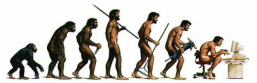
Human-Computer Interaction

Session 11
Multimodal & Perceptual Interfaces

The evolution of user interfaces

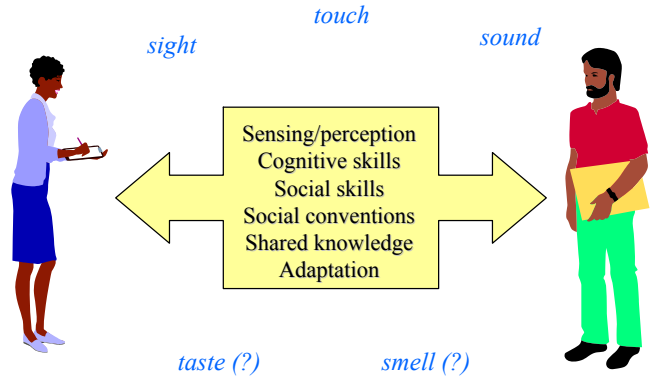


Year	Paradigm	Implementation
1950s	None	Switches, punched cards
1970s	Typewriter	Command-line interface
1980s	Desktop	Graphical UI (GUI), direct manipulation
1980s+	Spoken Natural Language	Speech recognition/synthesis, Natural language processing, dialogue systems
1990s+	Natural interaction	Perceptual,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

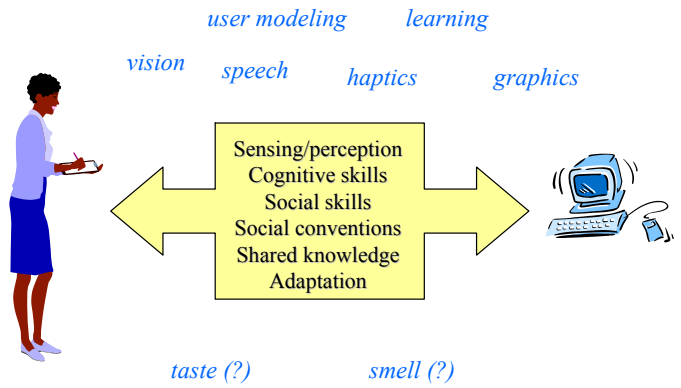
A „perceptual" interface?

- Highly 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 a combination of mouse, keyboard, monitor, speakers, etc.
- **Goal**: For people to be able to interact with computers in a fashion similar to how they interact with each other and with the physical world

Natural human interaction



Natural perceptual interfaces



What's 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*

Natural & enculturated modalities

- 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 are learned and habituated specific techniques, e.g. reading & writing or point-and-click

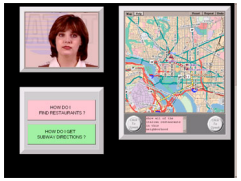
Modality & Multimodality

Definition:

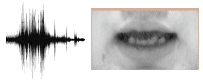
*A **Modality** is a communicative system that is characterized by a specific way of coding, transmitting, and interpreting information.*

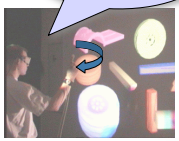
- Concerns the transmission of information from the user to the machine (input modalities) as well as from the machine to the user (output modalities)
- An user interface can be called **multimodal**, iff it provides more than one input modalities and/or output modalities

Multimodal interfaces



Speech recognition
+ lip reading





Example

Why building multimodal interfaces?

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

Bandwidth & efficiency of information codings
- can communicate more information per time unit

Adequancy of information coding/multi-functionality
- different kinds of information can be conveyed by different modality differently well
  - propositional (content) vs. iunteractional/regulating (turn-taking, feedback, attention)
  - symbolic vs. iconic vs. indexical

Alternative ways of communicating *(universal design)*
- pays attention to different user groups (e.g. blind) in different situations (e.g. environmental noise)

Potential advantages

Robustness
- Less stress and abrasion in each modality

Adaptivity
- Allows to utilize the best modality under changing conditions

Redundancy
- Reduce error rate by putting same information into different modalities
- Mutual disambiguation of modalities

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
  - When under cognitive load, users tend to employ multimodal ways of instructions, with information being separated across the modalities (e.g. less redundancy)

Frequent objections

- HCI should be characterized by (e.g. Shneiderman):
  - Direct manipulation
  - *Predictable* interactions
  - Giving responsibility and sense of accomplishment to users
- Won't work –"A.I. hard"
- Technological obstacle
  - But: lots of researchers worldwide, increasing interest, consistent progress
- Economic obstacle
  - But: hw/sw advances, commercial interest in biometrics, accessibility, recognition technologies, virtual reality, entertainment, ...

Multimodal Interfaces vs. GUIs

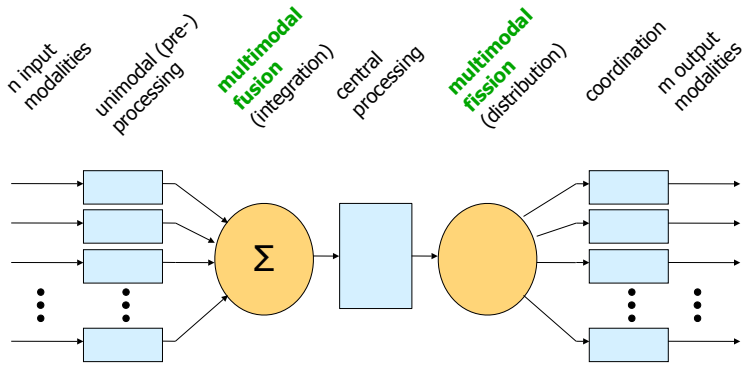
GUIs

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 structure



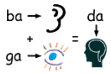
Language

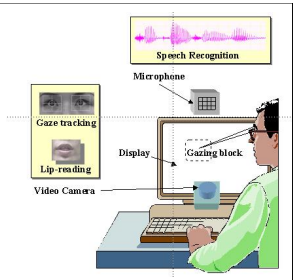
- *Symbolic* modality
  - words = signs with conventionalized meanings
  - modified in context
  - Exception: *Onomatopoetika* (Lautmalerei)
- Spoken Language = Speech
  - comprises additional non-symbolic information: prosody

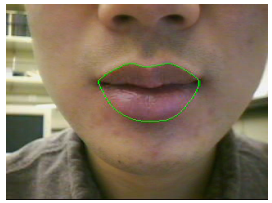
(NLP already covered in this lecture)

Lip reading

- Movements of the mouth during reading
  - audio-visual speech processing
- Utilized to increase speech recognition, esp. under background noise (e.g. in car)
  - recall: „McGurk-Effekt"







Bimodal speech rec.,
Rockwell Scientific Comp.

Gesture

- *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.

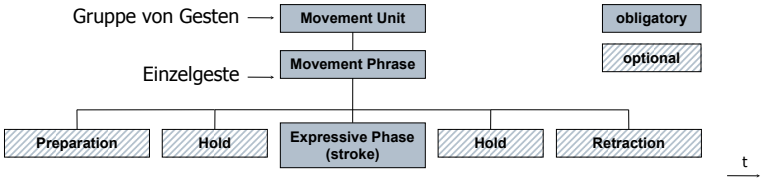
Iconic Gesture
form resembles its referent (object, event)

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

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

Gesture structure

A gesture typically consists of multiple phases

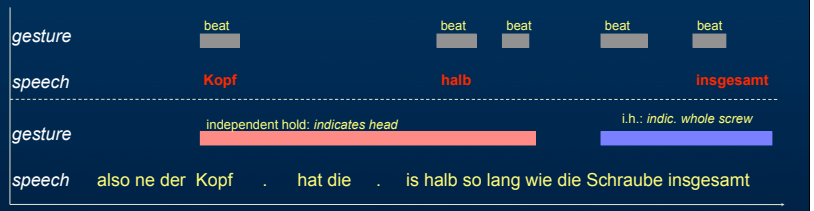


- preparation: bring hands up in starting position
- expressive Phase (stroke): meaning-carrying part
- retraction: bring hands back into a (possibly intermediate) rest position
- hold: no movement

Other functions of gesture

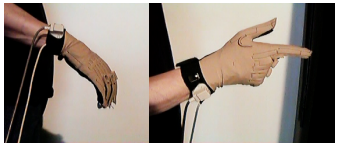
Reflect discourse structure

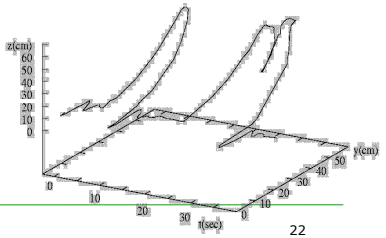
- Convey, and thus mark, discoursively focal elements
- Emphasize
  - Beats or beat-like movement qualities



Gesture recognition

- Technology: camera-based, active tracking (data gloves, sensors) or passive tracking (marker-based) (recall VL „Input Devices")
- *Segmentation problem*:
  How to segment strokes out of the continuous stream of movement signals?
- Possibilities: Expoit features like hand tension, symmetries, stops, particular form features, etc.



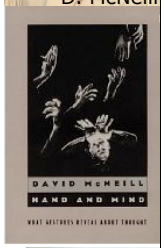


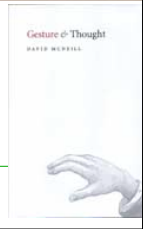
Multimodality: Gesture + Speech

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

- Phonological synchrony
  The s*troke* 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.


D. McNeill

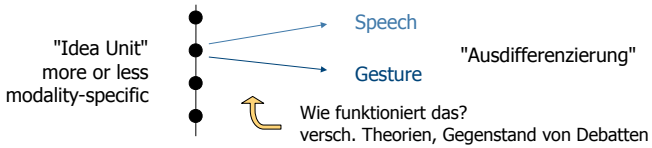




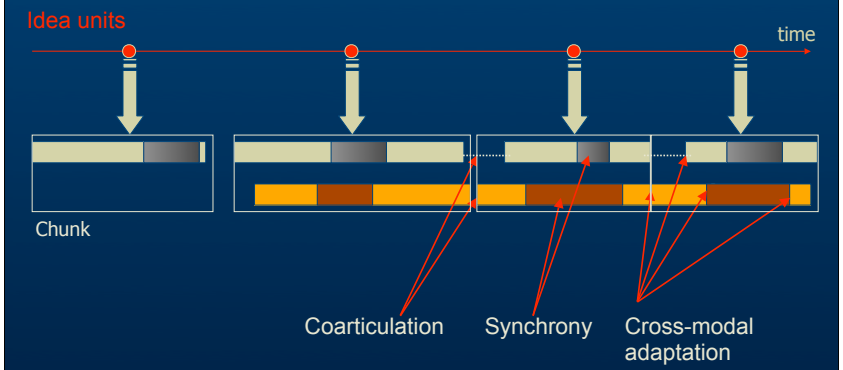
Gesture and speech

The close coupling of speech and gesture led to the theory that coverbal gesture and speech derive from one and the same underlying communicative „idea unit".

Communikation = Sequences of to-be-communicated idea units, which unfold to (or are packed into) speech and gesture.



Overall production of speech and gesture



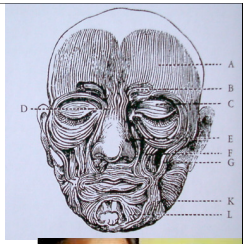
Facial Gesture (Mimik)

Lexicon definition (Duden)

"Gebärden- und Mienenspiel [des Schauspielers] als Nachahmung fremden oder als Ausdruck eigenen seelischen Erlebens"

Biological

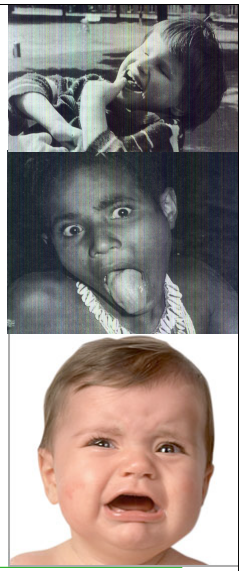
- Movement of the facial tissue and skin due to muscle movement
- also with other primats, but humans have the most differentiated facial gesture (more and finer muscles as e.g. chimpanzees)



Facial expression of emotions

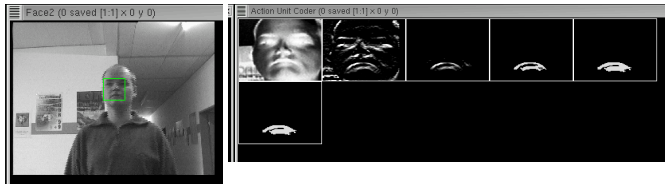
Facial expression conveys emotional states and contributes to communicative feedback

- Darwin:
  - little kids: Wut, Angst, Zuneigung, Freude, Neid, Schüchternheit, Unbehagen
  - + "cognitive" emotions in older children: Scham, Trauer, Verlegenheit, Resignation
- Often, 6 universal basis emotions distinguised
  - Freude, Trauer, Ekel, Überraschung, Wut, Angst
- ...or dimensional models (*Pleasure-Arousal-Dominance)*



Facial expression recognition

- Feature extraction: Finds specific, most indicative parts of the face (Augenbrauen, Augen, Nase, Mund), determines significant features points
- Classification of feature point configuration or movement:
  - emotions (freudig, ärgerlich, …)
  - „Activation Units" (Ekman & Friese)

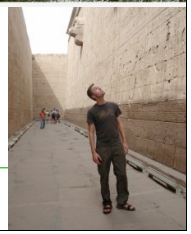


- *Facial Feature Tracking*

www.nevenvision.com (jetzt Google)

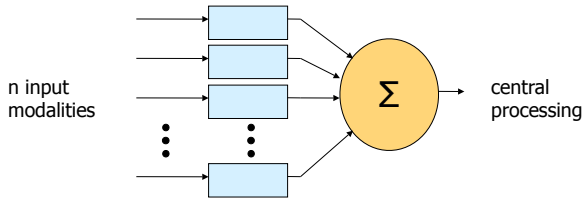
Gaze

- Increasingly considered as a modality on its own right
- important for determination of focus of attention, dialogue management (*turn-taking),* reference resolution
- Reflects internal states
  - gazing up: thinking or retrieval of memory information
  - gazing up + slighlty opened mouth: "what an idiot…"



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

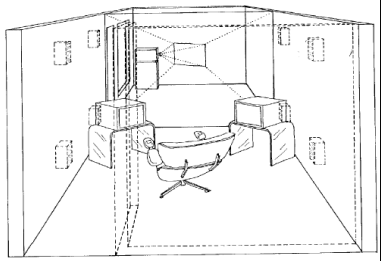
how to determine what relates to what across the multiple input modalities?

Multimodal fusion/integration

- Exploit
  - 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 the realm of natural language parsing (*"multimodal grammars/parsing"*)

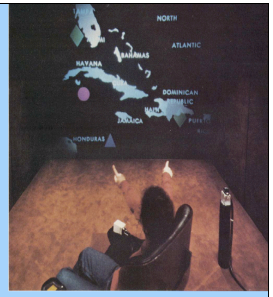
The beginning: MIT Media Room

- loudspeakers, frosted glass projection screen, TV monitors on either side of user's chair
- chair arms with one-inch high joystick sensitive to pressure and direction, touch sensitive pad
- Position-sensing cube attached to wristband



Put-That-There

(Bolt, 1980)



(Graphic taken from [1])

"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"

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.

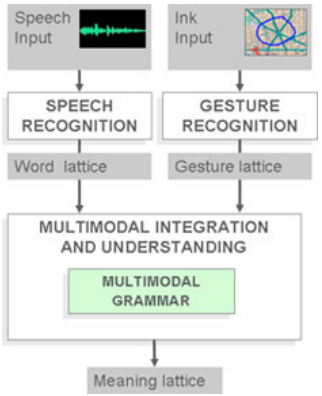
"Call that …the calendar"

→ 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

All utterances processed with hard-wired procedural semantics

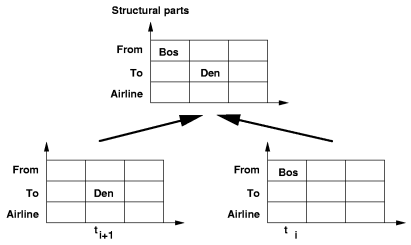
Example: AT&T Labs - Research

The Multimodal Access To City Help:
- 'show cheap italian restaurants in chelsea'.
- circle an area on the map + say 'show cheap italian restaurants in this neighborhood'
- circle an area + write 'cheap' and 'italian'



Frame-base integration

- Modeling user interactions as frames with a fixed set of slots for attribute-value pairs
- Modalities fill slots until the whole matrix (AVM) is filled
- Fixed structure, limited type of interactions



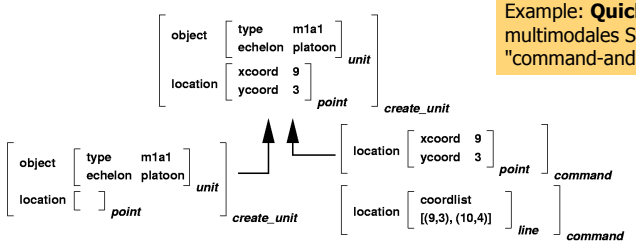
Example: **MATIS**, **M**ultimodal **A**ir **T**ravel **I**nformation **S**ystem

(„Melting pots")

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

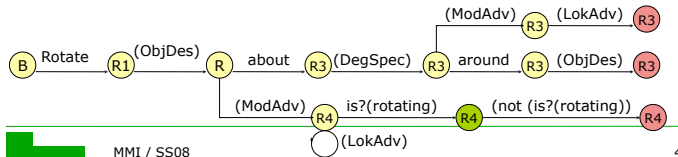
from pen/map

Integration with transition networks

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

Example: **tATN**

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

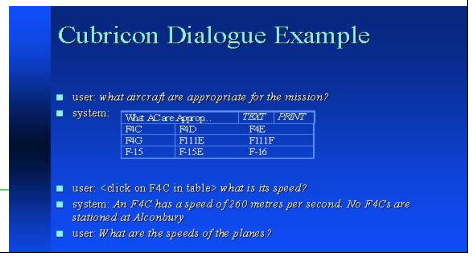


(Latoschik, 2001)

Example: CUBRICON (Neal & Shapiro, 1991)

- System integrating deictic and graphic gestures with simultaneous NL for *both* user input and system output
- interface capabilities
  - Accepts and understands references to entities in NL & pointing
  - Disambiguates unclear references and infers intended referent
  - Dynamically composes and generates synchronous spoken NL, gestures and graphical expressions in output

Calspan-UB Research Center Intelligent CONversationalist

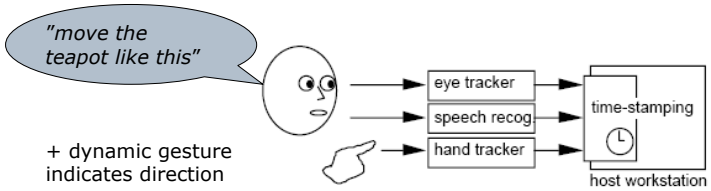


CUBRICON Knowledge Sources

- Multimedia parser: ATN network for NL + mouse gesture
- Used in understanding input and generating output
- Knowledge Sources:
  - Lexicon
  - Grammar: defines multimodal language
  - Discourse Model: Representation of "attention focus space" of dialogue. Has a focus list and display model – tries to retain knowledge pertinent to the dialogue
  - User Model: Has dynamic "Entity Rating Module" to evaluate relative importance of entities to user dialogue and task – tailors output and responses to user's plans, goals and ideas
  - Knowledge Base: Information about task domain, all objects and concepts represented in a single knowledge representation language (semantic net-based)

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



Processing input streams

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

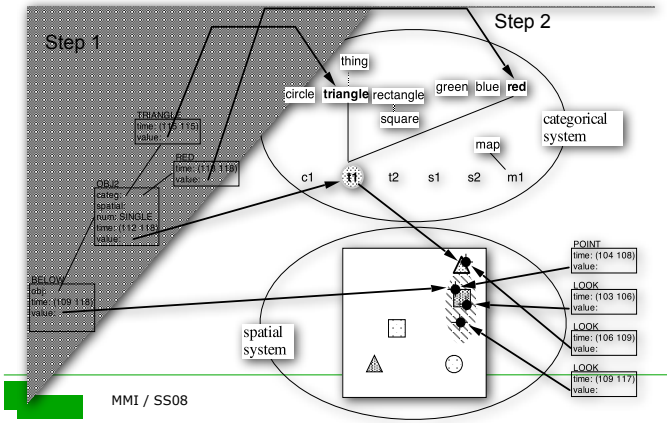
Step 2 - Evaluation
- Encode and evaluate the frames based on two models
- Every frame has method that controls search for frame values in KB

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

ICONIC: Evaluation

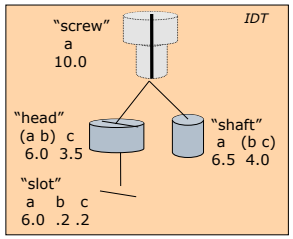
"…below the red triangle"

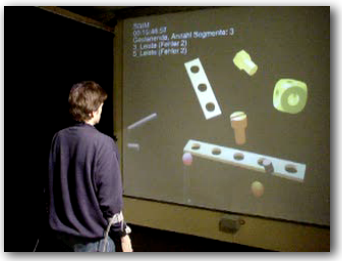
- Finds values for each frame in space/category systems
- Integrates spatial values from speech, gesture, eye



Shape-related expressions (Sowa 2006)

- translate gesture features into spatial representation of shape
- not limited to a single gesture, properties may accumulate over a series of movements and postures
- match shape representation with system's representation of how the objects look like





(Sowa & Wachsmuth, 2005)

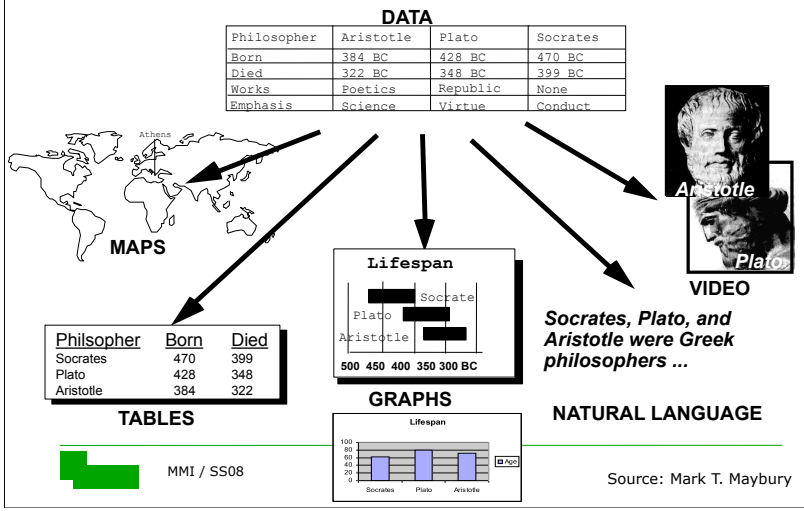
Multimodal fission

Two approaches in different domains

- **Multimedia:** Present information across different media that allow different modalities, usually those known from desktop computers: *text, graphics, animation, sounds, ...*
- **Anthropomorphic approach:** System embodied or interfaced via a humanoid figure/robot that serves as communication partner, using natural human modalities also for output generation: *speech, gesture, mimics, body posture, etc.*

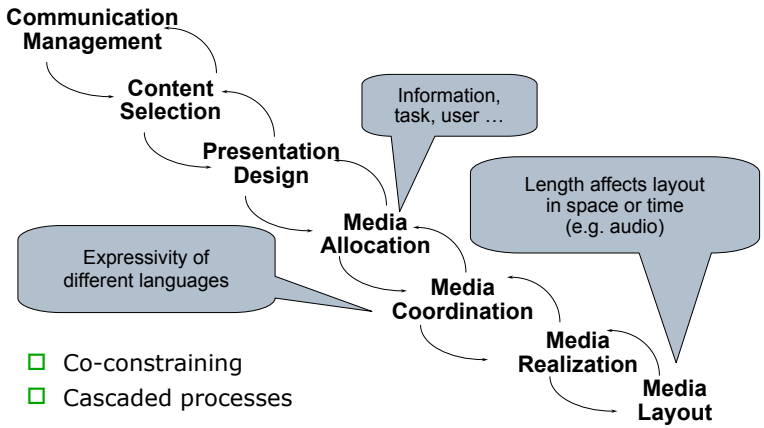
Multimedia Presentation Generation

Credo: *"No Presentation without Representation"*



Source: Mark T. Maybury

Common Presentation Design Tasks



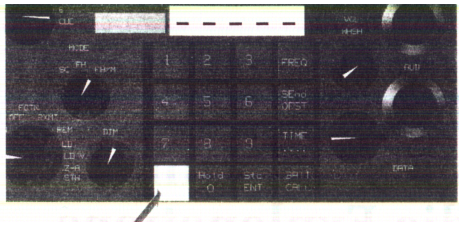
- Co-constraining
- Cascaded processes

Source: Mark T. Maybury

COMET

(**Co**ordinated **M**ultimedia **E**xplanation **T**estbed; Feiner et al. 1993)

- System explains how to diagnose a technical device
- First, content planning (what to be expressed), then microplanning the way of conveying it (how to express it)

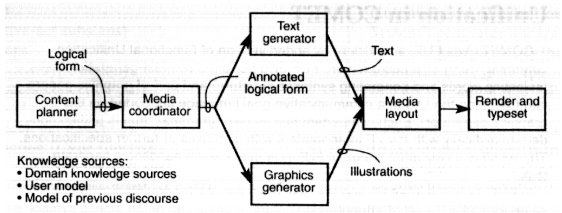


Press the CLR button to clear the display

Media coordination in COMET

Heuristics to decide which information to be encoded in which modality, depending on type of informationen:

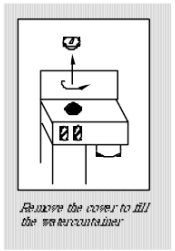
- Location, physical attribute (shape etc.) → graphics
- abstract action, relations (order, causality) → text
- concrete action → graphics + text

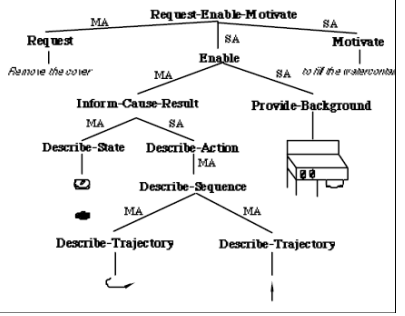


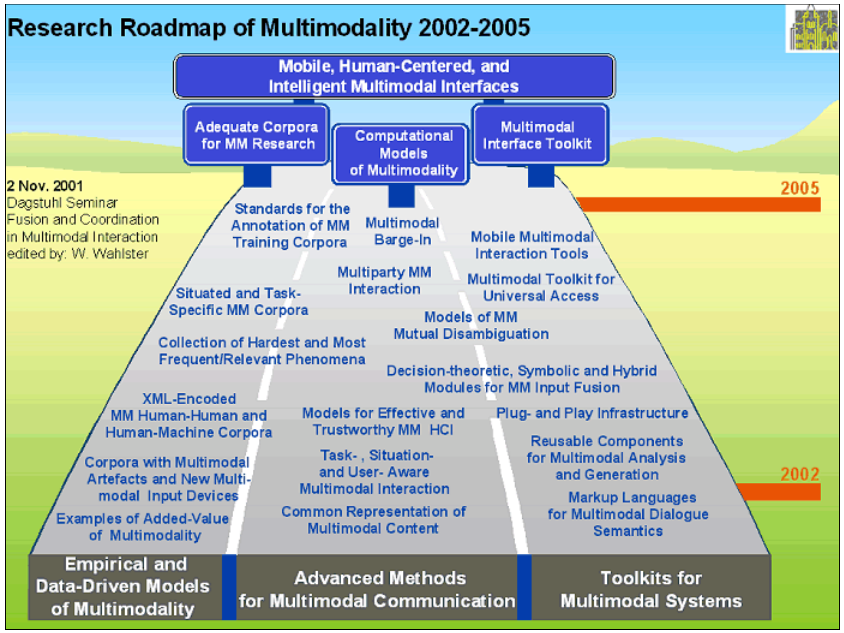
WIP: Use of communicative acts

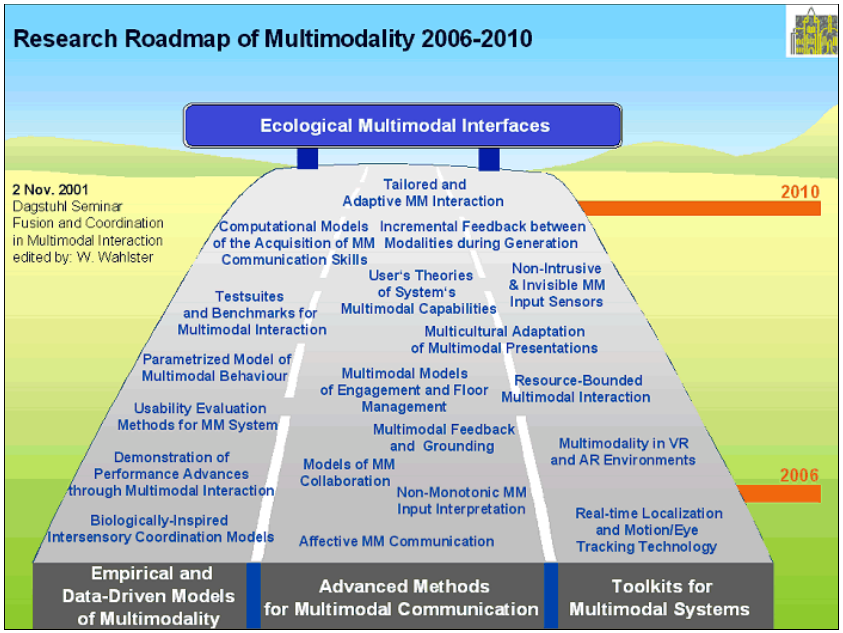
- Integrated planning process to create document plan
- Use of repository of communicative acts (cf. speech acts)
- Goal-refinement into subgoals
  - communicative (e.g., describe)
  - textual (e.g., S-request)
  - graphical (e.g., depict)

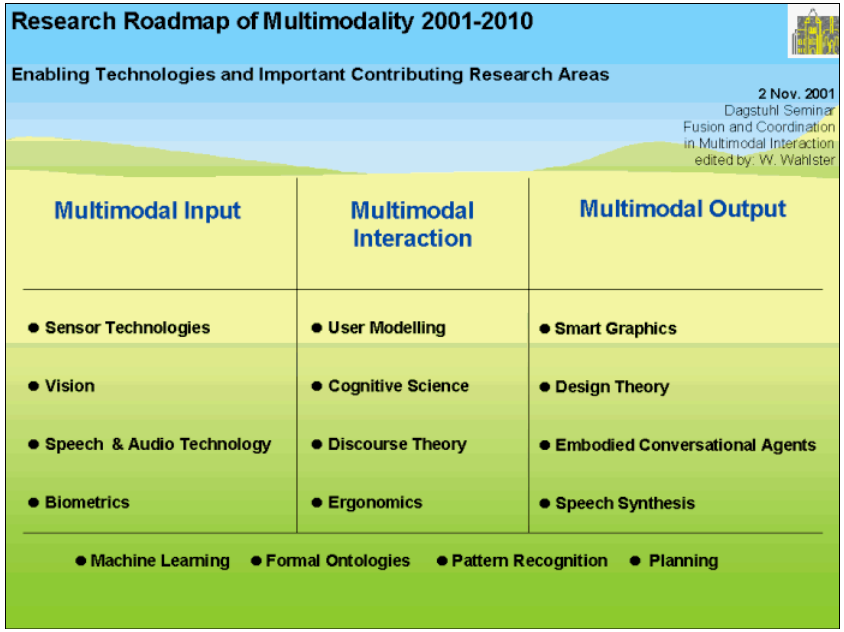
Wahlster et al., 1993; Andre & Rist, 1993











Next session: agent-based interfaces