	Implementation
1950s	None	Switches, punched cards
1970s	Typewriter	Command-line interface
1980s	Desktop	Graphical UI (GUI), direct manipulation
1980s+	Spoken Natural Language	Speech recognition/synthesis, Natural language processing, dialogue systems
1990s+	Natural interaction	Perceptual, gesture-based multimodal, interactive, conversational, tangible, adaptive
2000s+	Social interaction	Agent-based, anthropomorphic,social, emotional, affective, collaborative

Multimodal interfaces

- Highly perceptual, attentive, multimodal interfaces modeled after natural human-to-human interaction

perceives, attends to, and responds to various, even subtle cues

based on an integrative notion, not just use of mouse, keyboard, speech, etc. aside of each other

- **Goal**: For people to be able to interact with computers in a way similar to how they interact with each other and with the physical world

Is this a multimodal user interface?

- **NO** - all user actions are explicit commands, issued in different interchangable ways
- so, use of speech and point & click alternatively, but not integrated, multimodally

What is a „modality“ ?

physiological

sensory modality

Capability of sensory perception: *visual, auditory, tactil, olfactory, gustatory, vestibular*

motoric modality

Capability of acting or communicating: *verbal, manual, mimic, bodily*

technical

Modality as *interaction technique*

Combination <*d,L*> of an interaction *device d* with an interaction *language L*

What is a „modality“ ?

- Natural or fundamental modalities are part of the communicative faculties of a (social) being - including: *speech (sounds), gesture, mimics, body language (proxemics), prosody, etc.*
- The use of (even the natural) modalities is, at least partially, culturally dependent
  - *Exception*: expression of emotions through face, prosody, body posture, etc.
- Enculturated modalities: learned and habituated specific techniques, e.g. reading & writing or point-and-click

What is a „modality“ ?

Definition:

*A **modality** is a communicative system that is characterized by a specific way of coding, transmitting, and interpreting information.*

- Concerns the transmission of information from the user to the machine (input modalities) as well as from the machine to the user (output modalities)
- An user interface can be called **multimodal**, iff it provides input or output combining multiple modalities, so that the resulting communicative system is more powerful (modalities can be partly redundant in that)

What is „multimodality“ ?

Definition:

*An user interface can be called **multimodal**, iff it provides input or output combining multiple modalities*

- Goal: resulting multimodal communicative system should be more „powerful“ than each single modality alone
- Modalities may be redundant, encoding similar information, but in different ways with different dis-/advantages
- Additional power (and complexity) arises from the way in which the modalities are combined and related to each other (cross-modal relations)

Why is multimodality a good thing?

Bandwidth & efficiency of information codings

- can communicate more information per time unit

Redundancy & robustness

- less errors by putting same information into different modalities
- mutual disambiguation of modalities
- less stress and abrasion in each modality

Adequancy of information coding/multi-functionality

- different information conveyed in different modalities
  - propositional (content) vs. interactional (turn-taking, feedback)
  - symbolic vs. iconic vs. indexical

Adaptivity & *universal design*

- can utilize best modality under changing conditions
- allow different user groups (e.g. blind) in different situations (e.g. noisy)

Why is multimodality a good thing?

Naturalness & Intuitivity

- better adaptation to human user
- interacting can be more automatic/unconscious
- different users prefer different modalities, better acceptance espc. with unexperienced users

Error-proneness

- user intuitively select the modus which is least error-prone, change modality after errors
- user employ simpler instructions/language when interacting multimodally – reduces complexity by distribution of information
  - under cognitive load, users tend to employ multimodal ways of instructions, with less cross-modal coordination

- Study by Oviatt et al. (ICMI'04)
  - task: instruct the map system to coordinate emergency resources
  - different levels of difficulty

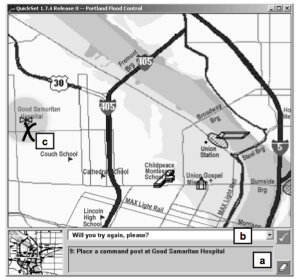
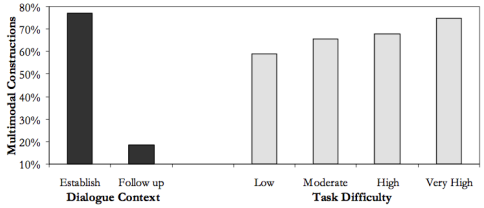


Figure 1. User interface

Difficulty	Message from Headquarters
Low	Situate a volunteer area near ***Marquam Bridge***
Moderate	Send a barge from ***Morrison Bridge barge area*** to ***Burnside Bridge dock***
High	Draw a sandbag wall along ***east riverfront*** from ***OMSI*** to ***Morrison Bridge***
Very High	Place a maintenance shop near the ***intersection of I-405*** and ***Hwy 30*** just ***east*** of ***Good Samaritan***

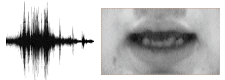


In cognitively difficult tasks:

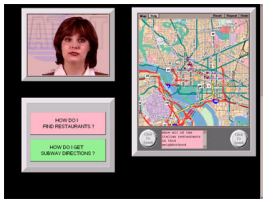
- more errors and longer reaction times
- people switch to multimodal (speech+pen) input

Kinds of multimodal interaction

Speech recognition
+ lip reading



input



output

Research Roadmap of Multimodality 2001-2010

Enabling Technologies and Important Contributing Research Areas

2 Nov. 2001
Dagstuhl Seminar
Fusion and Coordination
in Multimodal Interaction
edited by: W. Wahlster

Multimodal Input	Multimodal Interaction	Multimodal Output
● Sensor Technologies	● User Modelling	● Smart Graphics
● Vision	● Cognitive Science	● Design Theory
● Speech & Audio Technology	● Discourse Theory	● Embodied Conversational Agents
● Biometrics	● Ergonomics	● Speech Synthesis

● Machine Learning ● Formal Ontologies ● Pattern Recognition ● Planning

Multimodal Interfaces vs. GUIs

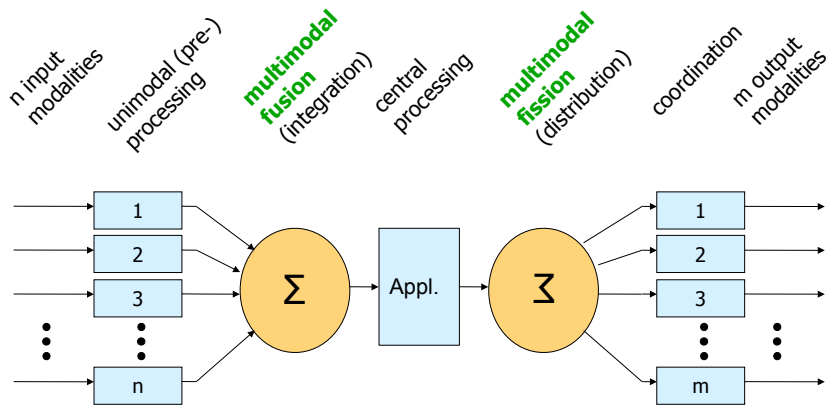
GUIs

1. Assume there is a single event stream that controls event loop with sequential processing
2. Assume that interface actions (e.g. selection of items) are atomic and unambiguous
3. Separable from application software and resides centrally on one machine
4. No temporal constraints, architecture not time sensitive beyond parallel mouse operations

Multimodal Interfaces

1. Typically process continuous and simultaneous input from parallel incoming streams
2. Process input modes using recognition-based technology, good at handling uncertainty and ambiguity
3. Large computational and memory requirements, typically distributed (e.g. multi-agent systems)
4. Time stamping of input, temporal constraints on mode fusion operations

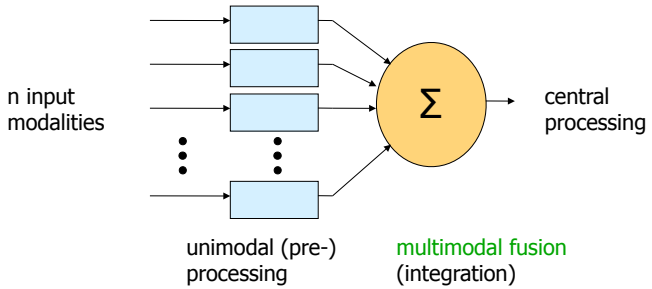
Multimodal interface: basic layout



Multimodal input processing

Multimodal input processing

- The sensing, processing and integration of multiple input modalities for the communication between a user and the computer



Multimodal fusion/integration

Two central problems (Srihari, 1995):

segmentation problem

how can a system be made to cope with `open input´? how can continuous input be segmented into units that can be processed in one system cycle?

correspondence problem

how to determine what relates to what across the multiple input modalities?

Multimodal fusion/integration

- Different approaches based on
  - temporal or structural (syntactical) relations
    Example: *"stell dieses <Zeigegeste> Ding dort hin"*
    → Does the gesture refer to the object (dieses) or the location (dort)?
  - semantic-pragmatic relations
    Example: „drehe diese <ikonische Geste> Leiste so herum"
    → Does the rotation gesture refer to the object or the action?
- Common approach: adoption and extension of techniques from natural language parsing, i.e. multimodal grammars/parsing

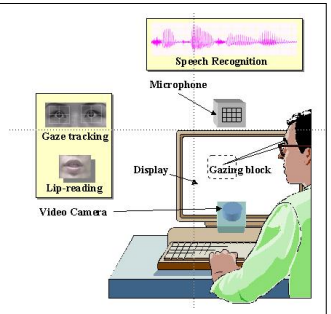
Language

- *Symbolic* modality
  - words = signs with conventionalized meanings
  - modified in context
  - Exception: *Onomatopoetika* (Lautmalerei)
- Speech
  - not only spoken language
  - additional modalities that bear non-symbolic information: prosody

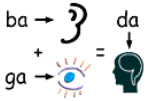
(for NLP, see previous lectures)

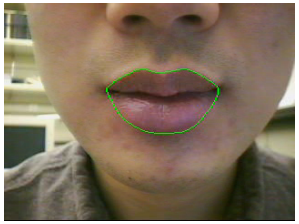
Audio-visual interfaces

- process speech + face video
- lip reading of movements of the mouth during speaking
- eye/gaze tracking



- Utilized to increase speech recognition and processing, esp. in noisy situation (e.g. car)
  - cognitively plausible (recall: „McGurk-Effekt")





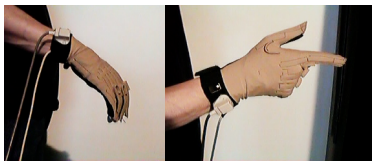
Bimodal speech rec., *Rockwell Scientific Comp.*

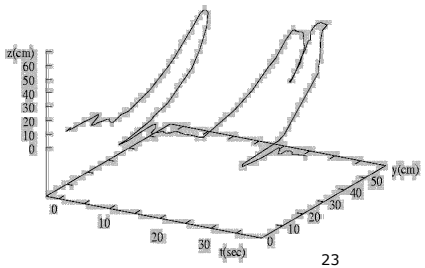
Gesture-based interfaces

- Use hands to interact with the system
  - direct manipulation: direct coupling and feedback
  - indirect manipulation: system mediates movements
  - gesture communication: hands used to communicate to the system
- Requires tracking, recognition & interpretation

Gesture-based interfaces

- Technology: camera-based, active tracking (data gloves, sensors) or passive tracking (marker-based)
- Segmentation problem: How to filter meaningful parts out of the continuous stream of movement signals?
  - Feature-based: hand tension, symmetries, stops, particular form features, etc.
  - Pattern-based: compare with known holistic patterns





Gesture-based interfaces

- **Communicative** Gesture
  - Non-manipulative (i.e. not wiping away something)
  - meaningful (i.e. not nervous fidgeting)

Gestures are movements (here, of the upper limbs) that are produced as a consequence of a communicative intent.

Iconic Gesture
form resembles its referent (object, event)

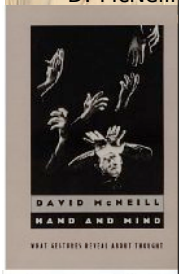
Deictic (indexical) Gesture
refers to an object in the (extra-gestural) context

Symbolic (emblematic) Gesture
arbitrary form, conventionalized meaning within a group of people

Multimodality: Gesture + Speech

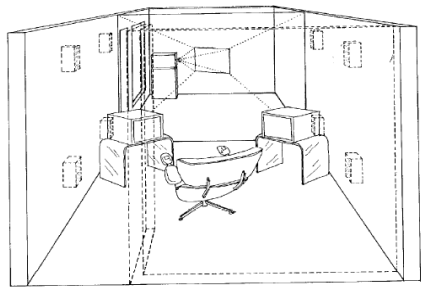
There is a close coupling between speech and gesture – summarized in three rules

- Phonological synchrony
  The *stroke* of a gesture precedes the most prominent syllable or is simultaneous with it
- Semantic synchrony
  Speech and gesture refer to the same overall meaning at the same time.
- Pragmatic synchrony
  When speech and gesture occur together, they fulfill the same pragmatic functions.



The beginning

- **MIT Media Room**(1980)
  - loudspeakers,
  - glass projection screen
  - TV monitors on either side of user's chair
  - joysticks at chair arms
  - touch sensitive pad
  - position-sensing cube attached to wristband
- First projects on multimodal interaction with computers



Put-That-There

(Bolt, 1980)

"Create":
"Create a blue square there."

"Make that …":
"Make that blue triangle smaller"
"Make that smaller"
"Make that like that"

"Move":
"Move the blue triangle to the right of the green square"
"Move that there"
(User does not even have to know what "that" is.)

"Delete":
"Delete that green circle"
"Delete that"

speech + pointing gestures

Processing of commands

"Create a blue square there."

→ Effect of *complete* utterance is a "call" to the *create* routine that needs the object to be created (with attributes) as well as x,y position input from wrist-borne space sensor.

"Call that …the calendar"

→ Recognizer sends code to host system indicating a naming command ("call") → x,y coordinates of item signal are noted by host → host switches speech recognition to training mode to learn the (possibly new) name to be given to the object

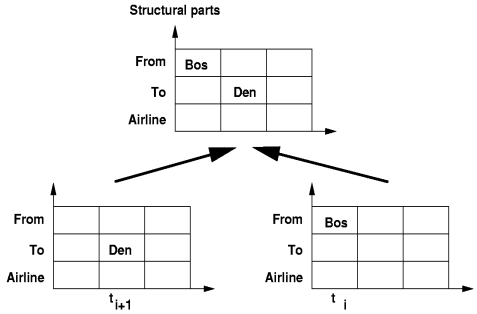
Hard-wired operational, procedural semantics

Multimodal fusion/integration

- Principled solution to correspondence problem?
  - How to fuse information from multiple modalities?
  - What kind of information about the modalities to fuse?
  - How to integrate with preprocessing of each modality?
- Different approaches distinguished according to
  - *what* is fused: pre-semantic vs. semantic
  - *when* fused: early vs. late
  - *how* to fuse: grammar-based vs. unification based

Frame-base integration

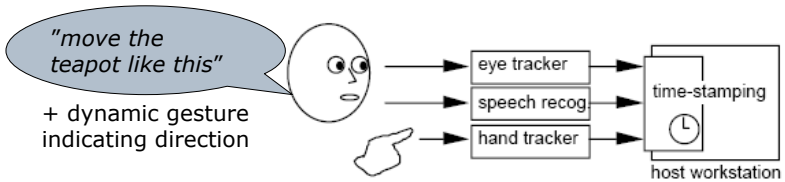
- Modeling user interactions as frames with a fixed set of slots for attribute-value pairs
- Modalities fill slots until whole matrix filled, use of dedicated procedures attachted to slots
- Fixed structure, limited type of interactions



late, semantic, simple unification

Example: ICONIC (Koons et al., 1993)

- Integrating simultaneous speech, gestural, and eye movement (for reference resolution for map and blocks world interaction)
- Problems: timing and abstraction
  - All three streams of data are collected on a central workstation and assigned time stamps, used later to realign data



speech + iconic gestures

Example: ICONIC (Koons et al., 1993)

Step 1 - **Parsing**
- Parse input data stream
- Generate frame-based description of the modality-specific data

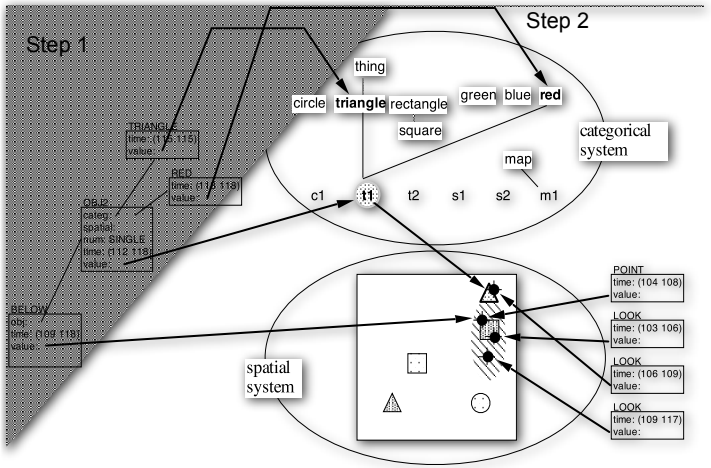
Step 2 - **Evaluation**
- Encode and evaluate the frames based on two models
- Every frame has method that controls search for values in KB

- Knowledge base comprises two representational systems, objects are represented in both
  - categorical system (semantic network)
  - spatial system (locations)

Example: ICONIC (Koons et al., 1993)

"…below the red triangle"

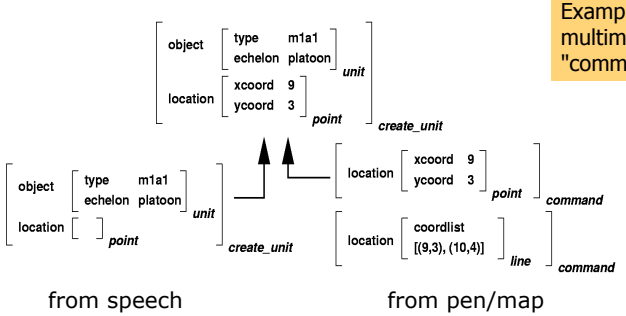
- finds values for each frame in space/category systems
- Integrates spatial values from speech, gesture, eye



Integration with typed AVMs

- Nested Attribute-Values-Matrices (AVMs)
- Use of different frame types
- Unifikation of frame structures
- Computational costly

Example: **QuickSet**, multimodales System für "command-and-control"



from speech

from pen/map

late, semantic, unification

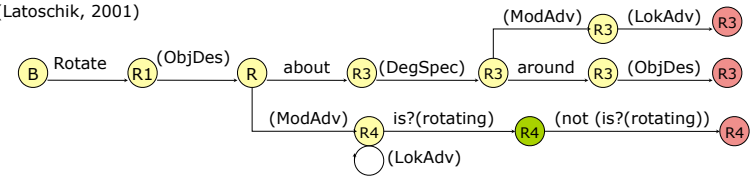
Integration with transition networks

- Parsing multimodal expression with state transition networks (STN, ATN)
- Alphabet of input symbols, e.g. words, gestures
- Problem: Multimodal actions are not sequential; need for flexible temporal relations between input symbols

Example: **tATN**

(Latoschik, 2001)

„Rotate [pointing] this thing about 30 degrees to the right."
„Rotate the yellow wheel like [rotating] this."

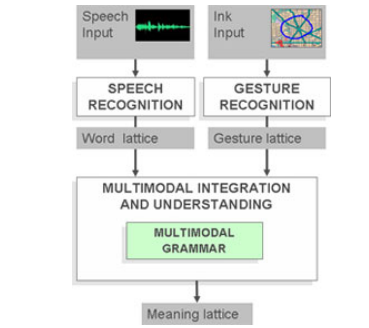


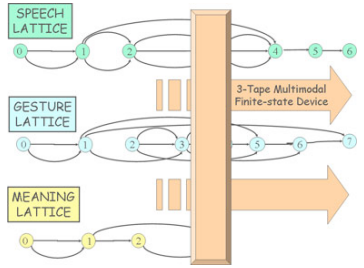
early, pre-semantic, grammar-based

Integrated approaches

- integrated model for speech parsing & understanding, gesture interpretation, multimodal parsing, integration & understanding
- multimodal grammar
  - compiled into finite state device
  - consumes input symbols from lattices representing speech and gesture inputs
  - writes out lattice representing their combined meaning

M. Johnston, AT&T Research

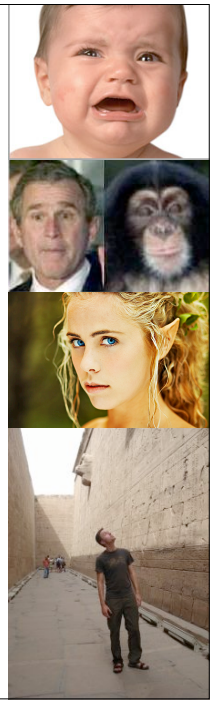




Other input modalities

- similar approaches have been used to include additional modalities in multimodal interfaces

- gaze
  - increasingly seen as modality itself
  - establishes focus of attention, regulates turn-taking, facilitates reference resolution, reflects internal (cognitive) state
- facial expression
  - emotional state (direct reflection of affective state and appraisal of perceived events)
  - modulates communicative acts (e.g. certainty, irony, fun)



Multimodal output generation

Multimodal fission

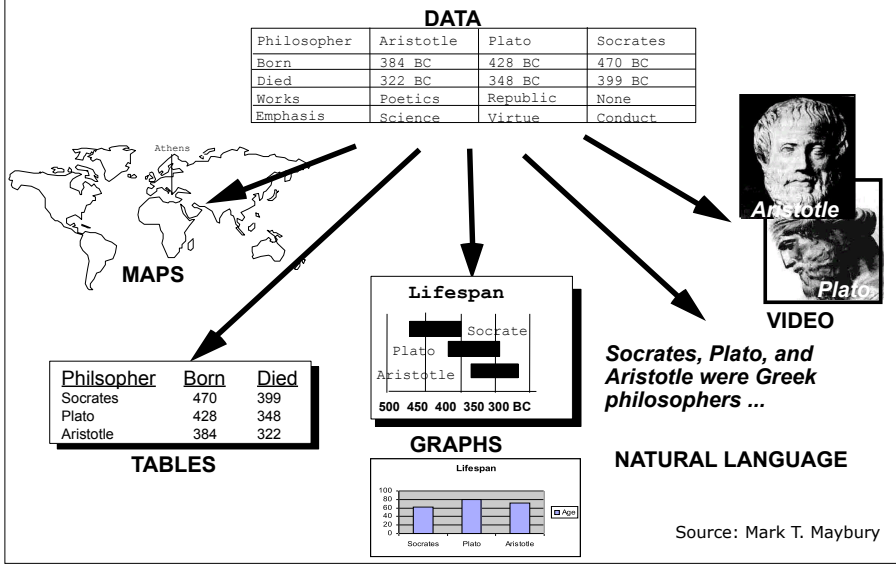
Used in different domains

- **Multimedia:** present information across different media that allow different modalities, usually those known from desktop computers: *text, graphics, animation, sounds, speech, videos, ...*

- **Embodied approach:** system embodied or interfaced via a humanoid figure/robot that serves as communication partner, using natural human modalities also for output generation: *visual speech, prosody, hand gesture, facial expressions, body posture, gaze, head gesture, ...*

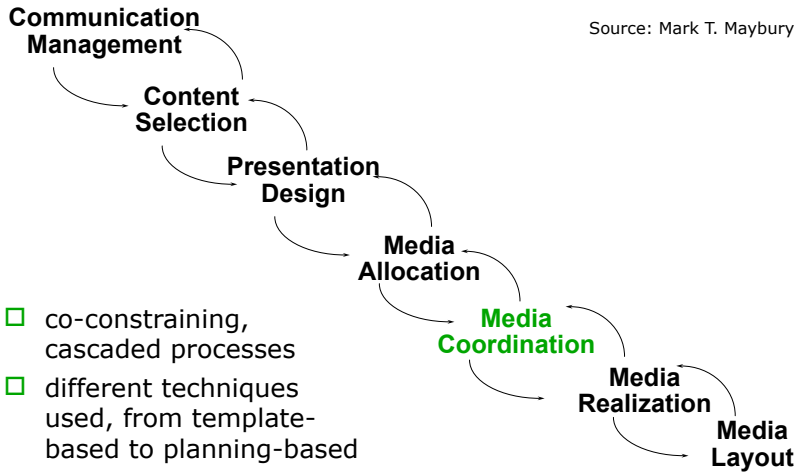
Multimedia Presentation Generation

"No Presentation without Representation"



Source: Mark T. Maybury

Multimedia Presentation Design Tasks



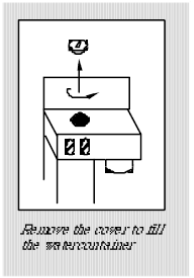
Source: Mark T. Maybury

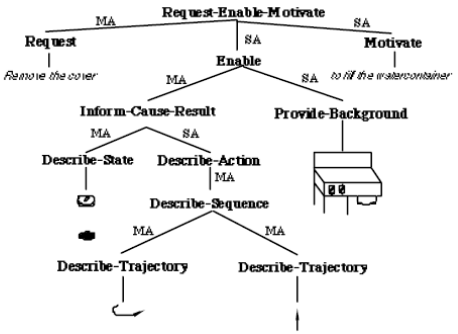
- co-constraining, cascaded processes
- different techniques used, from template-based to planning-based

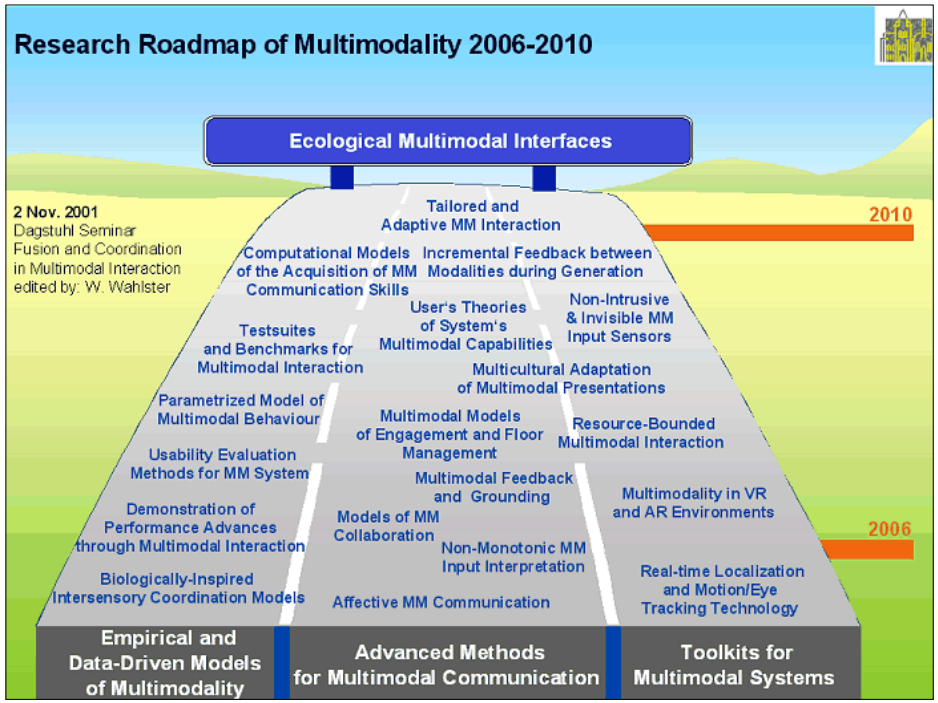
Example: WIP (DFKI, Saarbrücken)

- integrated planning process to create document plan
- repository of communicative acts (cf. speech acts)
- hierarchical goal-refinement into subgoal tree
  - communicative, textual, graphical acts
  - temporal & rhetorical relations between acts

Wahlster et al., 1993;
Andre & Rist, 1993







Next session: agent-based interfaces