TILBURG UNIVERSITY

The Open University

Natural Language Generation for Embodied Agents

Emiel Krahmer and Paul Piwek

Plan

1. Introduction: NLG for Embodied Agents
2. What is Natural Language Generation?
3. Case: Generation of Referring Expressions (GRE)
4. NLG for Embodied Agents: Requirements
5. Case: GRE for Embodied Agents
6. Realization: Audiovisual Speech and Gestures
7. Case: NLG for Teams of Embodied Agents
8. Evaluating NLG for Embodied Agents

Plan

1. Introduction: NLG for Embodied Agents
2. What is Natural Language Generation?
3. Case: Generation of Referring Expressions (GRE)
4. NLG for Embodied Agents: Requirements
5. Case: GRE for Embodied Agents
6. Realization: Audiovisual Speech and Gestures
7. Case: NLG for Teams of Embodied Agents
8. Evaluating NLG for Embodied Agents

TILBURG UNIVERSITY

The Open University

What is an ECA?

Embodied

Conversational

Agent

Examples of ECAs

- REA Real Estate Agent (Justine Cassell and collaborators)

Examples of ECAs

- The Mission Rehearsal Exercise (MRE) demonstrator at ICT, University of Southern California
- Virtual reality training environment for the US army.
- Demo movie

Examples of ECAs

- Web persona (WIP/PPP); Andre, Mueller & Rist, 1997). Presentation and tutoring [image]
- SmartKom (DFKI and partners) navigation assistant (mobile) [movie]
- Lexicle Smart Fridge [image fridge] [image agent]
- Lexicle's Cara Virtual Financial Sales Person (for First Direct)
- Charamel GmbH Living.Kiosk [image]

Examples of ECAs

- Product presentation
- Training
- Education
- Customer relations management
- Sales
- Entertainment/Games

- Companies: Zoesis (Mr. Bubb), PlanB, NoDNA, Charamel, Oddcast, Lexicle, Microsoft (Peedy the Parrot), Cantoche (Julie), …
- Research projects: NECA, MRE, MagiCster, …

Examples of ECAs

- RUTH (Douglas de Carlo and Matthew Stone at Rutgers)

- input:
  ((far ((register "HL") (accent "L+H*") (jog "TR")))
  (greater ((accent "!H*") (tone "H-") (blink) (jog)))
  (than ((register "HL-H") (brow "1+2")))
  (any ()) (similar ((accent "L+H*") (jog "D*")))
  (object ((pos nn) (tone "L-") (blink) (brow)))
  (ever ((register "L") (accent "H*") (jog "U*")))
  (discovered ((accent "L+!H*") (tone "L-L%") (blink))))
- output: “far greater than any object ever discovered” [Animation]
- Other systems: e.g., MAX of the Univ. of Bielefeld, Microsoft’s Agents, Cantoche’s LivingActor

Variety

- **Types of application**
  - As a servant for the user performing traditional computer tasks (e.g., the help function)
  - Engage in a role typical in a human-human real-life scenario (tutoring, sales, advice giver)
  - Represent individuals in virtual environments (gaming, chat)
- **Embodiment**
  - View: talking head, body torso
  - Avatar
  - Humanoid
  - Virtual character
- **Autonomy**
  - From scripted to fully autonomous

Why Embodied Conversational Agents?

- Enhance Human-Computer Communication
- Starting point: approximate **face-to-face conversation**
- Why? It is the primary setting for language use (Fillmore, 1981; Clark 1996):

  - Universal (available in all human societies; cf. written language)
  - Arguably the most common setting in daily life
  - It does not require special skills
  - Basic setting for children's acquisition of their first language

Properties of face-to-face conversation (Clark & Brennan, 1991)

- **Immediacy**: Participants can see and hear each other and there is no delay when perceiving each other's actions
- **Medium** (speech, gestures, eye gaze): evanescent, recordless and simultaneous
- **Control**: lies with the participants

Other types of communication are typically

more restrictive and less direct

than face-to-face conversation

NLG for Embodied Agents

- NLG: **Natural Language Generation**
- Required to engage in conversation, but also for more restricted settings.
- Generation versus interpretation (McDonald, 1992)
  - **Generation** – making choices: Given a communicative goal, choose the means to achieve that goal
  - **Interpretation** – hypotheses management: given an utterance, hypothesize what the speaker could have meant
- NLG for **embodied agents**
  - Integrate language generation with generation for other modalities (gaze, gestures, body posture) – multimodality
  - Taking into account the co-presence of the interlocutor

Ongoing Debate

- **Contra**: a human-like appearance of computer-interfaces gives rise to false expectations
- **Pro**: the presence of an ECA may lead to a noticeable improvement in the interaction (the persona effect)

- **Ethical issues**: “If you confuse the way you treat machines with the way you treat people, you may end up treating people like machines, which devalues human emotional experience, creativity, individuality, and relationships of trust” (Shneiderman, 1997)

Plan

1. Introduction: NLG for Embodied Agents
2. What is Natural Language Generation?
3. Case: Generation of Referring Expressions (GRE)
4. NLG for Embodied Agents: Requirements
5. Case: GRE for Embodied Agents
6. Realization: Audiovisual Speech and Gestures
7. Case: NLG for Teams of Embodied Agents
8. Evaluating NLG for Embodied Agents

Plan

1. Introduction: NLG for Embodied Agents
2. What is Natural Language Generation?
3. Case: Generation of Referring Expressions (GRE)
4. NLG for Embodied Agents: Requirements
5. Case: GRE for Embodied Agents
6. Realization: Audiovisual Speech and Gestures
7. Case: NLG for Teams of Embodied Agents
8. Evaluating NLG for Embodied Agents

What?

- Subfield of Artificial Intelligence and Computational Linguistics
- Programs that map non-linguistic to linguistic information
  - Weather data ⇨a weather report
  - Legal inference engine output ⇨ a verbally expressed verdict/recommendation
  - Patient records ⇨ report/summary for doctor/patient/…
  - Product database ⇨ customer-friendly descriptions of products
  - Client database ⇨ mail shot
  - Summarization?
- Input: non-linguistic representation from some application
- Output: text (and mark-up)
- Dominated by choice: deciding amongst alternative ways of expressing information (McDonald, 1992)

Input

- Knowledge source (domain knowledge; database)
- User Model
- Discourse History
- Communicative goal

WeatherReporter (Provide overview of the weather in a given month)
(Reiter & Dale, 2000)

- Numerical data: Year, day, month, time, barometric pressure, temperature, soil temp, etc.
- Hardwired: layperson interested in weather conditions
- Discourse History: none
- SummarizeMonth(January)

Output

- The month was rather dry with only three days of rain in the middle of the month. The total for the year so far is very depleted again.

Input

CLIME EP25414
Computerised Legal Information Management and Explanation

Answers questions about application of maritime regulations to specific ships (Evans et al, 2002)

- Knowledge source:
  - list of inference rules (instantiations) applying to the situation;
  - Partial ordering of rules according to their legal precedence;
  - Rule applications that would be triggered if some further facts had held.
- User model: Hardwired maritime surveyor
- Discourse history: (representation of the) question for which an answer is being generated.
- Communicative goal: Explain legal assessment outcome (Case)

Output

*I have a new oil tanker. It has a gross tonnage of 2000 tonnes. It is fitted with a fuel oil tank. Ballast water is stored in the fuel oil tank. *

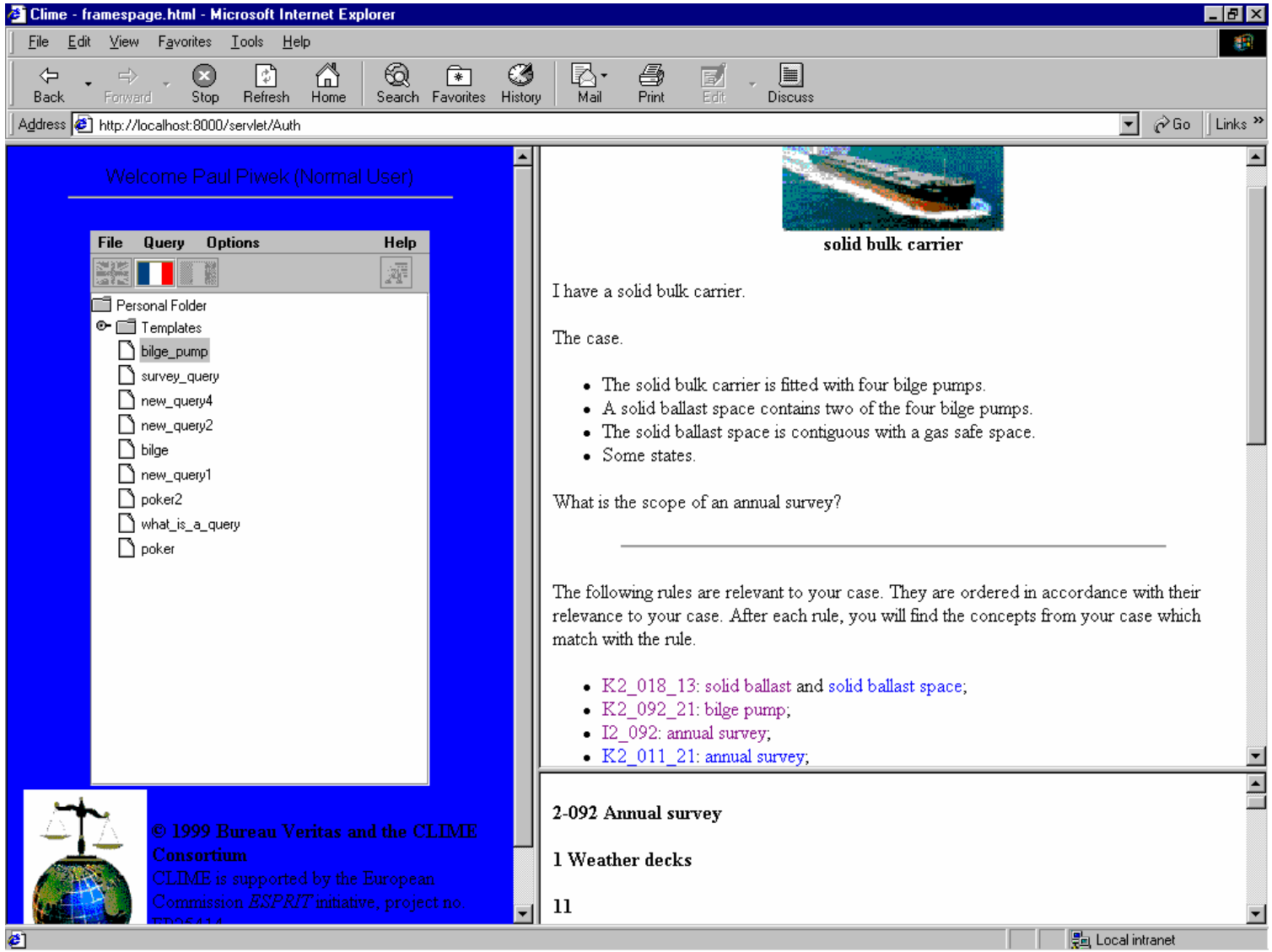
Tell me whether the regulations allow, disallow or are silent about the situation described.

This situation is disallowed. This is because:

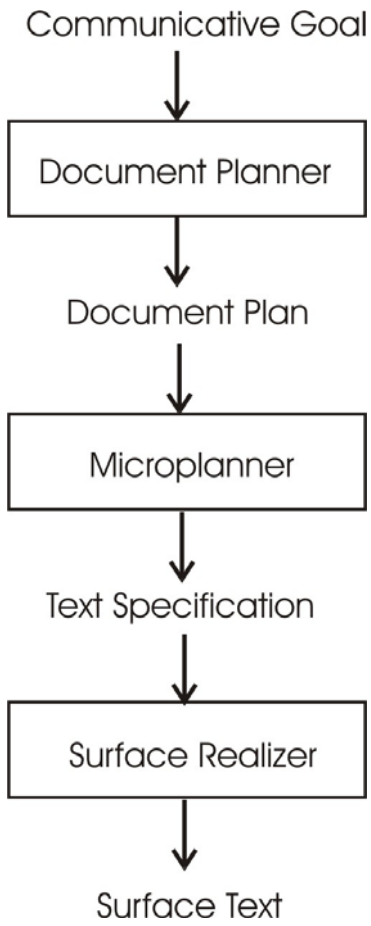
MARPOL AI P2 14 01: the fuel oil tank is part of the new oil tanker, the ballast water is located in the fuel oil tank and the gross tonnage of the new oil tanker is equal to or more than 150 tonnes.

However: If an oily water separating equipment is part of the new fuel oil tanker, then the situation is allowed (*MARPOL AI P2 16 03*).

Output


Clime - framespage.html - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit Discuss
Address http://localhost:8000/servlet/Auth
Go Links
Welcome Paul Piwek (Normal User)
File Query Options Help
Personal Folder
Templates
bilge_pump
survey_query
new_query4
new_query2
bilge
new_query1
poker2
what_is_a_query
poker
© 1999 Bureau Veritas and the CLIME Consortium
CLIME is supported by the European Commission ESPRIT initiative, project no.
solid bulk carrier
I have a solid bulk carrier.
The case.
The solid bulk carrier is fitted with four bilge pumps.
A solid ballast space contains two of the four bilge pumps.
The solid ballast space is contiguous with a gas safe space.
Some states.
What is the scope of an annual survey?
The following rules are relevant to your case. They are ordered in accordance with their relevance to your case. After each rule, you will find the concepts from your case which match with the rule.
K2_018_13: solid ballast and solid ballast space;
K2_092_21: bilge pump;
I2_092: annual survey;
K2_011_21: annual survey;
2-092 Annual survey
1 Weather decks
11
Local intranet

Generation Reference Architecture (Dale & Reiter, 2000)

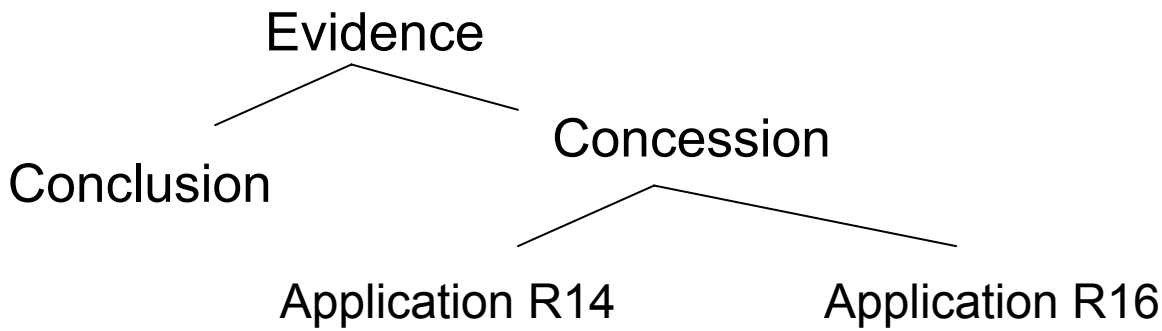


- **Pipeline**: perform tasks in fixed sequence
- **Document Planner**: decisions on content and structure (domain dependent)
- **Microplanner**: specify content and structure left by open by doc. planner; typically language dependent
- **Surface Realizer**: remainder, resulting in concrete text.
- Down the pipeline: more linguistic
- Upstream: More domain specific

Document Planner

- **Content determination**: what should be communicated?
  - Select winning rule (instantiation) and rules which are directly overruled by it.
  - Select rules which can change the assessment.

- **Document structuring**: How to group information (specifically, in rhetorical terms).



Microplanner

- **Lexicalisation**: selecting which words should be used to express the content
- **Referring Expression Generation**: deciding which expressions to use to refer to objects
- **Aggregation**: mapping the (rhetorical structure) produced by the document planner to sentences and paragraphs.

Microplanner: Lexicalization

I have a new oil tanker. It has a gross tonnage of 2000 tonnes. It is fitted with a fuel oil tank. Ballast water is stored in the fuel oil tank. \\

Tell me whether the regulations allow, disallow or are silent about the situation described.

This situation is disallowed. This is because:

MARPOL AI P2 14 01: the fuel oil tank is part of the new oil tanker, the ballast water is located in the fuel oil tank and the gross tonnage of the new oil tanker is equal to or more than 150 tonnes.

However: If an oily water separating equipment is part of the new fuel oil tanker, then the situation is allowed (*MARPOL AI P2 16 03*).

Microplanner: Lexicalization

[[[new-ship, i5], [gross-tonnage,i9], [ballast-water, i17],
[fuel-oil-tank, i16]],
[[measurable,i5, i9], [part-of, i16, i5], [in, i17, i16],
[is-eq-or-more, i9, 150], [oil-tanker, i5]],
[]]

MARPOL AI P2 14 01: the fuel oil tank is part of the new oil tanker, the ballast water is located in the fuel oil tank and the gross tonnage of the new oil tanker is equal to or more than 150 tonnes.

Microplanner: Lexicalization

[[part-of,[i16], [i5], pos],
[in, [i17], [i16], pos],
[is-eq-or-more,[gross-tonnage, [i5]],150,pos]]

[[np([i16]), is part of, np([i5])],
[np([i17]), is located in, np([i16])],
[the, gross tonnage, of, np([i5]), is equal to or more than,150]]

Microplanner: Generation of Referring Expressions

[[the fuel oil tank, is part of, the new oil tanker],
[the ballast water, is located in, the fuel oil tank],
[the, gross tonnage, of, the new oil tanker, is equal to or more than,150]]

Microplanner: Aggregation

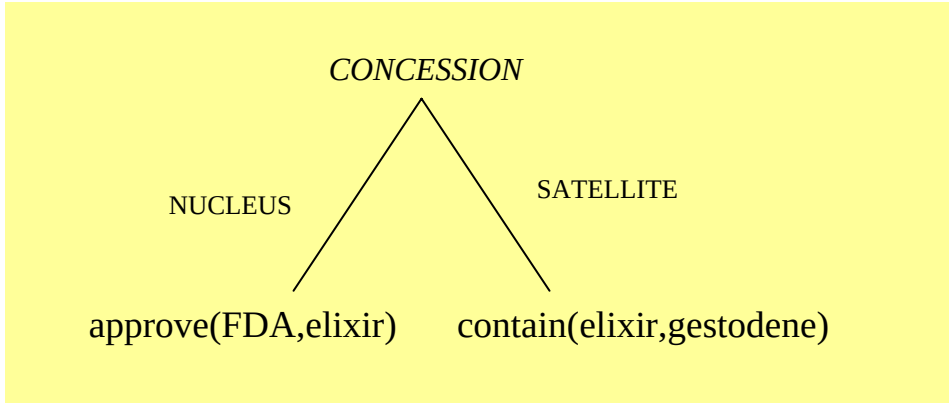
Sentences

[The ballast tank, is part of, the new oil tanker] and [The ballast tank, is not, a segregated ballast tank]

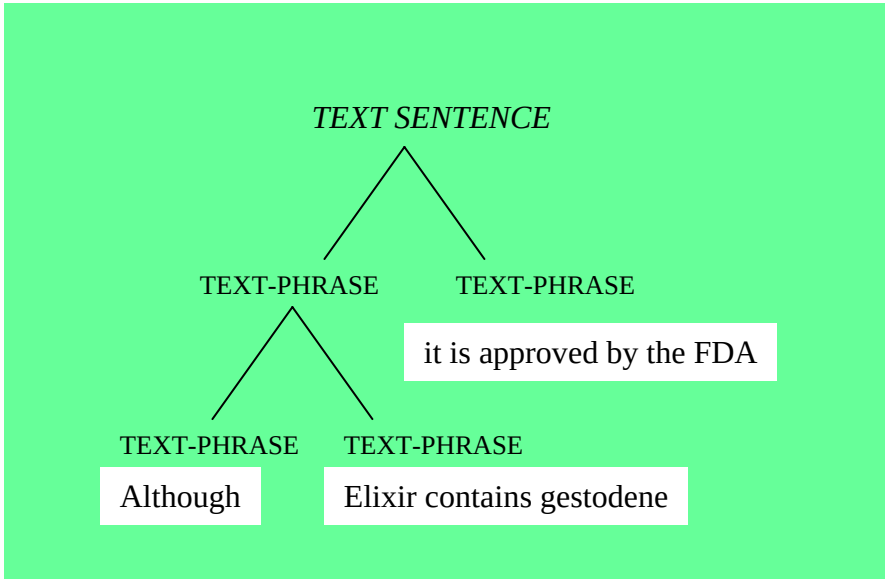
[The ballast tank, is part of, the new oil tanker, and, is not, a segregated ballast tank]

Expressing Discourse relations

Discourse structure

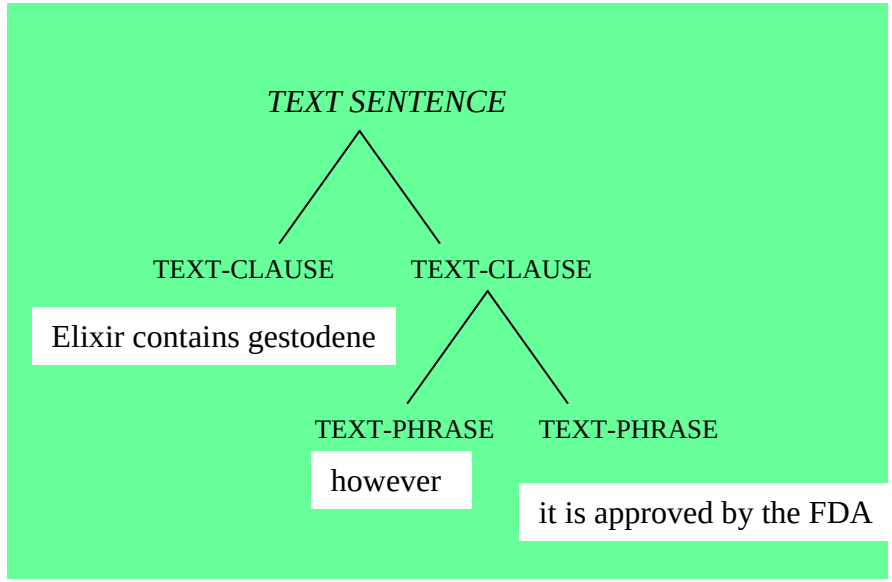


Document structure 1



"Although Elixir contains gestodene, it is approved by the FDA."

Document structure 2



"Elixir contains gestodene; however,it is approved by the FDA."

Microplanner: Aggregation

ICONOCLAST (Power et al., 2003)

- Explores the alternative ways in which a message can be expressed, and investigate reasons why one version might be better, or more suitable, than another;
- Generates drug leaflets;
- A given message is generated in a wide range of different styles.

Surface Realizer

- Linguistic Realization: converting abstract specs of sentences into real text.

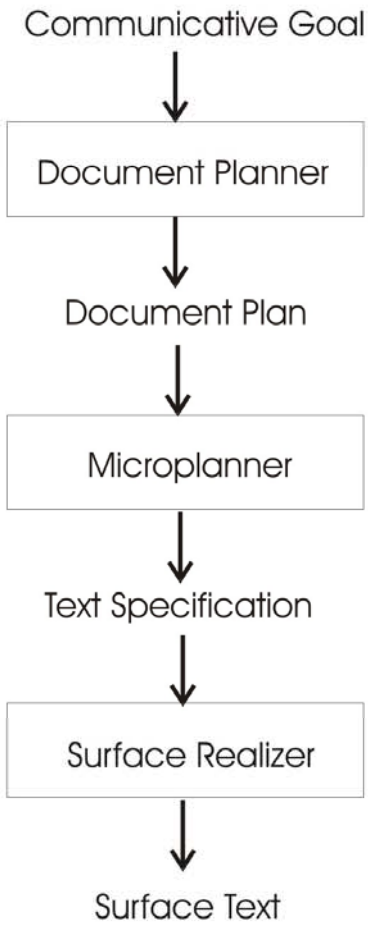
  Subectj: oil tanker & definite
  Predicate: navigate & tense = present & person =3
  Location: Open water & Prep: in
  $\Rightarrow$
  The oil tanker navigates in open waters.

  Available surface realizers: KPML, SURGE, REALPRO

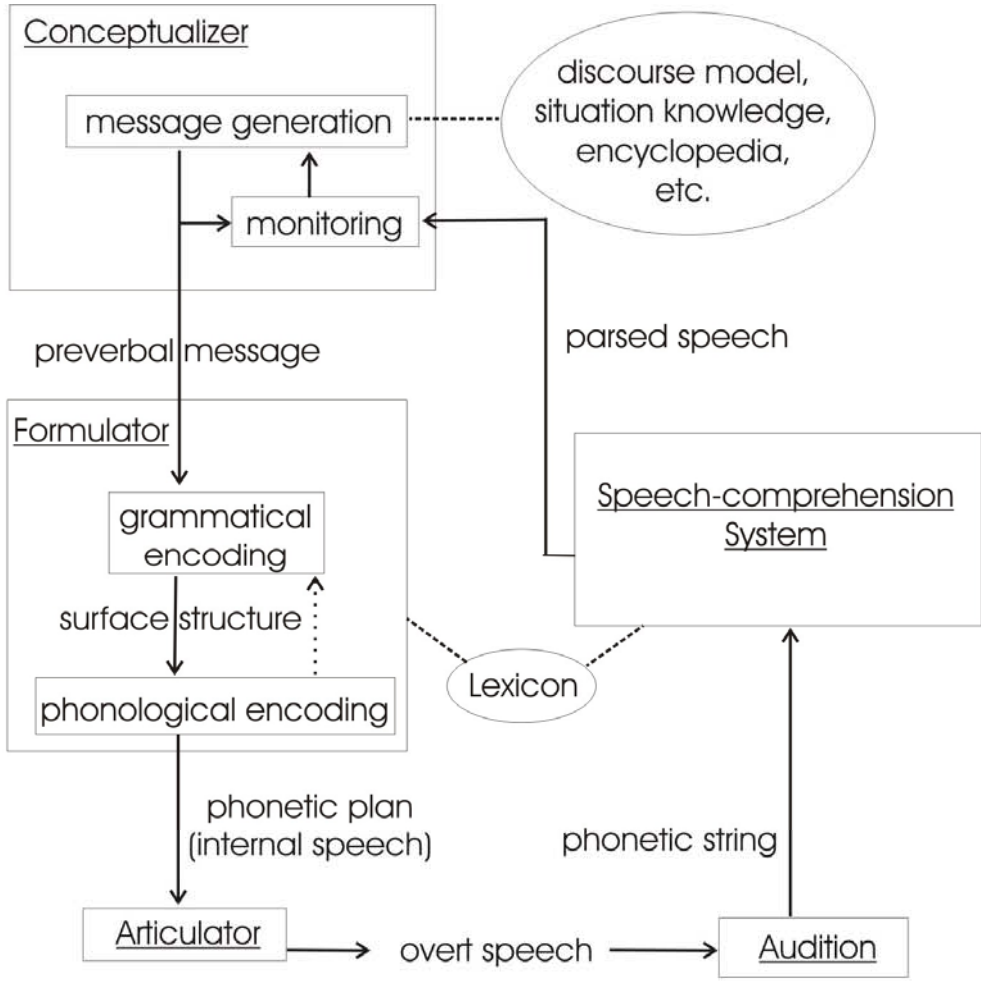
- Structure Realization: mapping abstract representations of structural organization (e.g., paragraphs and sections) into concrete mark-up (e.g., *HTML*) which can be interpreted by the document presentation component.

Other architectures

- The reference architecture is characterized by modularization and sequential processing
- Feedback
- Revision
- Integrated approaches: a single reasoning component, e.g., constraint solver, theorem prover or AI planner. Example: KAMP system (Appelt, 1985).



NLG Architecture (from Reiter & Dale, 2000:60)



Blueprint for the speaker (From Levelt, 1989:9)

Plan

1. Introduction: NLG for Embodied Agents
2. What is Natural Language Generation?
3. Case: Generation of Referring Expressions (GRE)
4. NLG for Embodied Agents: Requirements
5. Case: GRE for Embodied Agents
6. Realization: Audiovisual Speech and Gestures
7. Case: NLG for Teams of Embodied Agents
8. Evaluating NLG for Embodied Agents

Plan

1. Introduction: NLG for Embodied Agents
2. What is Natural Language Generation?
3. Case: Generation of Referring Expressions (GRE)
4. NLG for Embodied Agents: Requirements
5. Case: GRE for Embodied Agents
6. Realization: Audiovisual Speech and Gestures
7. Case: NLG for Teams of Embodied Agents
8. Evaluating NLG for Embodied Agents

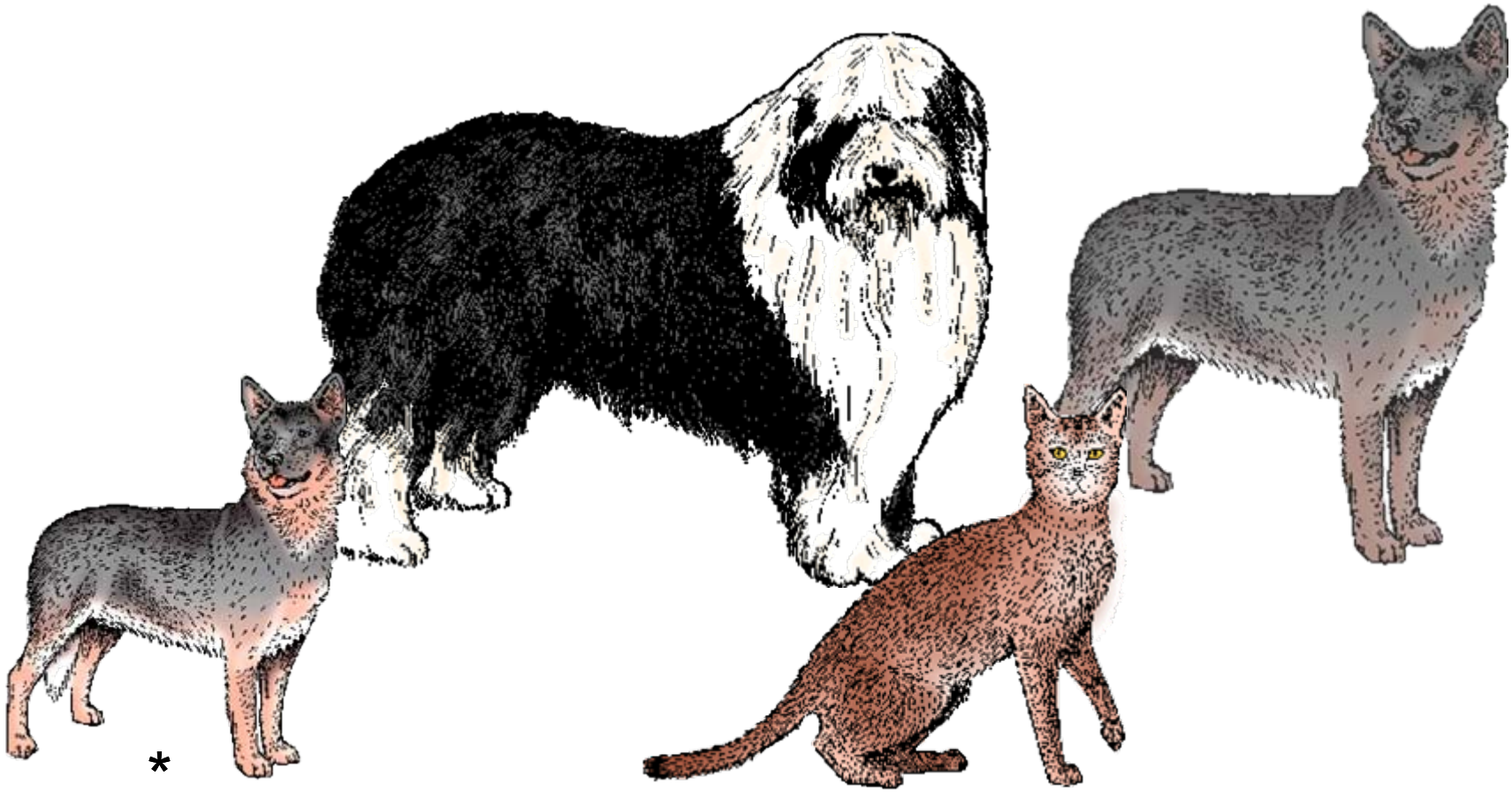
Introduction

- “I’d rather be married to a really rich old guy than to a really old rich guy.”
- Referring expression: linguistic expressions referring to objects
- Referring to objects is one of the most common tasks in NLG.
- Examples: a really old rich guy,
  the black square, the three stooges, it, ...

GRE - Generation of Referring Expression

- Algorithms focus on generation of *distinguishing definite descriptions ("the N").*
- Content Determination Task: given a **target object** *v* and a set of **distractors**, decide which properties **distinguish** the target object from its distractors.
- On the basis of the list of selected properties a distinguishing description can be **realized** in natural language.

Example domain

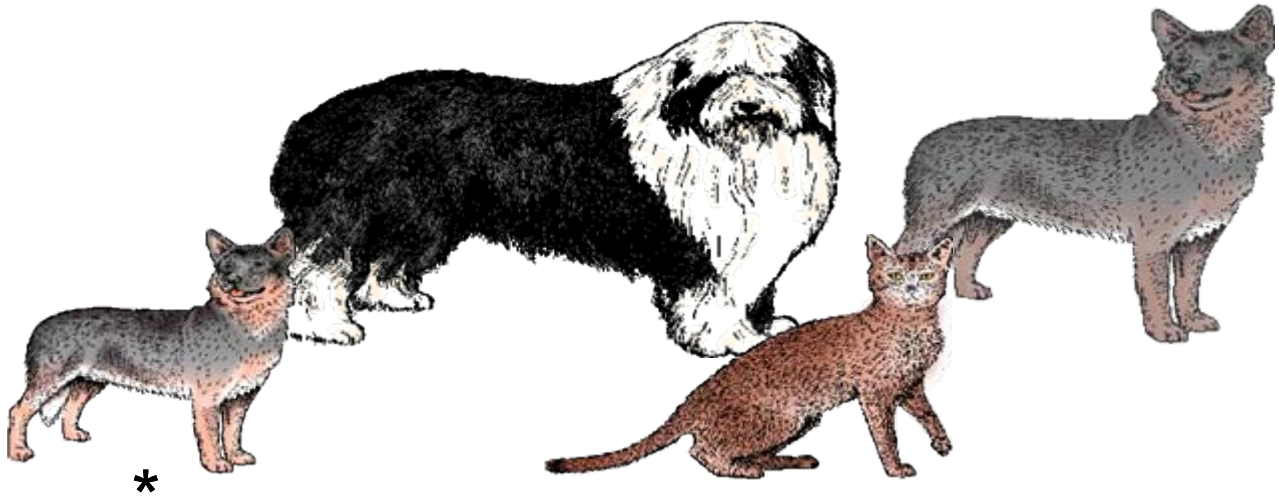


Full brevity generation strategy

- **Full brevity** (Dale 1992, Dale and Reiter 1995): shortest possible description.

- **Strategy**:
  1. Try to generate a distinguishing description with one property.
  2. If this fails, look at all possible combinations of two properties.
  3. *Etc.*

- **Relevant properties:** dog, cat, small, large, brown, black-white



*

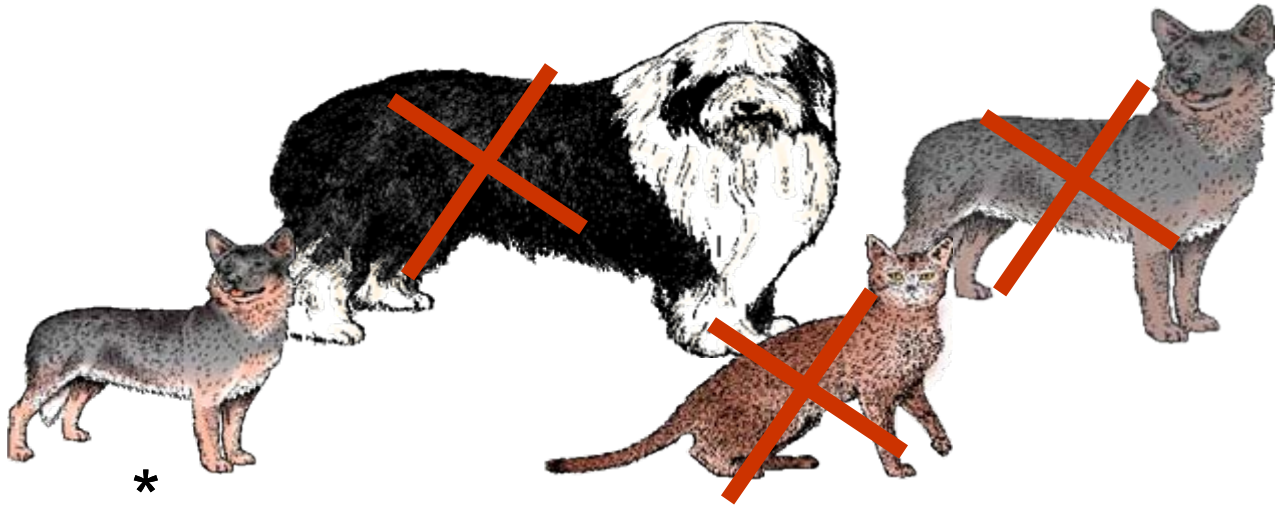
- **Worked example**:
  1. No *single* property rules out all distractors.
  2. Combination of “dog” and “small” does.

- **Output**: “the small dog”

- Full brevity works fine, but:
  1. Minimal descriptions are rare in human communication
  2. Computationally expensive (NP complete)

Incremental Generation Strategy

- Key insight: Human speakers and listeners prefer certain *kinds* of properties (or **attributes**) when describing objects from a given domain.
- E.g., **absolute** properties are preferred over **relative** properties.
- **Preferred attributes (P)**: predefined ordering of attributes for a given domain, e.g., <type, color, size>
- **Incremental algorithm**: iterate through P, add property to description under construction if it rules out at least one remaining distractor.



*

- **Worked example**

  1. <type, dog>
  2. <color, gray>
  3. <size, small>

- **Output**: “the small gray dog”

Incremental Algorithm: discussion

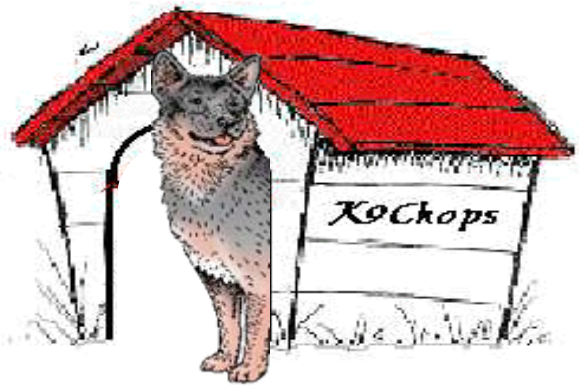
- Incremental algorithm is *de facto* standard for GRE algorithms.

- Advantages:
  1. Empirically motivated
  2. Computationally effective (polynomial); no backtracking

- Disadvantages:
  1. May result in overlong descriptions; not flexible.
  2. Does not cover the entire range of existing descriptions; various (incompatible) extensions have been formulated for descriptions such as “the dogs”, “the gray dog and the two black cats”, “the animals without a flee collar”, etc.
  3. Relational descriptions are difficult to integrate.

Example domain (2)

(Incremental) generation of relational descriptions



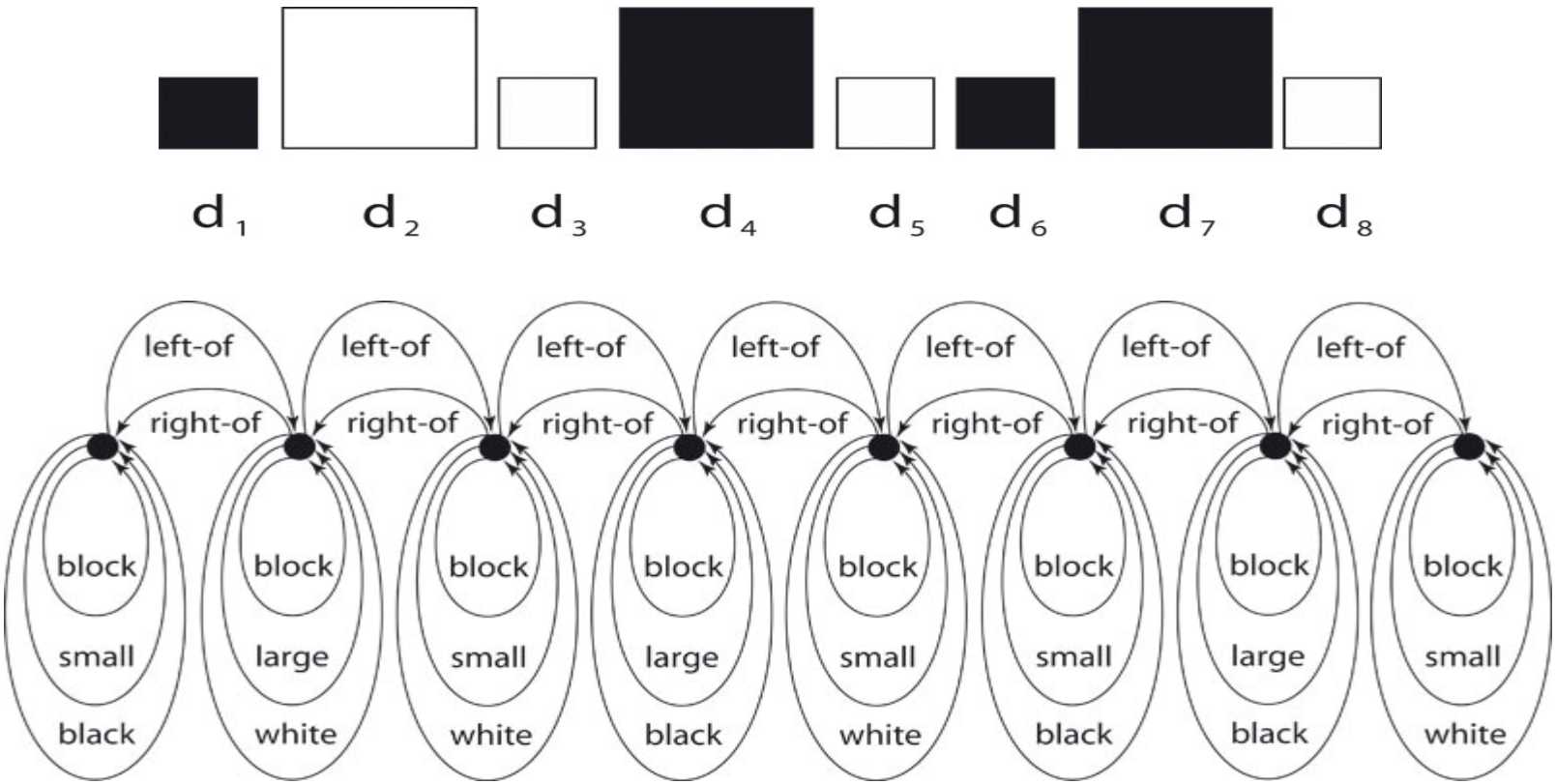
*

- Strategy: first apply algorithm to target, then to its *relatum* (e.g., Dale & Haddock 1991)
- **Problem one**: infinite recursions (the dog in the doghouse containing a dog which is *etc*.).
- **Problem two**: forced incrementality (the dog next to the tree in front of the garage etc.).

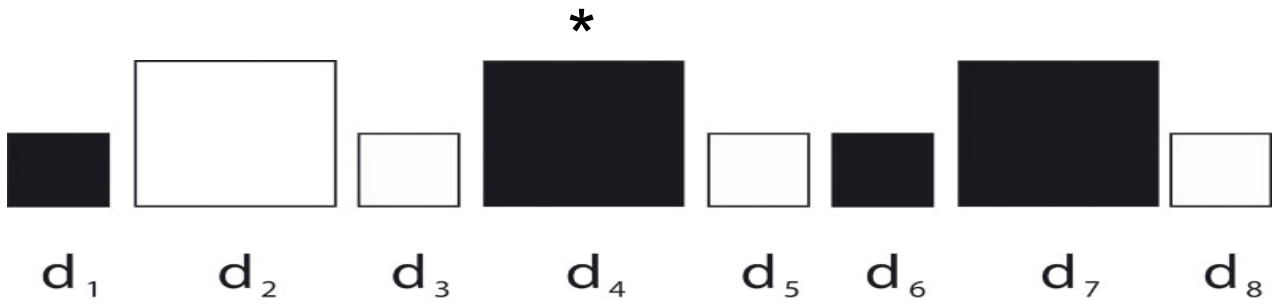
Graph-based GRE (Krahmer et al. 2003)

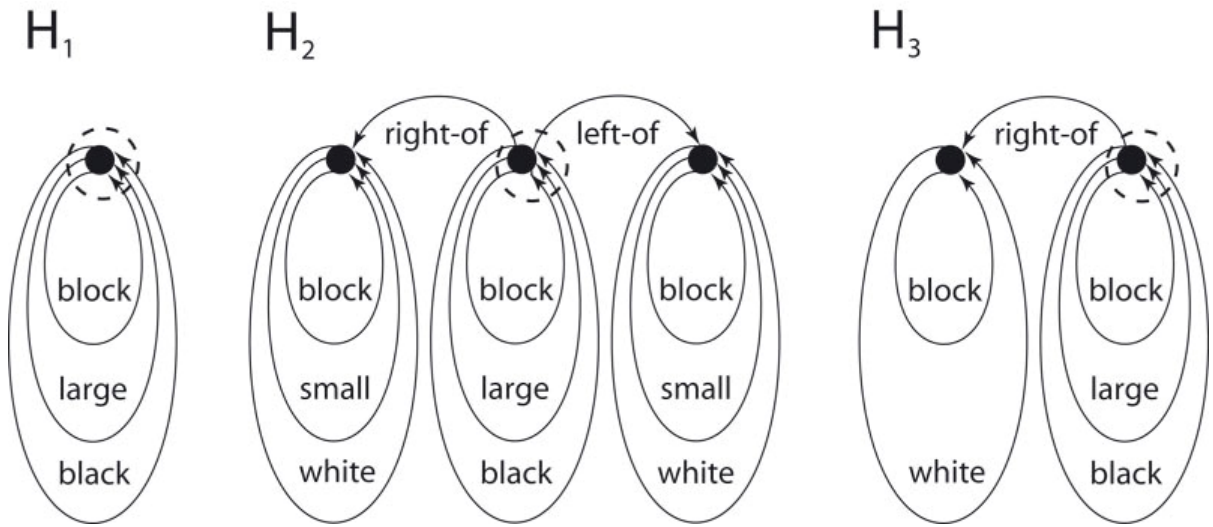
- Formalize domain as a labeled directed graph.
- Content selection becomes a subgraph construction problem.
- Motivations:
  1. Many attractive and well-understood algorithms for dealing with graphs.
  2. Allows for combination and comparison of different extensions of IA.
  3. Clean solution of relational description problem.

A scene and its graph representation



Three referring graphs





Finding referring graphs

- Given a domain graph $G = \langle V_G, E_G \rangle$ and a target node v from V_G, look for subgraph $H = \langle V_H, E_H \rangle$ that can only be "placed over" v.
- Formalized "placed over" notion in terms of **subgraph isomorphisms**.
- H can be "placed over" G iff there is a subgraph G' such that H is isomorphic to G'
- H is **isomorphic** to G' [notation: $H \subseteq_\pi G$] iff there is a bijection π: $V_H \rightarrow V_{G'}$ such that for all vertices $v, w \in V_H$ and all labels l:

$$(v, l, w) \in E_H \leftrightarrow (\pi.v, l, \pi.w) \in E_{G'}$$

Problem formalization

- Given graph *H* and vertex *v* in *H*, and a graph *G* and vertex *w* in *G*, define:
  - *(v,H)* **refers** to *(w,G)* iff (*i*) H is connected, (*ii*) $H \subseteq_\pi G$, and (*iii*) π.v = w
  - *(v,H)* **uniquely refers** to *(w,G)* iff *(v,H)* **refers** to *(w,G)* and there is no w' in G (different from w) such that *(v,H)* **refers** to *(w',G)*
- Task: Given a domain-graph *G* and target vertex *w* in *G*, find a pair *(v,H)* which **uniquely refers** to *(w,G)*.

Cost Functions

- Using **cost functions** to prioritize among different referring graphs.
- **Monotonicity constraint**: adding an edge can never be cheaper
- Cost (H) = Cost (V_H) + Cost (E_H)
- **Option 1**: All edges and nodes cost 1 point **[full brevity]**
- **Option 2**: Cost (type) < Cost (color) < Cost (size) **[incremental]**
- More options exist...

Algorithm (sketch)

- Implemented using a **branch & bound** search strategy (alternative search strategies also possible).

- Input: domain graph $G = <V_G, E_G>$ and a target node v from V_G.

1. Start with graph under construction H, only consisting of v.
2. Systematically extend this graph, adding edges adjacent to v.
3. The algorithm returns the cheapest distinguishing subgraph H of G that refers to v if one exists.

Complexity and Implementation

- **Worst-case complexity**: testing for subgraph isomorphism is NP-complete
- Potential remedies:
  - Define **upper bound** K on number of edges
  - **Planarize** G
- In general: more diverse labeling of scene graph implies quicker solution
- **Implemented** in Java (J2SE, version 1.4)
- Performance evaluation: Krahmer et al. (2003:61)
- E.g., in a graph with 128 vertices and 896 edges distinguishing graphs found on average in 1,023 ms.

Discussion

- GRE is ubiquitous in NLG
- State-of-the-art: Incremental Algorithm (Dale and Reiter 1995)
- Recent alternative: graph-based GRE (Krahmer et al. 2003)
  - Relational descriptions fit in naturally (no principled distinction between (unary) properties and relations)
  - Compatible with different search strategies (“meta-algorithm”)
  - Allows for integration of various alternative extensions of Incremental Algorithm (see van Deemter and Krahmer 2005)
  - Enables multimodal GRE [*more about that in a minute*]

Plan

1. Introduction: NLG for Embodied Agents
2. What is Natural Language Generation?
3. Case: Generation of Referring Expressions (GRE)
4. NLG for Embodied Agents: Requirements
5. Case: GRE for Embodied Agents
6. Realization: Audiovisual Speech and Gestures
7. Case: NLG for Teams of Embodied Agents
8. Evaluating NLG for Embodied Agents

Introduction

- Traditionally, NLG has concentrated on the generation of **text**.
- But generated language for embodied agent requires **audiovisual realization**:
  - Audiovisual speech
  - Prosody
  - Facial Expressions
  - Gestures
- Two potential strategies:
  - **Pipeline**: first generate text, then plan audiovisual realization
  - **Integration**: generate content and audiovisual realization in tandem

Audiovisual speech

- Embodied agents typically integrate some form of speech synthesis (usually, either diphone synthesis or phrase concatenation).

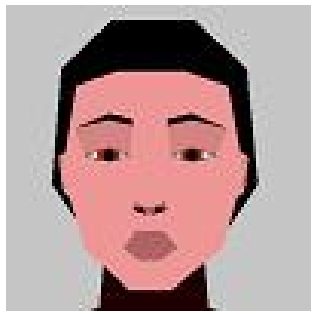
- Visual speech:

  1. Auditory speech consists of **phonemes** (elementary speech sounds).
  2. Visual speech, by analogue, may be based on **visemes** (elementary mouth positions).
  3. Sample speech sound with a certainy frequency (e.g., 40ms).
  4. Display corresponding viseme; intermediate stages via linear interpolation or morphing.

- Complication: /s/ in /soap/ looks different from /s/ in /seed/ (**co-articulation**)

Other audiovisual manipulations

- **Eye movements** ➔ continuously
- **Eye blinks** ➔ every 4.8s on average
- **Movements** ➔ continuously
  - Modelled using Perlin noise mrl.nyu.edu/~perlin/



- *All non-functional / non-communicative*

Audiovisual prosody

- Audiovisual prosody = Not what is said, but *how* it is said...
  - Auditory prosody = intonation, pitch accents, phrasing, voice quality….
  - Visual prosody = facial expressions, arm and hand movements, ….
- *Auditory* prosody is well-studied.
- (Auditory) prosody:
  - Has a big impact on naturalness perception.
  - Facilitates information processing.
  - May even influence meaning (truth-conditions)

Prosody and meaning

- **Pitch accents**:
  - John only introduced Bill to Sue. (Rooth)
  - Ik voel me serieus genomen. (Komrij)
- **Intonational phrasing**:
  - 1 + \\ 2 x 3 [= 7]
  - 1 + 2 \\ x 3 [= 9]

Generating spoken language

- Accent and boundary placement is co-determined by information status and syntactic structure.
- Theune et al. (2001): (auditory) prosody computation is easier in NLG than based on text.
- All relevant information is available within the NLG system:
  - Discourse model: which entities/concepts have been discussed
  - Syntactic information

Prosody computation in NLG

- "I'd **rather** be **married** to a really **rich** old **guy** than to a really old **rich** guy."

- **Discourse model:**
  - New information: pitch accent
  - Given information: no pitch accent
  - Contrastive information: pitch accent

- **Syntactic information**:
  - Where do accents "land" in a particular phrase (e.g., [the blue square]$_F$)
  - Where can intonational phrases occur

Thus...

- Integrated approach seems to work better (better performance evaluations).
- Might also work for visual cues in Embodied Agents (e.g., Cassell et al. 2001, Pelachaud et al. 2001).
- So far, considered *spoken* language.
- What about non-speech output such as gestures?

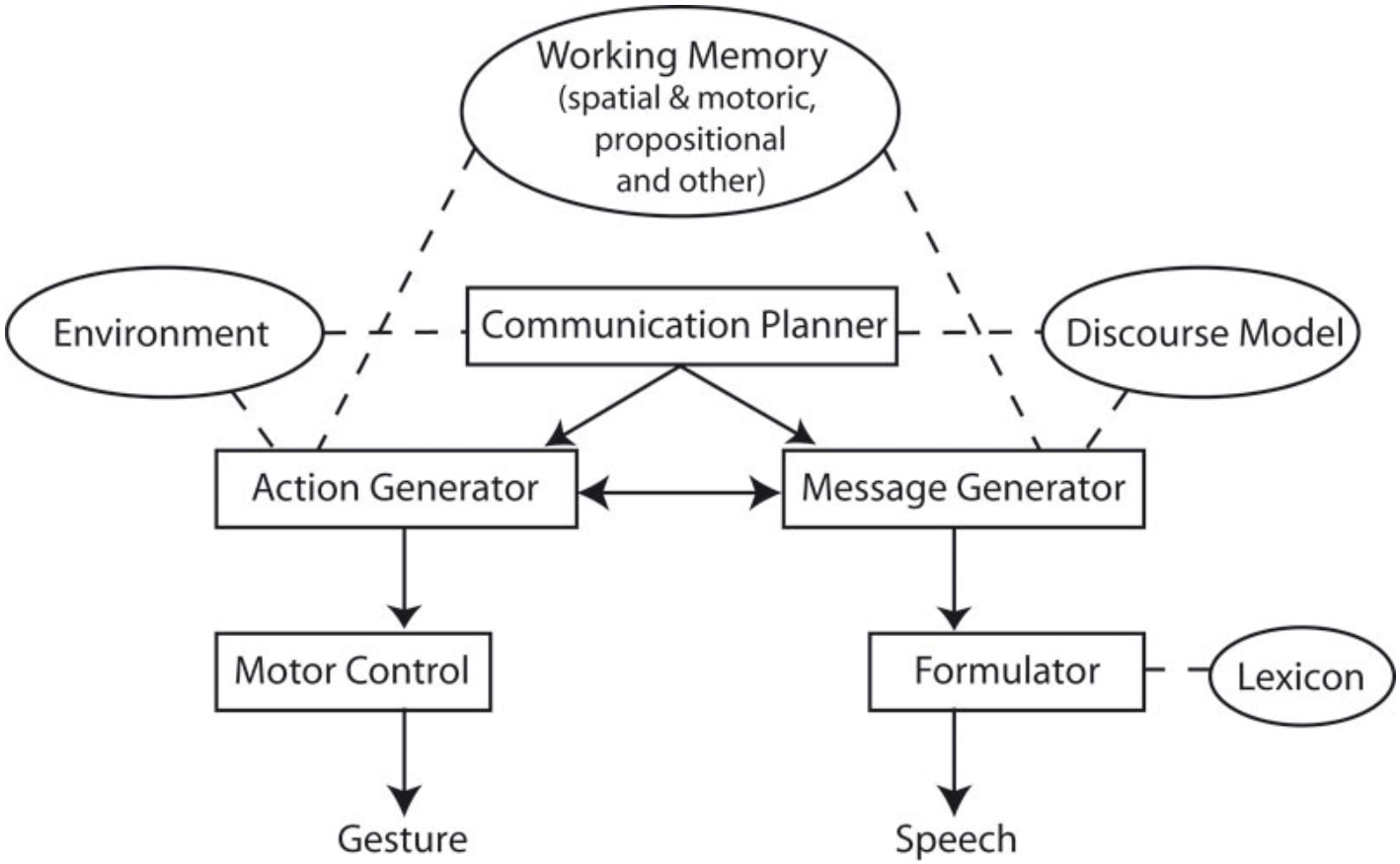
Functions of gesture

- Why do humans gesture?
  - Lexical access [avoid “c”, Rauscher et al. 1996]
  - Thinking [sit on hands, Krauss et al. 2001]
  - For hearer [Alibali et al. 2001]
  - …
- Relation between gesture and language is not well-understood.
- But general consensus is that gesture and speech are closely intertwined (e.g., Kendon 1994, McNeill 1992).

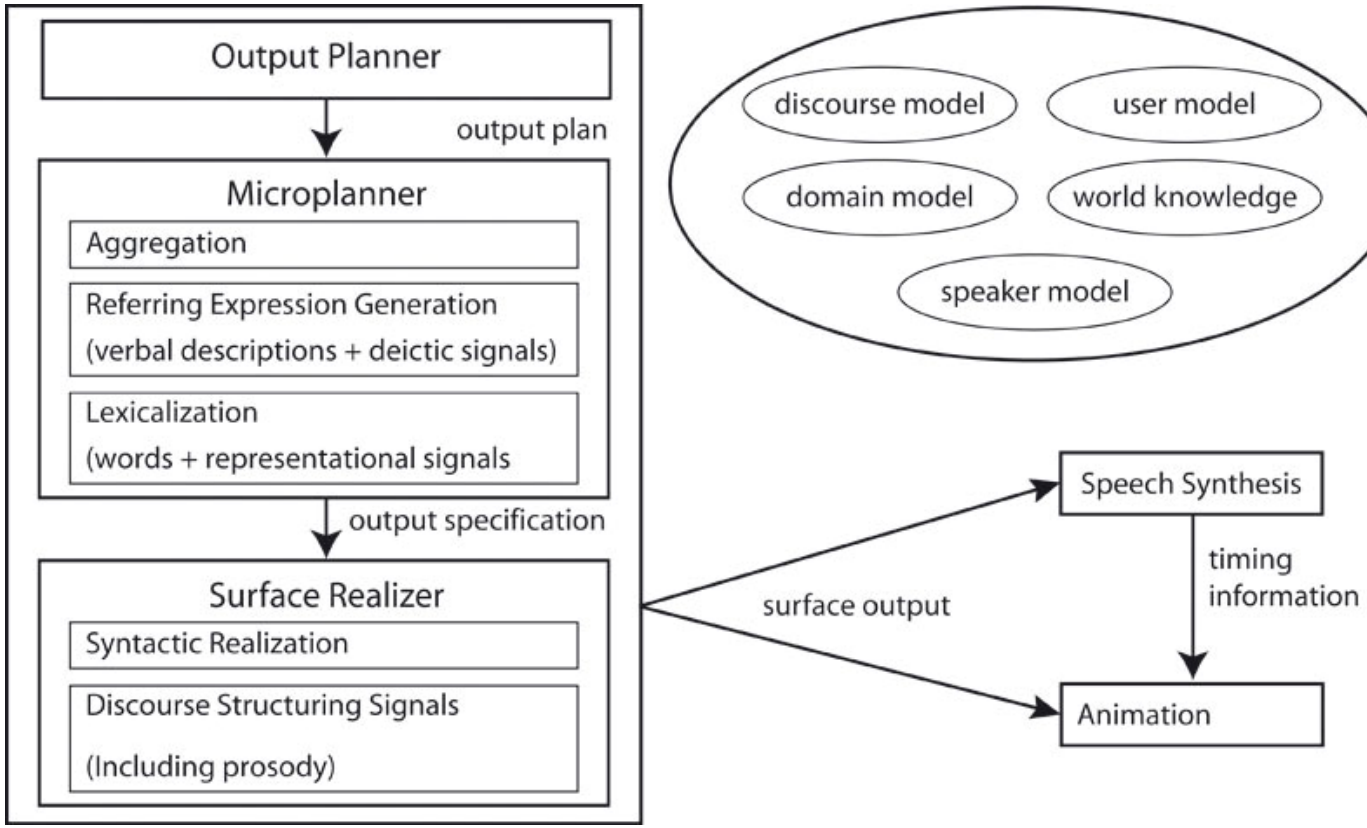
Speech and gesture

- Question: how closely?
- Various extensions of Levelts' 1989 model including gesture stream:
  - McNeill and Duncan (2000): no seperate stream, speech and gesture arise simultaneously from "growth points"
  - Krauss et al (1996): gestures arise in working memory
  - De Ruiter (2000): gestures arise in the conceptualizer ("sketch")
  - Kita & Őzyürek (2003): gestures arise after planning

Multimodal speech production (Kita and Özyürek 2003)



Multimodal NLG architecture (Theune et al. 2005)



Discussion (1)

- NLG for embodied agents requires special care:
  - Audiovisual speech
  - Audiovisual prosody
  - Gestures
- Integrated models (co-generation of speech and gesture and audiovisual prosody) seem to work best, both for modelling human speech production and for speech production for embodied agents.
- Optimal integration still an open issue (in both cases).

Discussion (2): different gestures

- Different *kinds* of gestures might have different functions as well as different sources in speech production models

- Kinds of gestures (e.g., McNeill 1992):

  - Beat gestures (“flick of the hand”) [aka batons]
  - Representational gestures
    - Iconic: related to semantic content (shape, motion, etc.)
    - Deictic: pointing gestures

Plan

1. Introduction: NLG for Embodied Agents
2. What is Natural Language Generation?
3. Case: Generation of Referring Expressions (GRE)
4. NLG for Embodied Agents: Requirements
5. Case: GRE for Embodied Agents
6. Realization: Audiovisual Speech and Gestures
7. Case: NLG for Teams of Embodied Agents
8. Evaluating NLG for Embodied Agents

Introduction

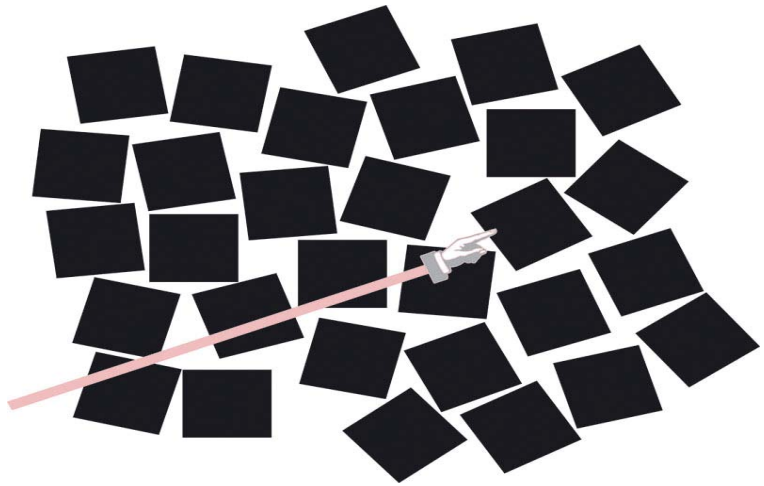
- GRE for embodied agent: combination of spoken language and (deictic) gesture.

- Motivating example application: Smartkom HomeOffice

Multimodal referring expressions

- Here: natural language referring expression which may include a deictic pointing gesture.

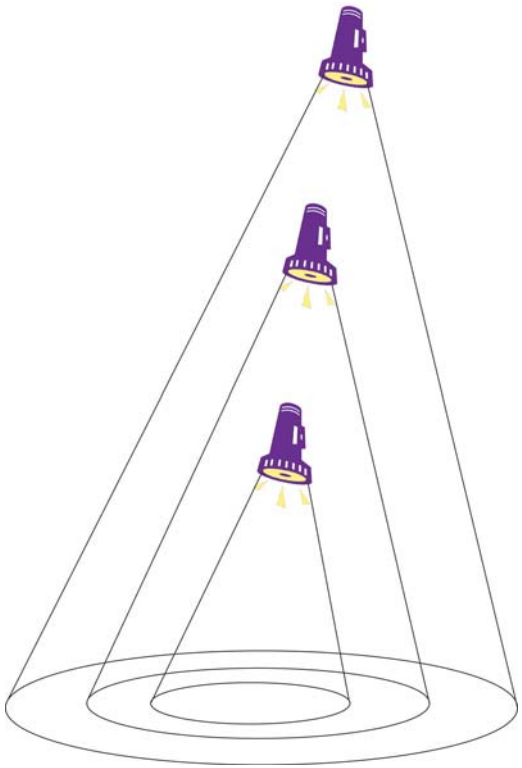
- Advantages:
  - In human-human communication multimodal referring expressions are rather common.
  - Sometimes a purely linguistic description is too complex.



Earlier work

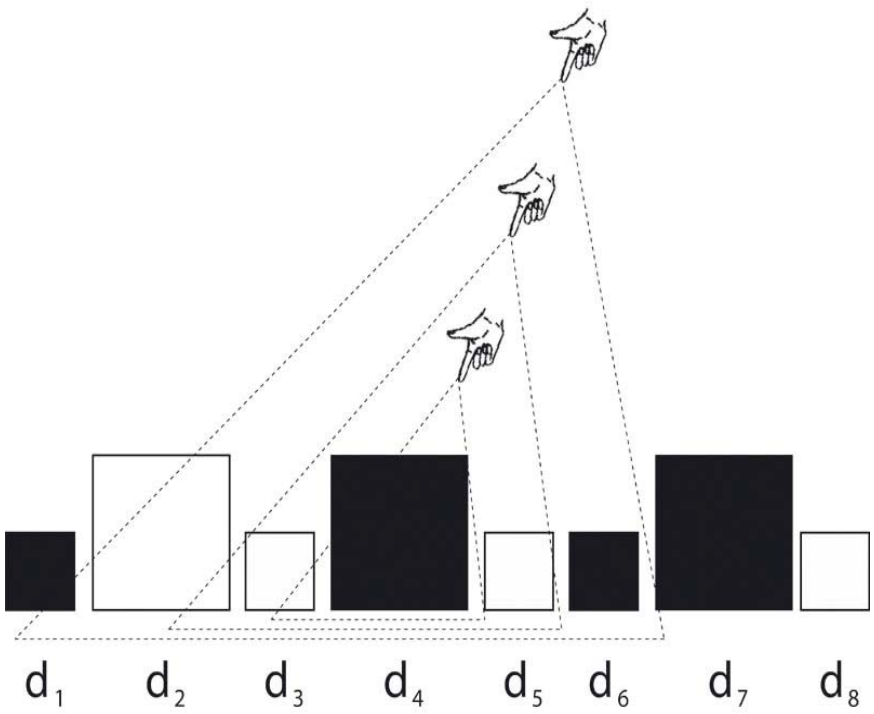
- Cohen 1984, Claassen 1992, Huls et al. 1995, Andre & Rist 1996, Lester et al. 1999, among others.
- Pointing is always precise and unambiguous (results in relatively simple descriptions, e.g., *this block*).
- Decision to point based on simple, context-independent criterion, for instance:
  - Claassen (1992): always point when no distinguishing linguistic description can be found
  - Lester et al. (1999): always point when pronominal reference (“it”) is possible.

Alternative



- Decision to point based on a trade-off between costs of pointing and costs of linguistic properties.
- Allow for gradations in pointing precision.

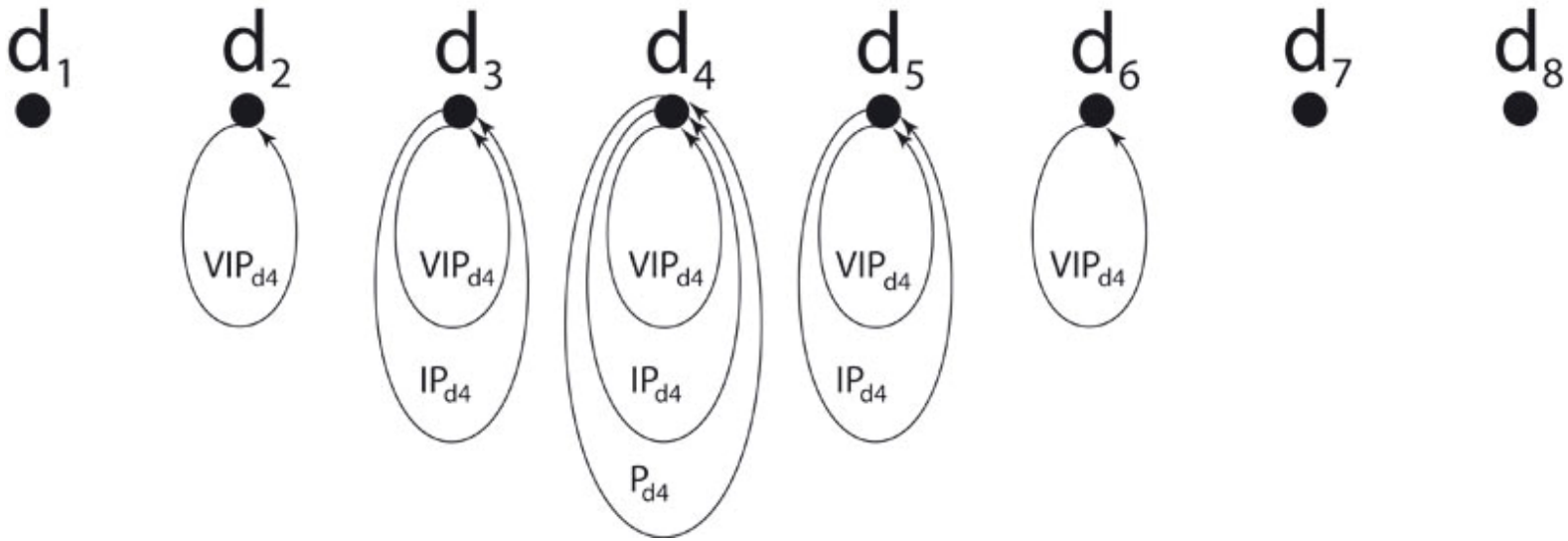
Flashlight model for pointing (Krahmer & van der Sluis 2003)



- Prediction: amount and kind of linguistic properties co-vary with kind of pointing gesture.
- Assumption: precise pointing is more ‘expensive’ than imprecise pointing.
- Neurological evidence: Smyth and Wing 1984, Bizzi and Mussa-Ivalid 1990.

Multimodal graphs

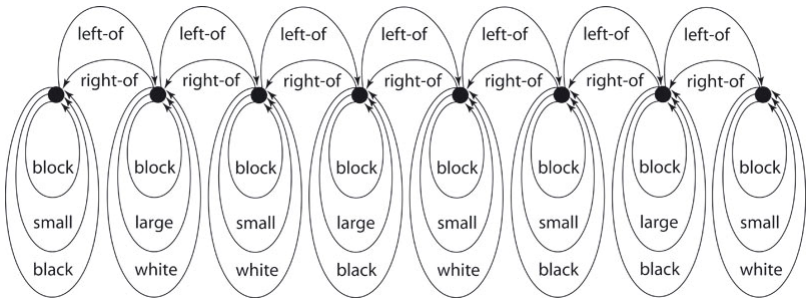
- Idea: potential pointing gestures can be represented as edges in a gesture graph.

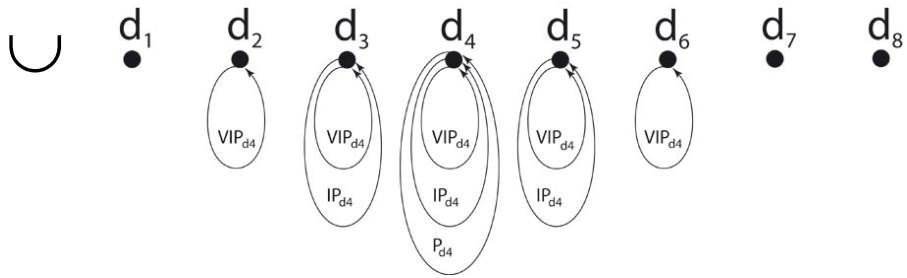


Finding multimodal referring graphs

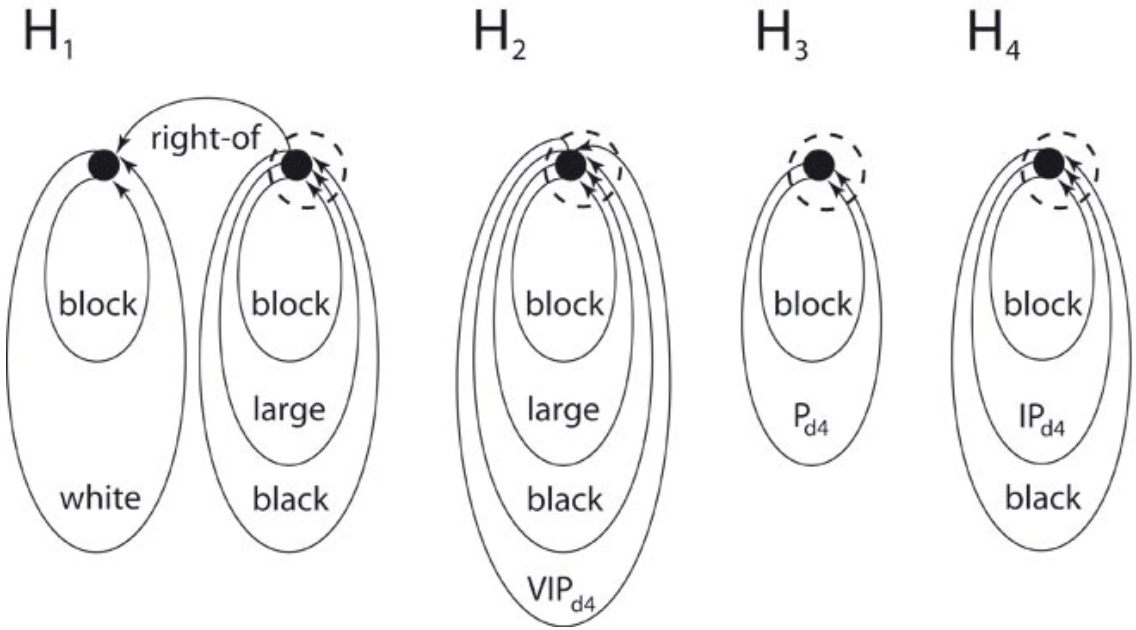
- Selecting a precise pointing edge rules out all distractors, a very imprecise one rules out peripheral objects.

- Merge domain graph (fixed) with gesture graph (variable) and look for referring graphs using graph-based GRE algorithm.





Some multimodal referring graphs

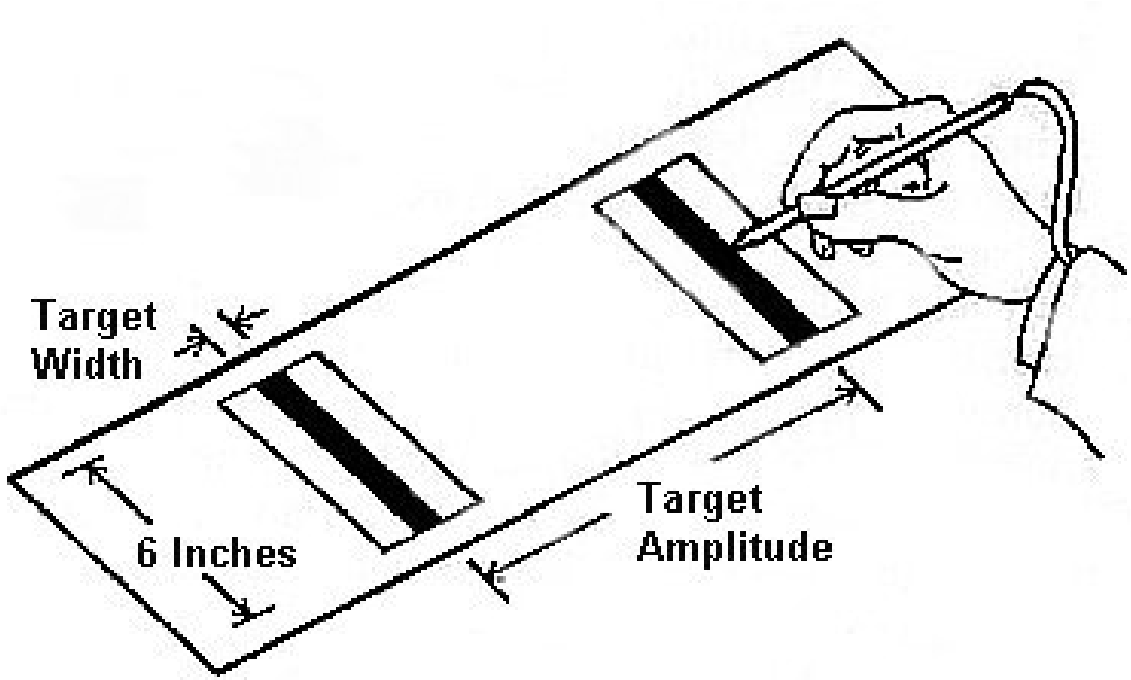


- Q: Which one to select?
  A: The cheapest...

Costs

- ...of **linguistic edges**. As before...
- ...of **pointing edges**. Arguably, dependent on two factors:
  - Size S of the target object (bigger is easier).
  - Distance D of hand to target object (closer is easier).
- Determine costs in terms of Fitts' law (1954).

Fitts (1954)



[source: MacKenzie (1992)]

The Index of Difficulty

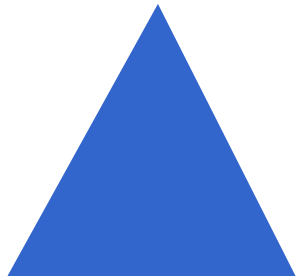
- Fitts' Index of Difficulty

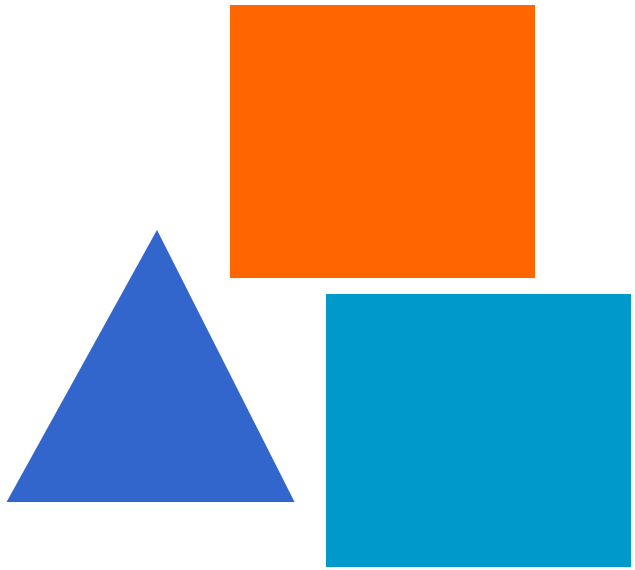
$$ID = \log_2 (D/S + 1)$$

- D = Distance [Amplitude]; S = Size [Width]
- ID is good predictor of MT (movement time)
- Reinterpretation: D = distance from current position to target position of the hand [VIP is closer, hence cheaper]

Evaluation (van der Sluis and Krahmer 2004)

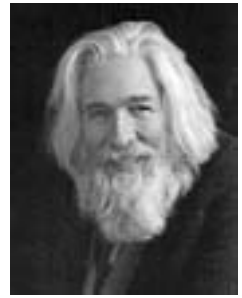
- Production experiment
- Participants believed they were testing a “digital stick”
- Two conditions: near and far
- Task: object identification
- Two kinds of targets: colored geometrical figures and people





TILBURG UNIVERSITY

The Open University



Results

	Near	Far
Object	0.3	2.24
Person	0.2	2.60

Average number of properties as a function of distance and target.

More results

- Persons (vs objects):
  - More effort (‘uh’, ‘um’)
  - More location phrases (“top-left”) in far condition
  - More redundant properties in far condition

- Second experiment in which pointing was not forced: confirmation of these general results.

Discussion

- Model for multimodal referring expressions based on a few, independently motivated assumptions:
  - Foundation: graph-based generation (Krahmer et al. 2003).
  - Linguistic properties have certain costs (Dale and Reiter 1995).
  - Flashlight model for pointing, costs derived from Fitts' law (Fitts 1954).
- No a priori pointing criterion: trade-off between costs of linguistic and pointing edges.
- Integrated approach works best.
- Generalize to other (iconic) gestures?

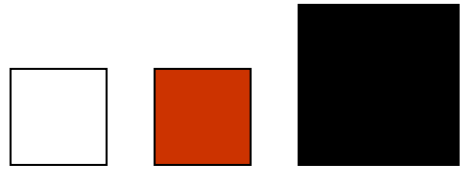
Plan

1. Introduction: NLG for Embodied Agents
2. What is Natural Language Generation?
3. Case: Generation of Referring Expressions (GRE)
4. NLG for Embodied Agents: Requirements
5. Case: GRE for Embodied Agents
6. Realization: Audiovisual Speech and Gestures
7. Case: NLG for Teams of Embodied Agents
8. Evaluating NLG for Embodied Agents

Motivating example

- Consider:

  Q: Which square is the largest?
  A: The black square



- **Question 1**: Which word in answer “the black square” is the most important (most informative)?

- **Question 2**: How do human speakers signal the importance of this word?

About brows

- Speakers of Germanic languages (Dutch, English, …) may use pitch accents to signal **prominence** of words: *in focus*.
- Rapid eyebrow movements (*flashes*) may play similar role (e.g, Birdwhistell 1970, Condon 1976).
- Ekman (1979), *About brows*:
  - Emotional: slow, well-understood
  - Conversational: rapid, understudied
- Link with (verbal) prosody (1): *Metaphor of up and down* (Bolinger 1985:202ff).
- Link with (verbal) prosody (2): BLUE square.

More about brows

- Early empirical study: Cavé et al. (1996). They found significant correlation between F_0 (pitch) and (*left*) eyebrow.
- Left-sidedness confirmed by Keating et al. (2003), Schmidt and Cohn (2001) and [*interestingly*] Pennock et al. (1999)
- Certainly no one-to-one mapping between pitch and eyebrows.
- **Talking Heads**: no consensus about timing / placement of flashes.
- [I know that Harry prefers POTATO chips, but what does JULIA prefer?]

(JULIA prefers)$_{theme}$ (POPCORN)$_{rheme}$ Pelachaud et al. (1996)
(JULIA prefers)$_{theme}$ (POPCORN)$_{rheme}$ Cassell et al. (2001)

Research questions

- What *is* the contribution of eyebrows for the perception of focus?
- Are there other visual cues for focus as well?
- What is the precise relation between auditory and visual cues?

Collection of speech materials

- [Swerts, Krahmer, Avesani (2002), *Journal of Phonetics*]
- Subjects: 8 Dutch speakers [and 8 Italian ones from Tuscany].
- Played a dialogue game (“kwartetten”), cards with colored geometrical figures.
- Target utterances: *blauw vierkant.*
- **Define**:
  - A property is **given** (G) if it was mentioned in the previous turn.
  - A property is **contrastive** (C) if the object described in previous turn had a different value for the relevant attribute. [contrastive ➔ in focus]

Contexts

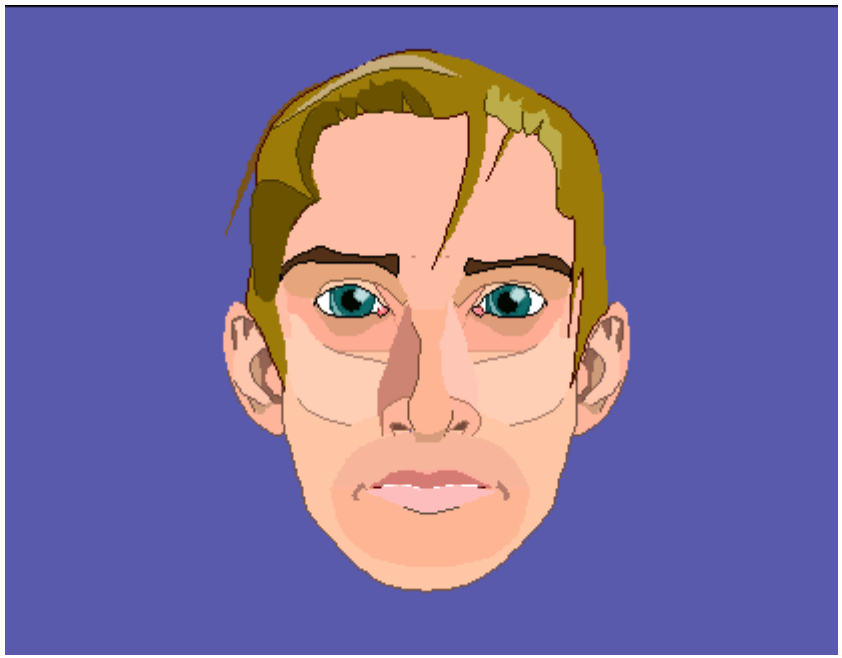
CC	A: rode cirkel B: blauwe vierkant
CG	A: rode vierkant B: blauwe vierkant
GC	A: blauwe driehoek B: blauw vierkant

Distributional analysis: contrastive (C) words receive a pitch accent, while given (G) words do not.

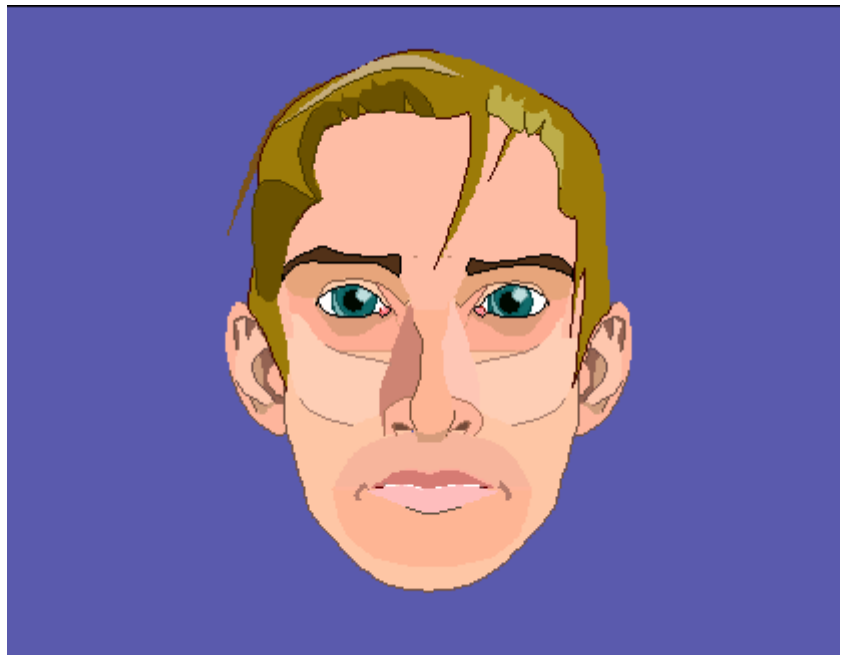
Talking Heads Stimuli

- Dutch: 6 voices (4 human / 2 synthetic)
- Animations made with CharToon environment (Ruttkay et al. 1999).
- Brows: 100ms up, 100ms raised, 100ms down.
- Duration is comparable to human flashes (±375ms).
- Placement: either on first or on the second word.
- *Nota bene*: some stimuli are 'incongruent' e.g., C<u>G</u> (pitch accent on first, flash on second word)

Example stimuli (Dutch)



CG



CG

Pretest: Preferences and Prominence

- Both eyebrow movements and pitch accents clearly perceivable.
- Dutch subjects generally prefer eyebrow movement on accented word. (Hence, disprefer incongruent stimuli !)
- Prominence:
  - presence of eyebrow movement **boosts** prominence of word,
  - and **downscales** prominence of surrounding words.
- So, eyebrows are perceivable and influence prominence. But are they functional?

Experiment: Functional analysis

- Method: *dialogue reconstruction*. Subjects hear and see “blauw vierkant” and have to determine whether preceding figure was

  1. Red triangle (CC)
  2. Red square (CG)
  3. Blue triangle (GC)

- Three prosodic contexts and 3 eyebrow conditions:
  CC, CC, CG, CG, GC, GC

- 25 subjects, native speakers of Dutch

Results

context	Focus perceived on blue	square	both
CC	.30	.27	.43
CC	.14	.47	.39
CG	.75	.15	.10
CG	.70	.20	.10
GC	.17	.61	.22
GC	.18	.60	.22

Discussion

- Both auditory cues (pitch accents) and visual cues (eyebrow movements) influence focus perception, but effect is much larger for auditory cues.
- Visual cues mainly contribute when auditory cues are unclear (CC).
- Advantages of analysis-by-synthesis method:
  - Full control over the relevant parameters.
  - Can directly apply findings in a Talking Head.
- Potential disadvantage: results may be incomplete. Other non-verbal cues for prominence? Nods? Gestures?
- Remedy: Supplement analysis-by-synthesis with *analysis-by-observation*.

TILBURG UNIVERSITY

The Open University

Collection of audio-visual stimuli

- 20 subjects utter CV nonsense words.
- /ma ma ma/ and /ga ga ga/.
- One syllable marked for promince (e.g., /ma MA ma/).
- Two conditions: neutral and exaggerated.

TILBURG UNIVERSITY

The Open University

Example stimuli (congruent)

/ma ma ma/

/ga ga ga/

Experiment: Perception of congruent stimuli

- Selected five speakers, and offered their utterances to subjects, in three conditions.
- 15 subjects per condition determined most prominent syllable.
- Results:

Condition	Correct %
Sound and Vision	97.11
Sound	97.33
Vision	92.89

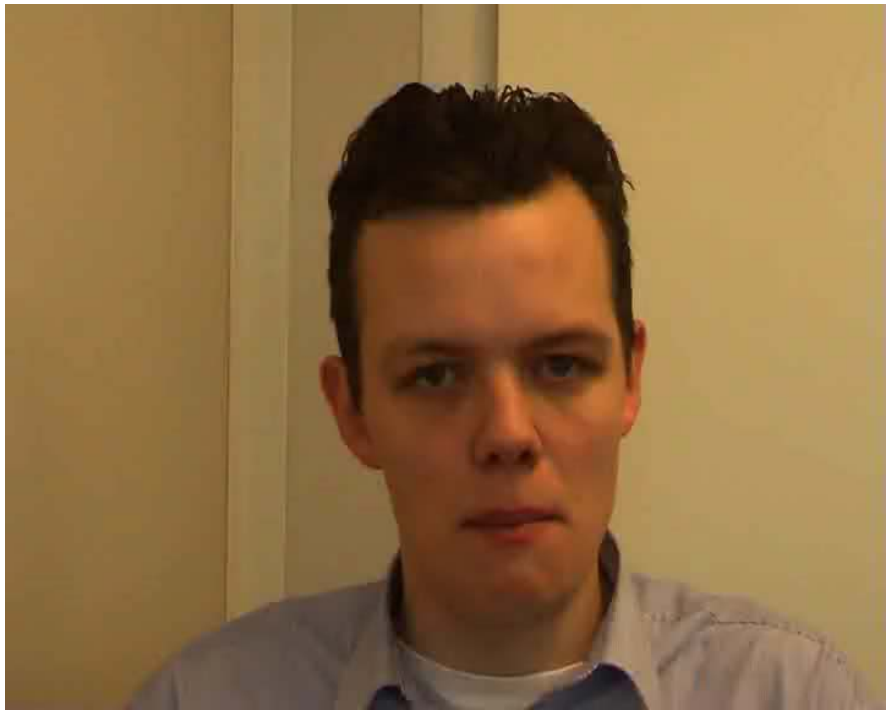
Discussion

- Slightly more errors for /ga/ than for /ma/ (n.s.).
- Slightly more errors for normal than for exaggerated (n.s.)
- Most surprising result: vision only.
- But: **Ceiling effect**…
- Second perception test to tease differences between auditive and visual cues more apart.
- Using *incongruent* (mixed) stimuli.

TILBURG UNIVERSITY

The Open University

Example stimuli (incongruent)



Results

Auditory accent	Visual accent	Perceived accent 1	2	3
1	2	.92	.07	.01
1	3	.92	.00	.08
2	1	.20	.78	.01
2	3	.00	.92	.08
3	1	.37	.03	.60
3	2	.00	.23	.77

Discussion

- Visual analysis of cues: eyebrows, but *also* nods, mouth opening, ...
- Perception results in line with analysis by synthesis study
  - Auditive cues are dominant
  - Incongruent stimuli lead to much more confusion
- The clearer the visual cue, the more confusion arises.
- Disadvantages:
  - Nonsense syllables...
  - Possible mixing artefact compared to congruent stimuli
  - No insight in processing of (in)congruent stimuli

Reaction times experiment

- Collection of material:
  - 6 speakers
  - Stimulus: “Maarten gaat maandag naar Mali”
  - 4 conditions: (1) no accent, (2) Maarten, (3) Maandag, (4) Mali
- Stimuli
  - Both congruent *and* incongruent were systematically mixed
  - Auditory accent (4 conditions) x Visual accent (4 conditions)
  - Yields 16 stimuli per speaker.
- 40 subjects (right handed, age between 18 – 35)
- RT tool: Pamar

TILBURG UNIVERSITY

The Open University

Example stimuli (congruent)

Example stimuli (incongruent)

Results

- Confirming earlier experiments:
  - Auditory accent is strongest cue for perceived accent
  - Visual accent contributes most when auditory information is poor.

- Incongruencies confuse (signigicantly longer RTs), e.g., for

auditory accent = perceived accent = Mali

Visual accent	RT (ms)
-	218
Maarten	290
Maandag	311
Mali	**109**

Discussion

- Congruent stimuli are processed quicker than incongruent ones (especially for first and third word).
- In general: confirmation of results from earlier two experiments, plus:
  - 'Natural' stimuli
  - All stimuli mixed
  - Includes RT to measure 'confusion'
- So far: how does visual information, combined with speech, influence perception.
- Question: does visual information *directly* influence speech.

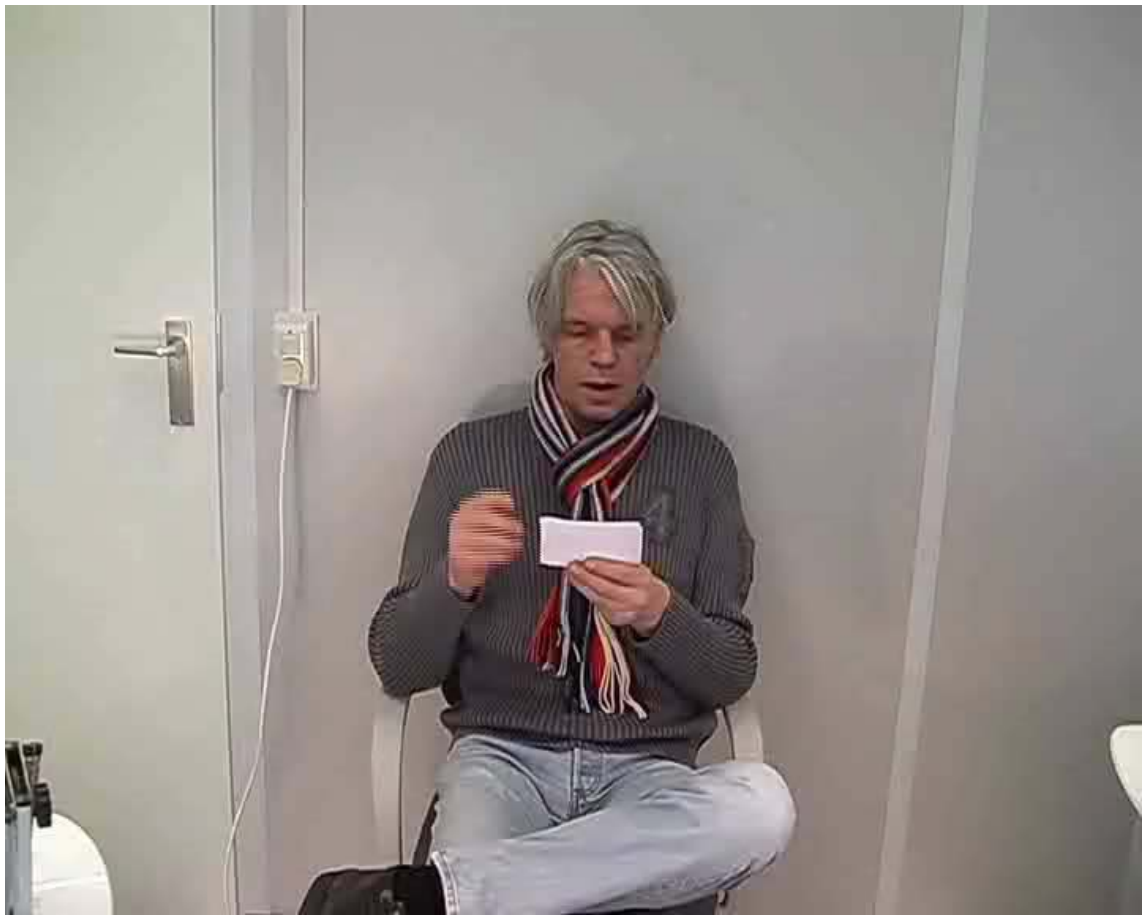
Hearing gestures?

- Collection of materials
  - 11 speakers
  - Stimulus: “Amanda gaat naar Malta”

- Conditions on instruction cards:
  - Auditory accent on Amanda / Malta / neither
  - Visual cue on Amanda / Malta / neither
    - Head nod
    - Beat gesture
    - Eyebrow

- Two rounds of 20 cards, subjects try as often as they like.

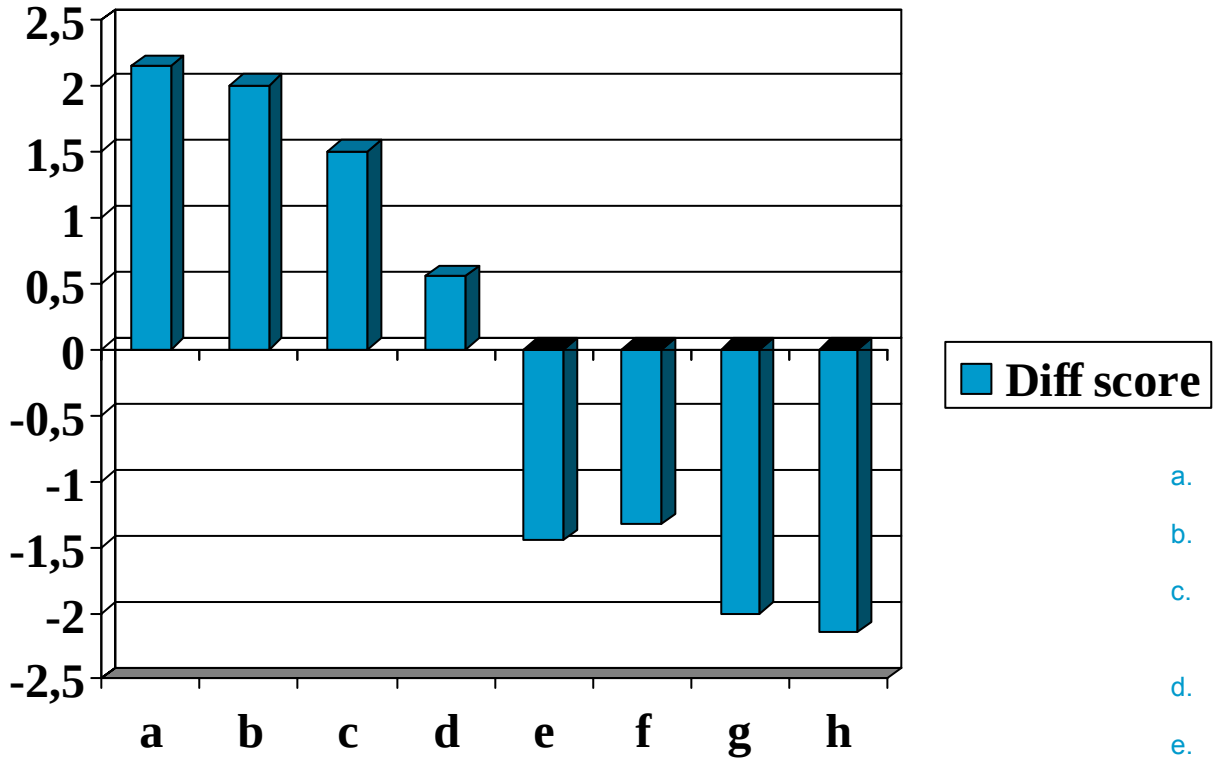
Example



Data processing

- All 440 utterances (20 x 2 x 11) were labelled by 3 persons (only speech, blind for condition).
- For both ‘Amanda’ and ‘Malta’ labelled accents from 0 (no accent) to 2 (clear accent).
- Good correlations between labellers (avg. r = .67)
- Summed 3 accent scores (gives range from 0 to 6)
- Computed **difference score** via *Amanda*-score minus *Malta*-score (gives range from 6 to -6)

Results (*beat gesture*)



a. = accent + b-gest Amanda
b. = accent Amanda
c. = accent Amanda + b-gest Malta
d. = b-gest Amanda
e. = b-gest Malta
f. = accent Malta + b-gest Amanda
g. = accent Malta
h. = accent + b-gest Malta

Discussion

- Both accent and beat gesture have significant effect on diff. score.
- Same for eyebrows and (to a lesser extent) .
- In sum: head nods, beats and eyebrows directly influence speech production.
- [Perhaps due to coordination, e.g., Turvey 1990]
- They can be *heard*...

Perceptual relevance

- *Seeing* a speaker leads to increased prominence (in particular loudness), Grant & Seitz 2000
- Perception study, stimuli with and without corresponding visual images.
- *Seeing* a gesture and (to a lesser extent) an eyebrow leads to increased prominence perception.

General Discussion (1)

- Visual cues for focus: eyebrows, head nods, beat gestures, mouth opening,... [exp. 1, 2, 4].
- Visual cues boost prominence perception [exp. 1, 4].
- Visual cues may *directly* influence the auditory ones [exp. 4]
- Visual cues contribute to focus perception but primarily when speech cues are underdetermined [exp. 1, 3].
- Incongruencies between visual and auditory information are disprefered [exp. 1], lead to confusion [exp. 2] and to longer processing times [exp. 3].

Methodological remarks

- Tried various experimental paradigms, all with their own pros and cons.
- *Combination* of paradigms is obviously beneficial (e.g., analysis-by-synthesis combined with analysis-by-observation).
- Combination of production with perception studies is crucial.
  (Requires many subjects: 4 experiments had 45 speakers and 110 listeners...)
- Incongruent stimuli can be more informative than congruent ones, e.g., if you are interested in relative importance of cues (cf. McGurk and MacDonald 1976, Goldin-Meadow and Wagner 2004).

Plan

1. Introduction: NLG for Embodied Agents
2. What is Natural Language Generation?
3. Case: Generation of Referring Expressions (GRE)
4. NLG for Embodied Agents: Requirements
5. Case: GRE for Embodied Agents
6. Realization: Audiovisual Speech and Gestures
7. Case: NLG for Teams of Embodied Agents
8. Evaluating NLG for Embodied Agents

Plan

1. Introduction: NLG for Embodied Agents
2. What is Natural Language Generation?
3. Case: Generation of Referring Expressions (GRE)
4. NLG for Embodied Agents: Requirements
5. Case: GRE for Embodied Agents
6. Realization: Audiovisual Speech and Gestures
7. Case: NLG for Teams of Embodied Agents
8. Evaluating NLG for Embodied Agents

NECA

NET ENVIRONMENT FOR EMBODIED EMOTIONAL CONVERSATIONAL AGENTS

Martin Klesen, Thomas Rist,
Marc Schröder

ITRI

Paul Piwek,
Kees van Deemter

Brigitte Krenn,
Hannes Pirker, Neil Tipper

Stefan Baumann,
Martine Grice

Erich Gstrein, Bernhard Herzog
Barbara Neumayr

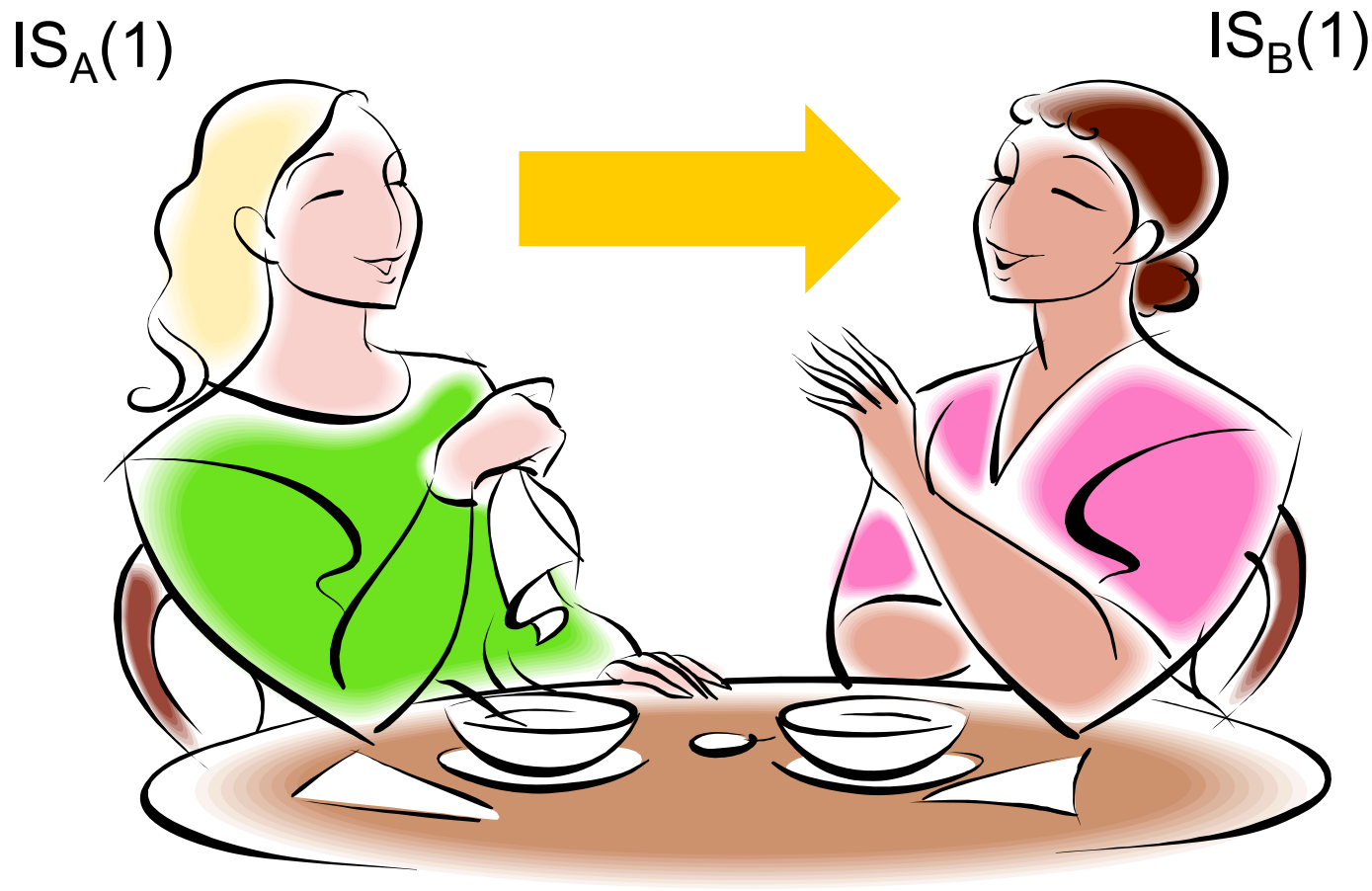
NECA

NET ENVIRONMENT FOR EMBODIED EMOTIONAL CONVERSATIONAL AGENTS

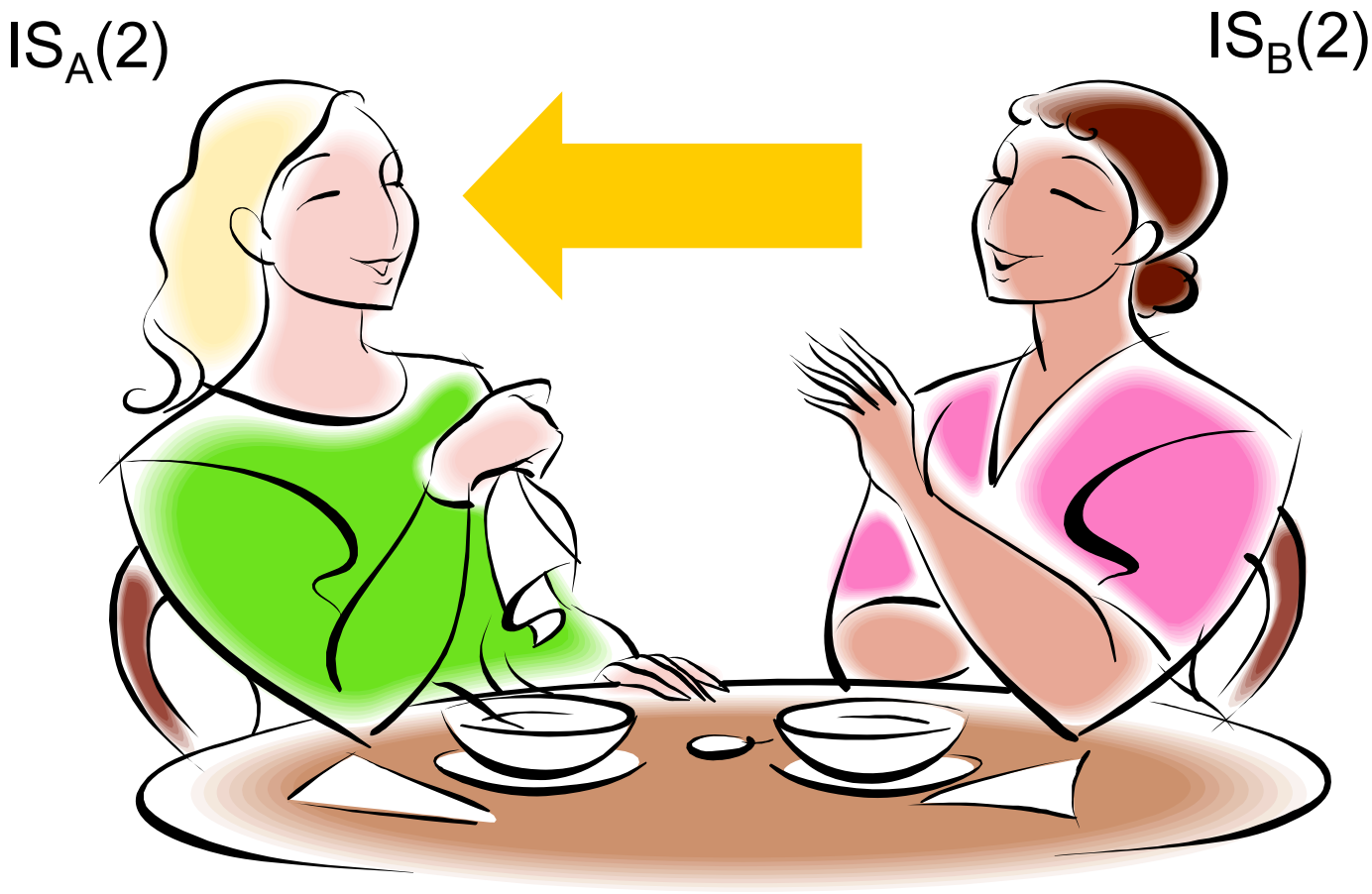
Goal: Automated generation of dialogues for embodied/animated characters

Status: Completed in May 2004


$IS_A(1)$
$IS_B(1)$


$IS_A(1)$
$IS_B(1)$


$IS_A(2)$
$IS_B(2)$


$IS_A(2)$
$IS_B(2)$


$IS_A(3)$
$IS_B(3)$

Changing the Perspective: Scripted Dialogue

Automated Generation of Scripted Dialogue

Pioneered in André et al. (2000)

Why Scripted Dialogue?

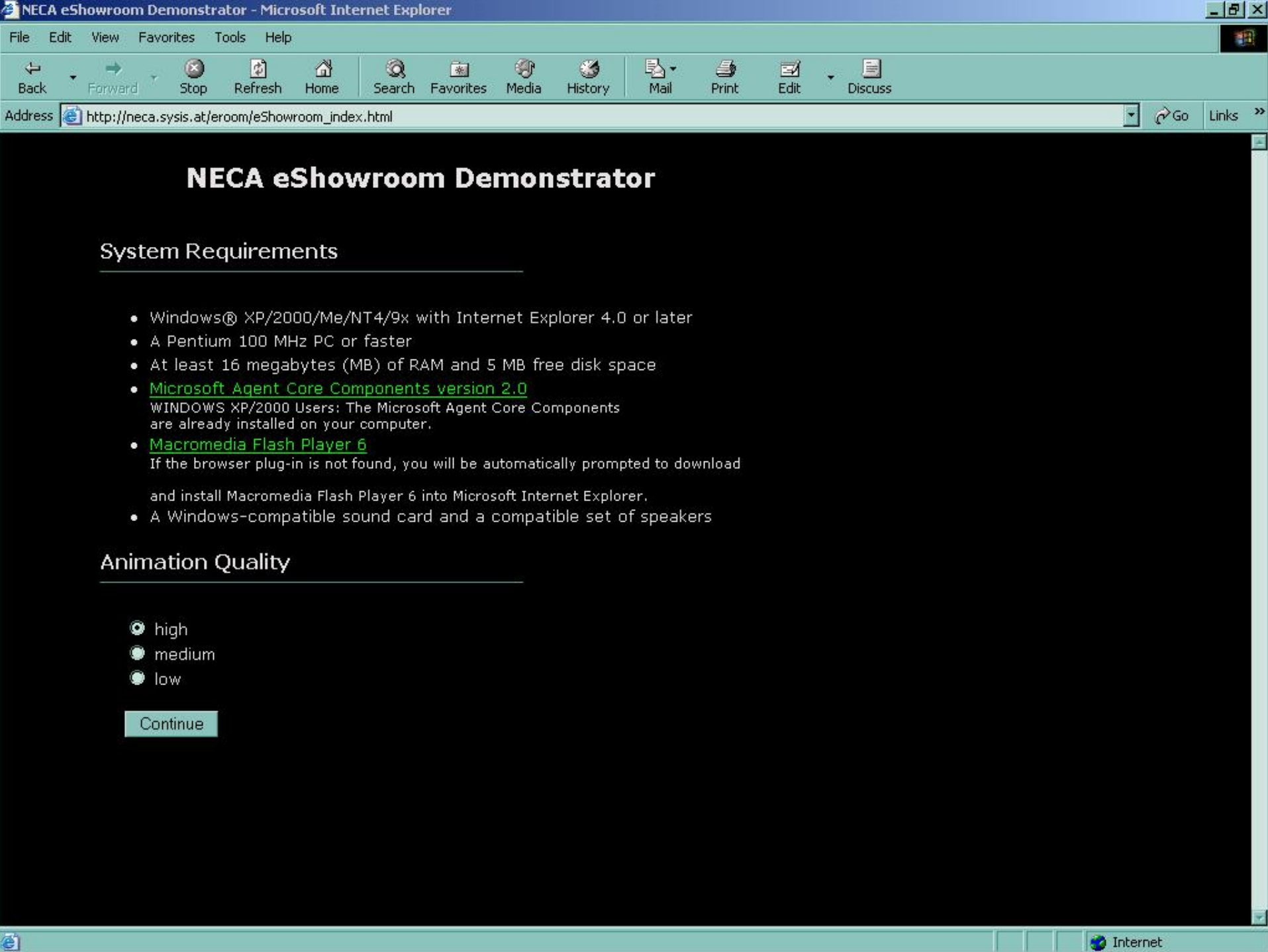
- **Limitation**
  - It can not be applied if we require real-time interaction with a user or the environment.

Why Scripted Dialogue?

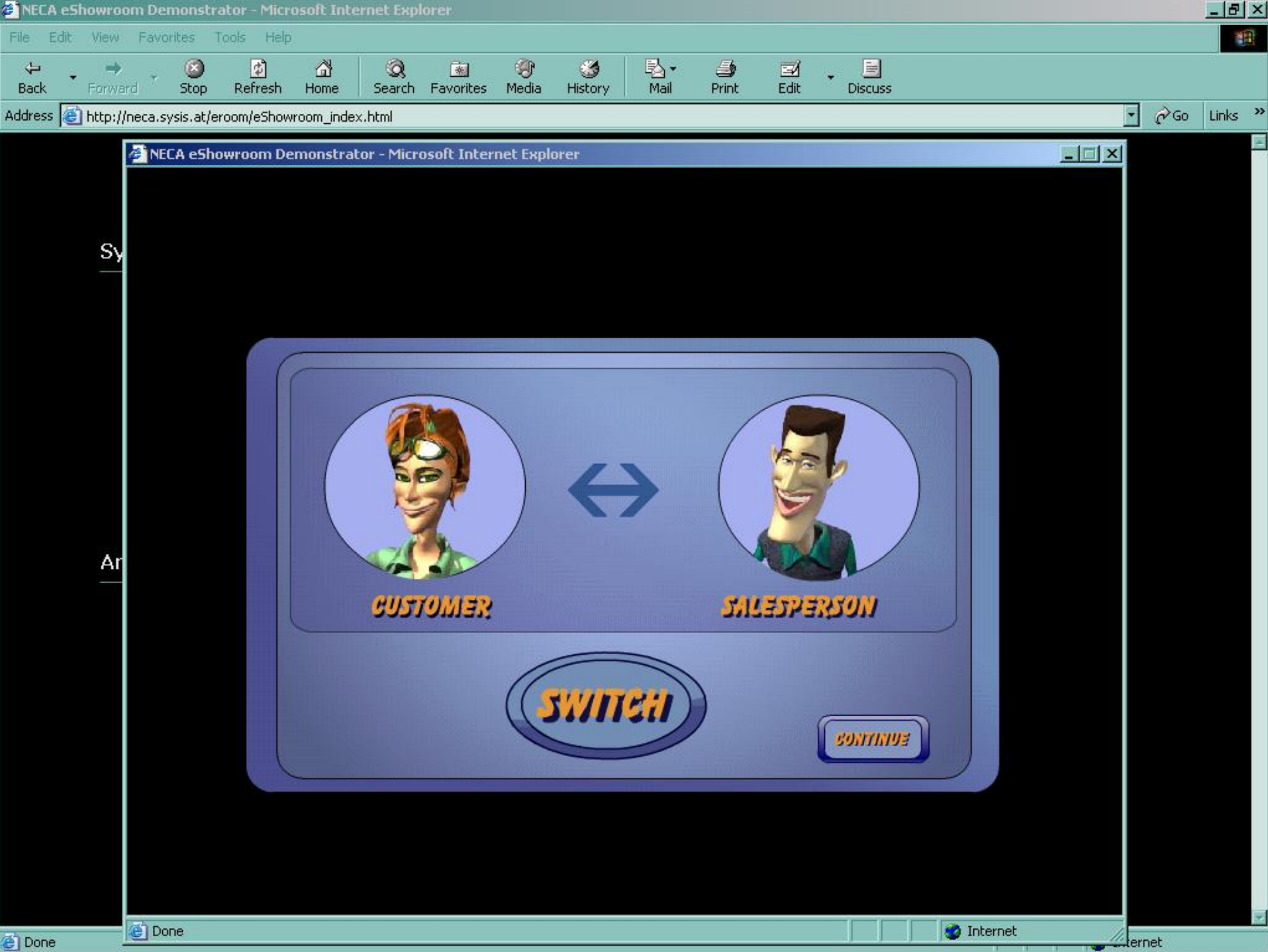
- **Limitation**
  - It can not be applied if we require real-time interaction with a user or the environment.
- **Where is/can it be applied?**
  - Media: television, film, radio, **the net**, games, ...
  - Purposes: entertainment, advertising, instruction, education (vicarious learning: e.g., Lee et al., 1998; Craig et al., 2000).
  - Mixed approaches (e.g., in training environments, games,...).

Why Scripted Dialogue?

- **New opportunities**
  - No parsing, interpretation required;
  - Real-time generation is not required;
  - New possibilities for optimization because the order of generation is decoupled from temporal order of performance.


NECA eShowroom Demonstrator - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites Media History Mail Print Edit Discuss
Address http://neca.sysis.at/eroom/eShowroom_index.html
Go Links
NECA eShowroom Demonstrator
System Requirements
Windows® XP/2000/Me/NT4/9x with Internet Explorer 4.0 or later
A Pentium 100 MHz PC or faster
At least 16 megabytes (MB) of RAM and 5 MB free disk space
Microsoft Agent Core Components version 2.0
WINDOWS XP/2000 Users: The Microsoft Agent Core Components are already installed on your computer.
Macromedia Flash Player 6
If the browser plug-in is not found, you will be automatically prompted to download and install Macromedia Flash Player 6 into Microsoft Internet Explorer.
A Windows-compatible sound card and a compatible set of speakers
Animation Quality
high
medium
low
Continue
Internet


NECA eShowroom Demonstrator - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites Media History Mail Print Edit Discuss
Address http://neca.sysis.at/eroom/eShowroom_index.html
Go
NECA eShowroom Demonstrator - Microsoft Internet Explorer
CUSTOMER
SALESPERSON
SWITCH
CONTINUE
Done
Internet
Done


NECA eShowroom Demonstrator - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites Media History Mail Print Edit Discuss
Address http://neca.sysis.at/eroom/eShowroom_index.html
Go Links
NECA eShowroom Demonstrator - Microsoft Internet Explorer
CUSTOMER
SALESPERSON
SWITCH
CONTINUE
Done
Internet
Done


NECA eShowroom Demonstrator - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites Media History Mail Print Edit Discuss
Address http://neca.sysis.at/eroom/eShowroom_index.html
Go Links
NECA eShowroom Demonstrator - Microsoft Internet Explorer
IMPOLITE ILL TEMPERED
POLITE ILL TEMPERED
IMPOLITE GOOD HUMORED
POLITE GOOD HUMORED
RITCHIE
CONTINUE
Done
Internet
Done

NECA eShowroom Demonstrator - Microsoft Internet Explorer

File Edit View Favorites Tools Help

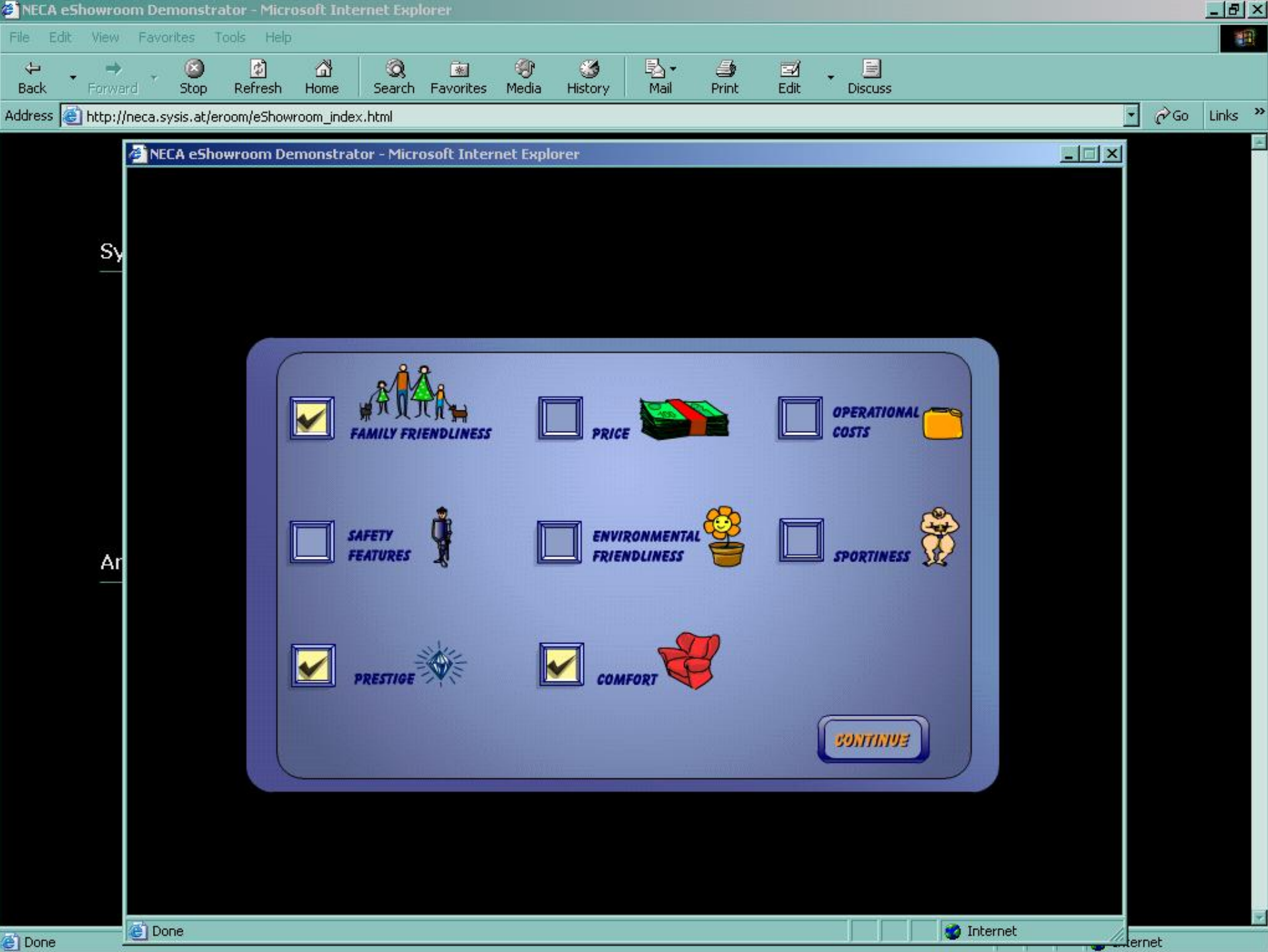
Back Forward Stop Refresh Home Search Favorites Media History Mail Print Edit Discuss

Address http://neca.sysis.at/eroom/eShowroom_index.html Go Links

NECA eShowroom Demonstrator - Microsoft Internet Explorer

Done Internet

Done


NECA eShowroom Demonstrator - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites Media History Mail Print Edit Discuss
Address http://neca.sysis.at/eroom/eShowroom_index.html
Go Links
NECA eShowroom Demonstrator - Microsoft Internet Explorer
FAMILY FRIENDLINESS
PRICE
OPERATIONAL COSTS
SAFETY FEATURES
ENVIRONMENTAL FRIENDLINESS
SPORTINESS
PRESTIGE
COMFORT
CONTINUE
Done
Internet
Done

NECA Flash Player - Microsoft Internet Explorer

File Edit View Favorites Tools Help

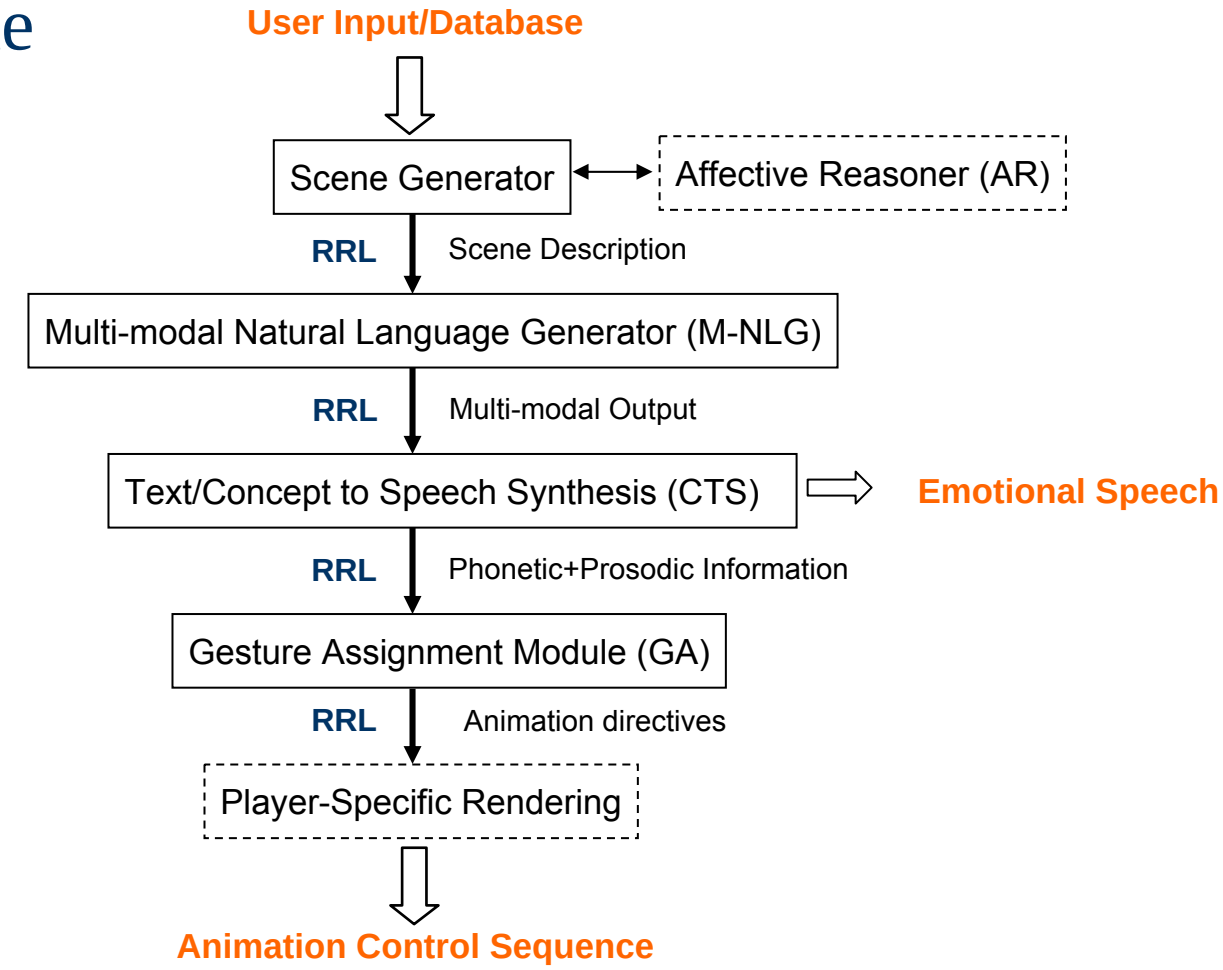
Open Again

Done

Internet

Start 4:03 PM

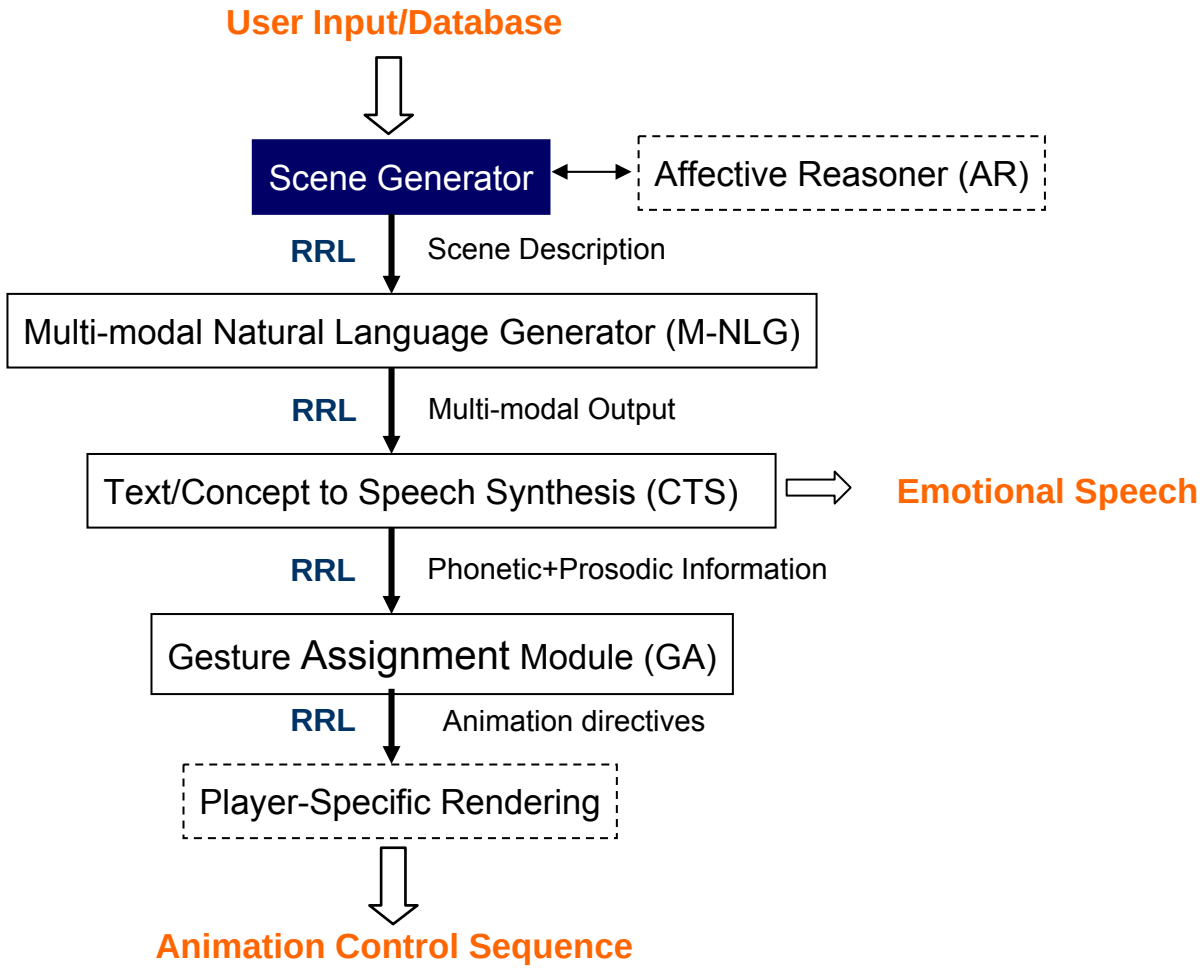
NECA Pipeline



Database (Java JAM)

```
FACT attribute "car-1" "horsepower“ "80hp";
FACT impact "car-1" "horsepower“ "sportiness" "pos";
FACT importance "horsepower" "sportiness" "high";
FACT role "Ritchie" "seller";
FACT role "Tina" "buyer";
FACT trait "Ritchie" "politeness" "impolite";
```

DFKI


User Input/Database
Scene Generator
Affective Reasoner (AR)
RRL
Scene Description
Multi-modal Natural Language Generator (M-NLG)
RRL
Multi-modal Output
Text/Concept to Speech Synthesis (CTS)
Emotional Speech
RRL
Phonetic+Prosodic Information
Gesture Assignment Module (GA)
RRL
Animation directives
Player-Specific Rendering
Animation Control Sequence

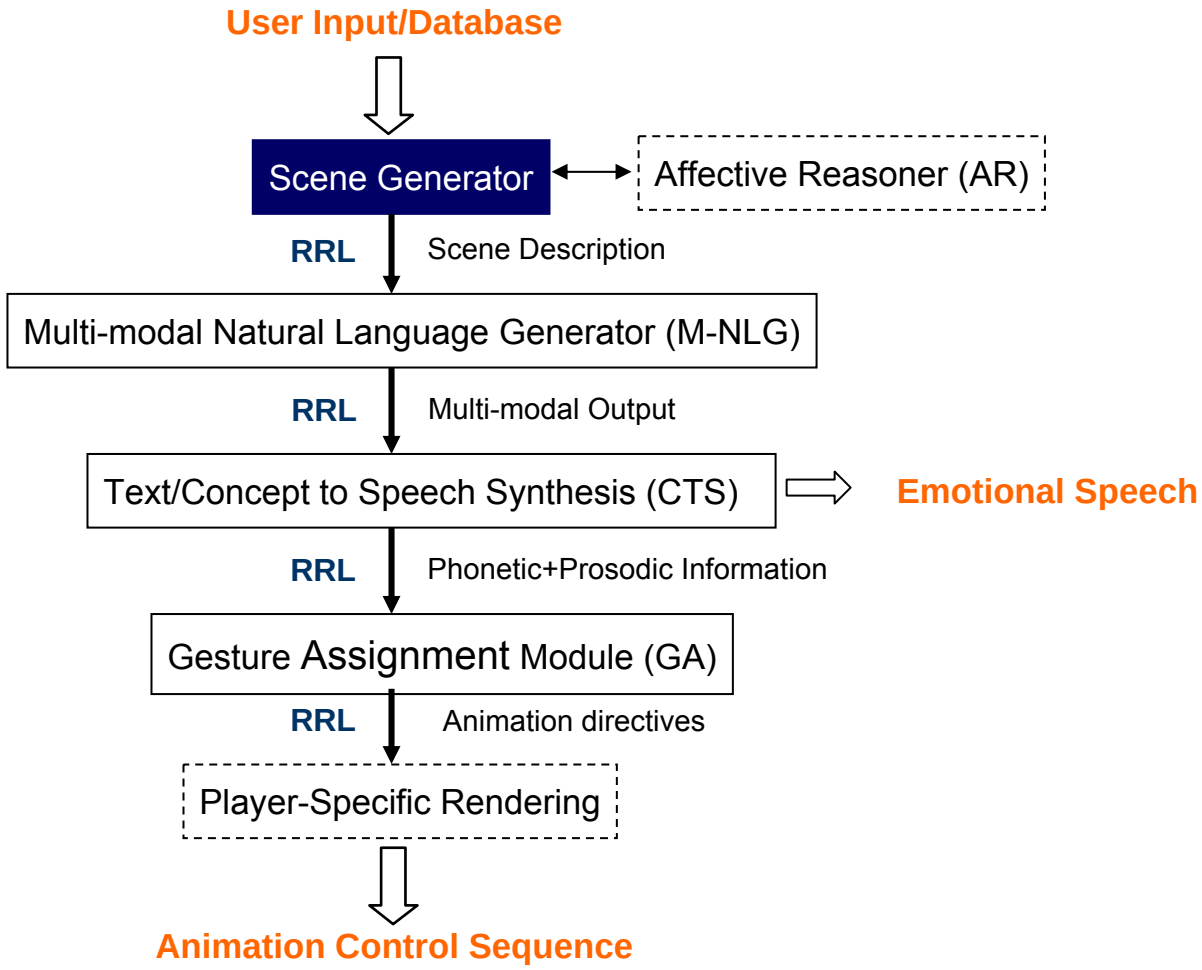
TILBURG UNIVERSITY

The Open University

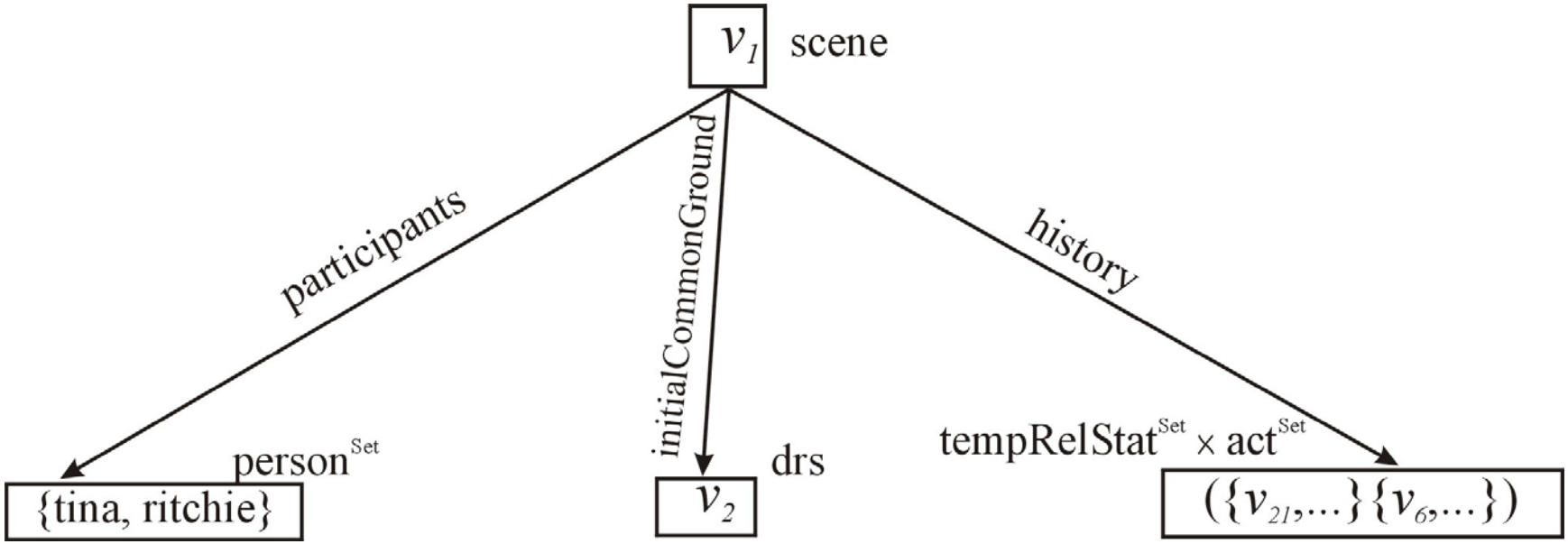
DFKI

***RRL** (XML): Rich Representation Language*

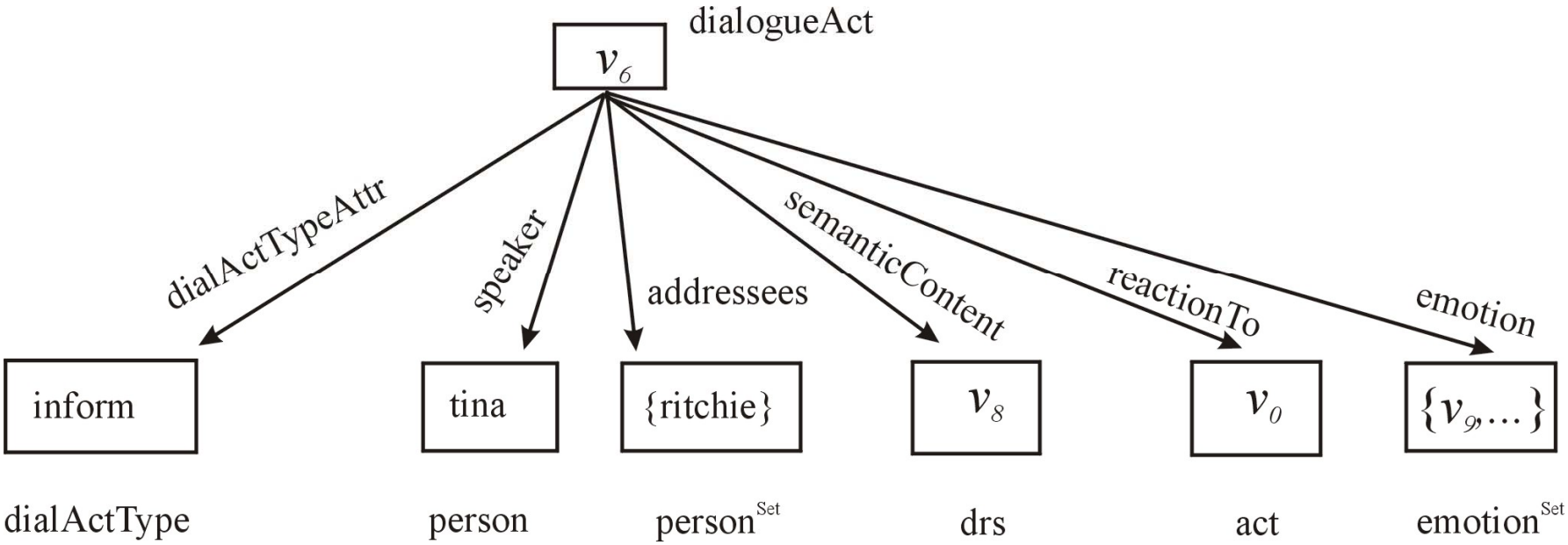
for Description of Agent Behaviour



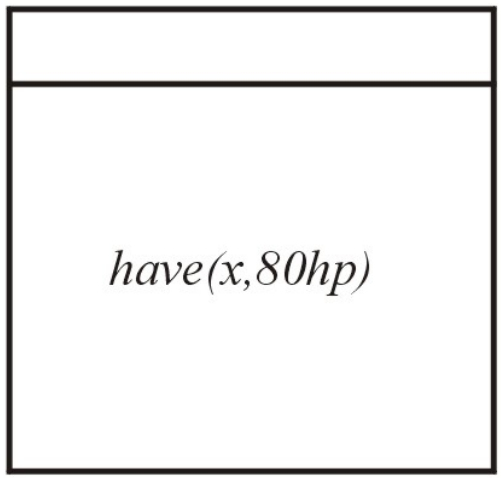
Scene Description



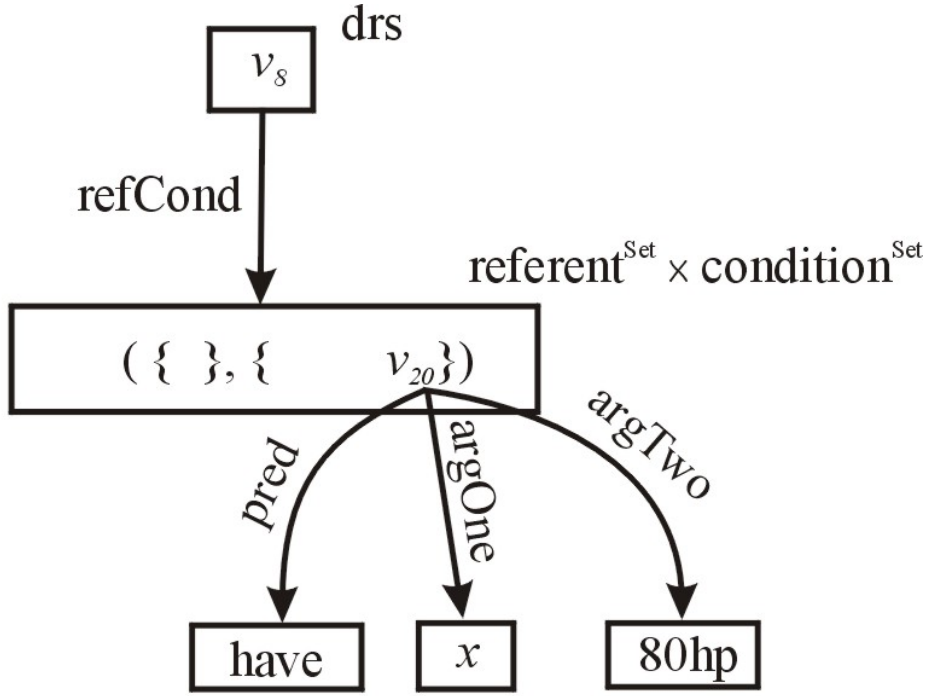
Dialogue Act



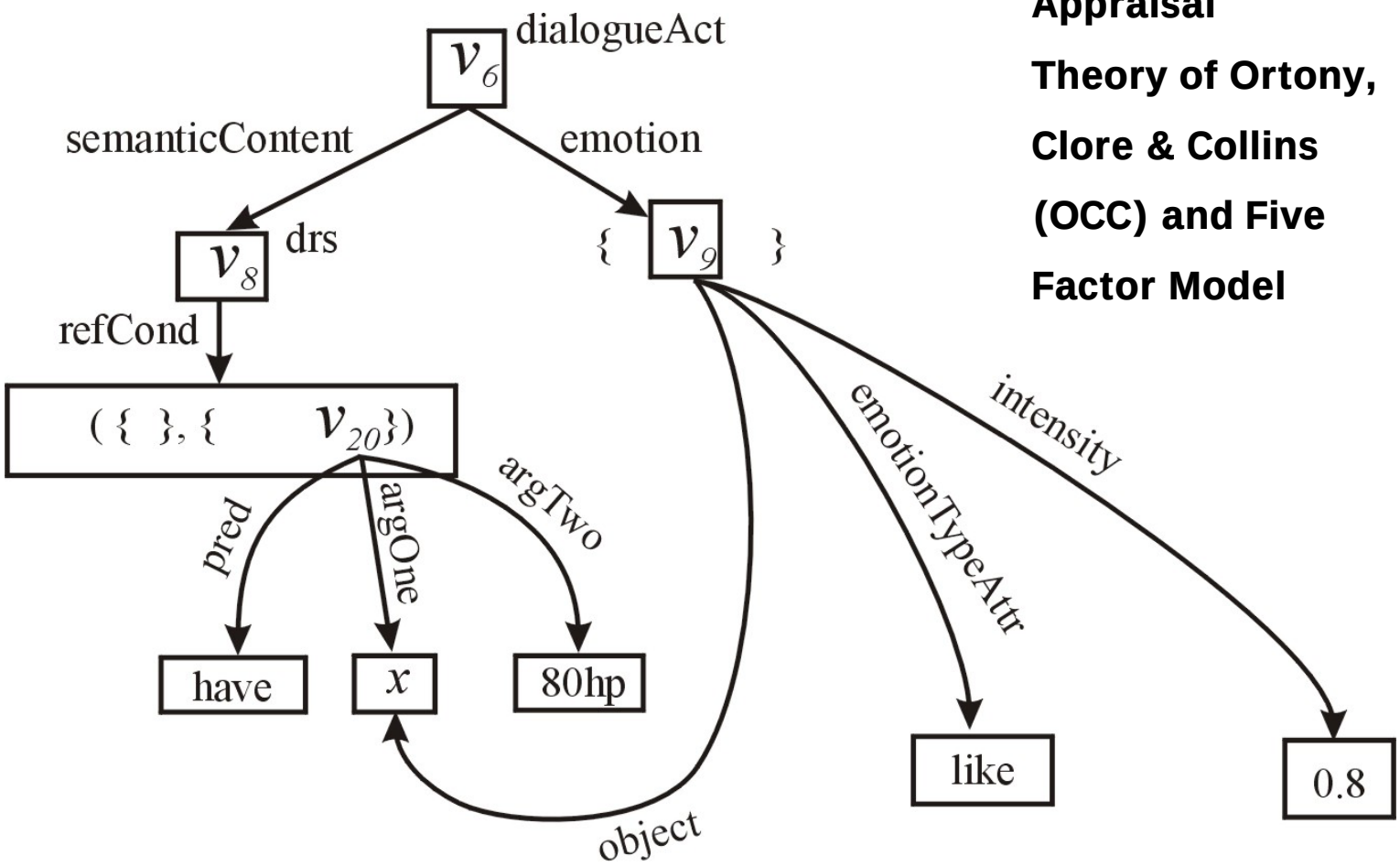
Semantic Content



a. DRS in standard notation

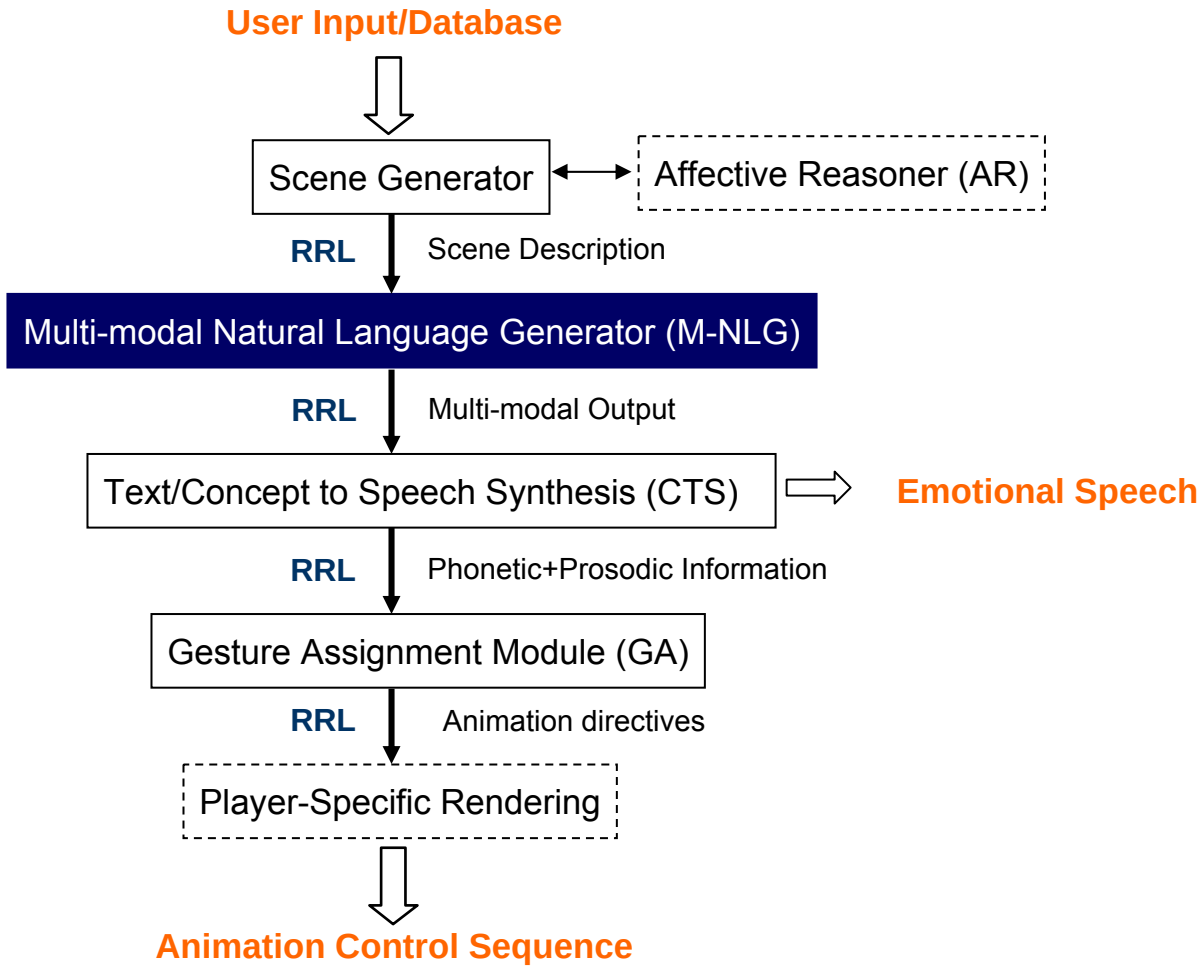


b. DRS encoded in a network representation



Appraisal Theory of Ortony, Clore & Collins (OCC) and Five Factor Model

ITRI


User Input/Database
Scene Generator
Affective Reasoner (AR)
RRL
Scene Description
Multi-modal Natural Language Generator (M-NLG)
RRL
Multi-modal Output
Text/Concept to Speech Synthesis (CTS)
Emotional Speech
RRL
Phonetic+Prosodic Information
Gesture Assignment Module (GA)
RRL
Animation directives
Player-Specific Rendering
Animation Control Sequence

NECA MNLG Output

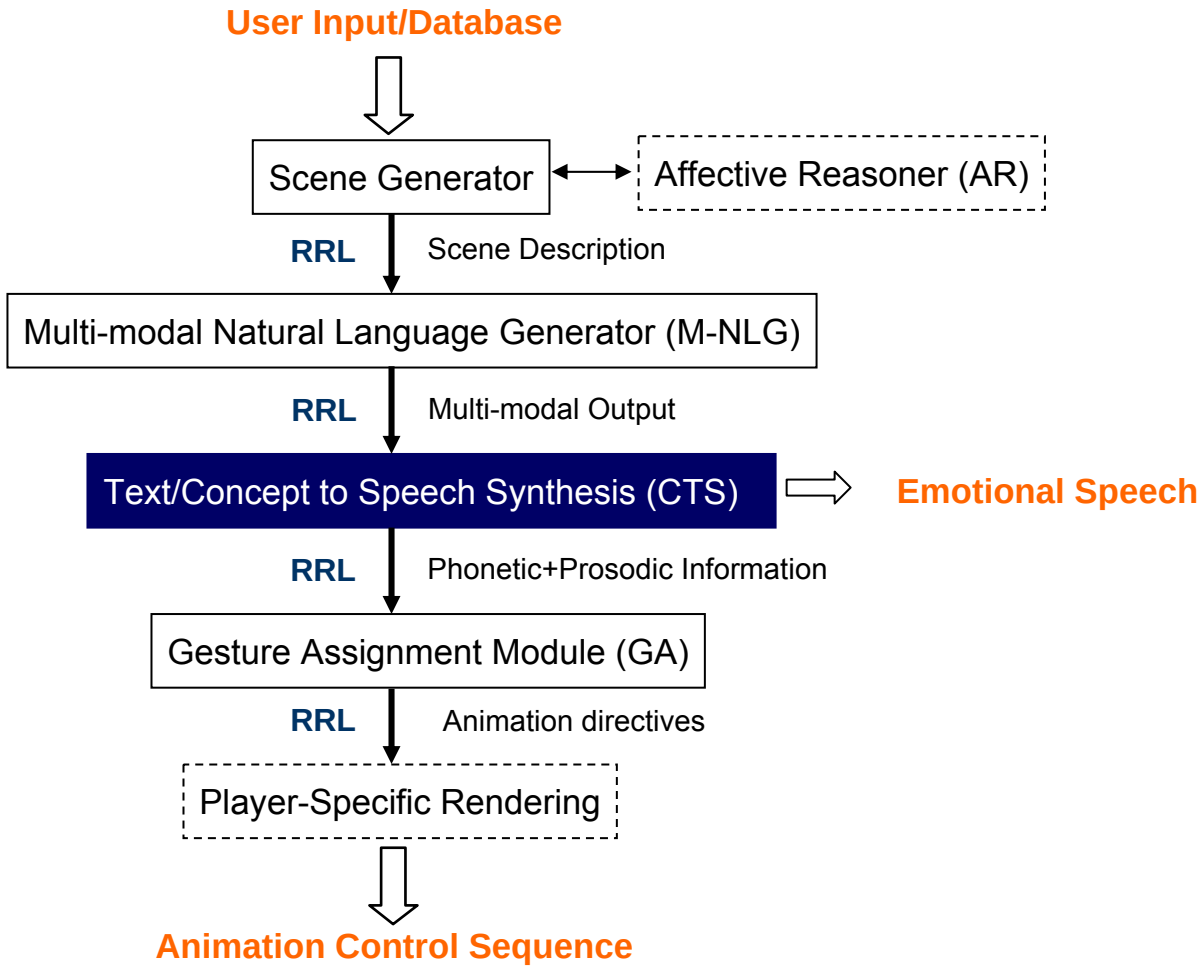
<gesture meaning="takingcommand“ alignto=“s1” aligntype=“seq_before” modality="body" identifier="hips" id="g1"/>

<gesture meaning="deictic" modality="face“ alignto=“s1” aligntype=“par” identifier="LookCar" id="g3"/>

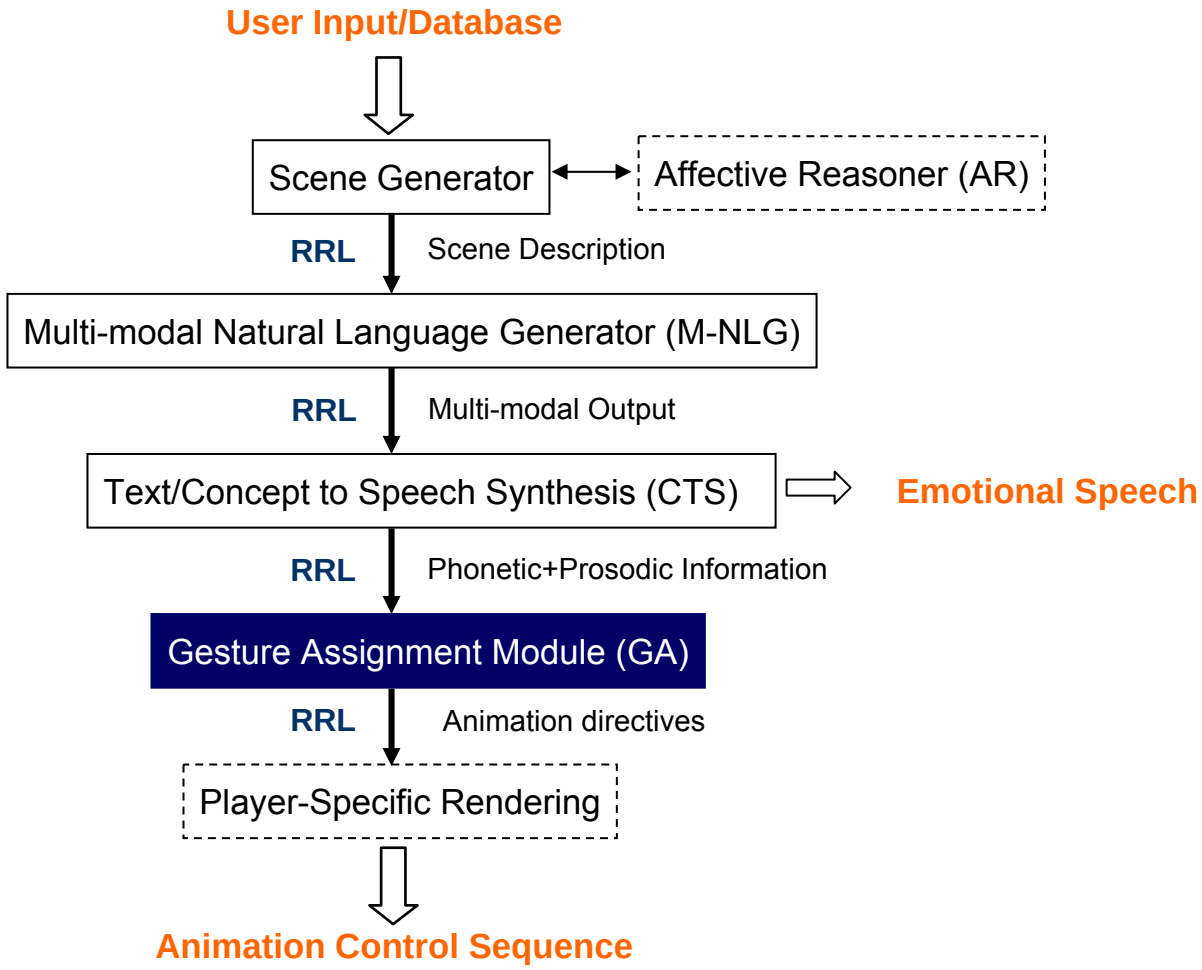
<sentence id="s1">How much fuel does this car consume?</sentence>

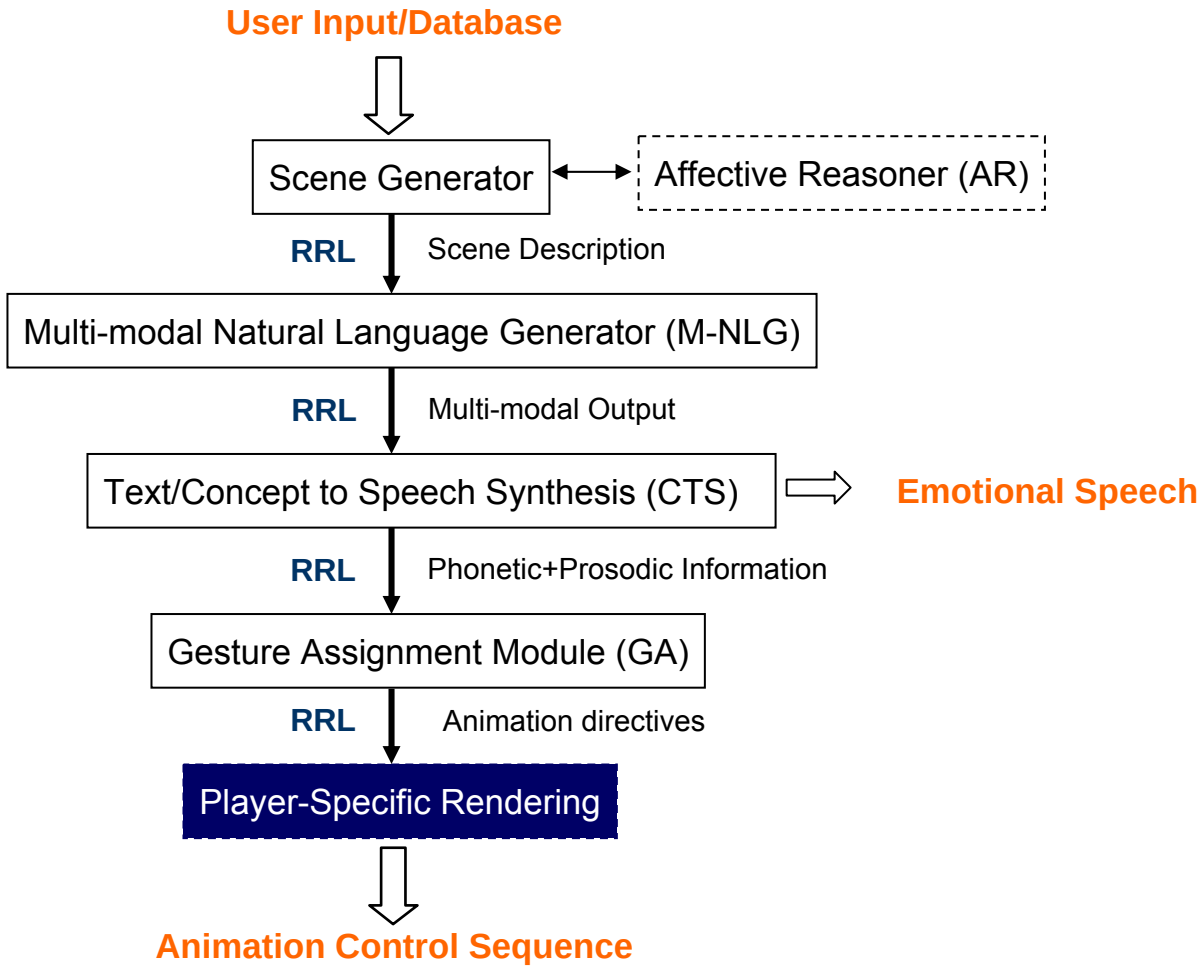
<gesture meaning="establishcontact" alignto=“s1” aligntype=“seq_after” modality="face" identifier="LookAddressee" id="g2"/>

DFKI/Saarland



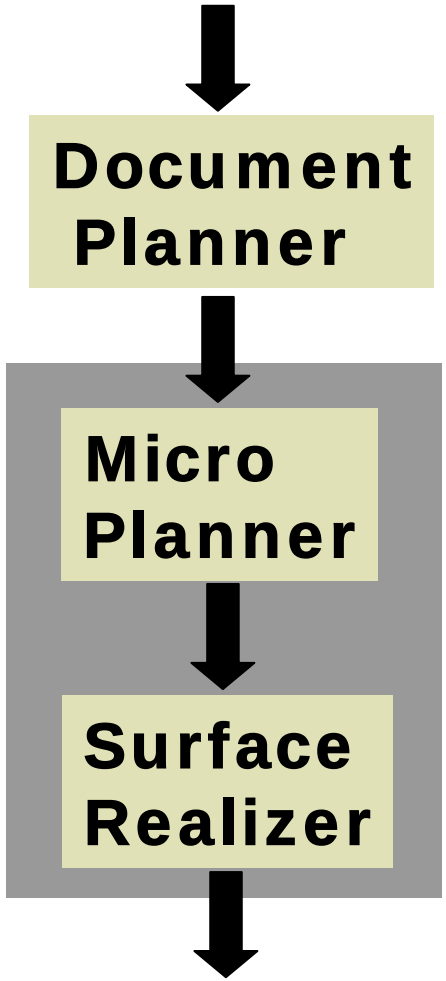
OEFAI

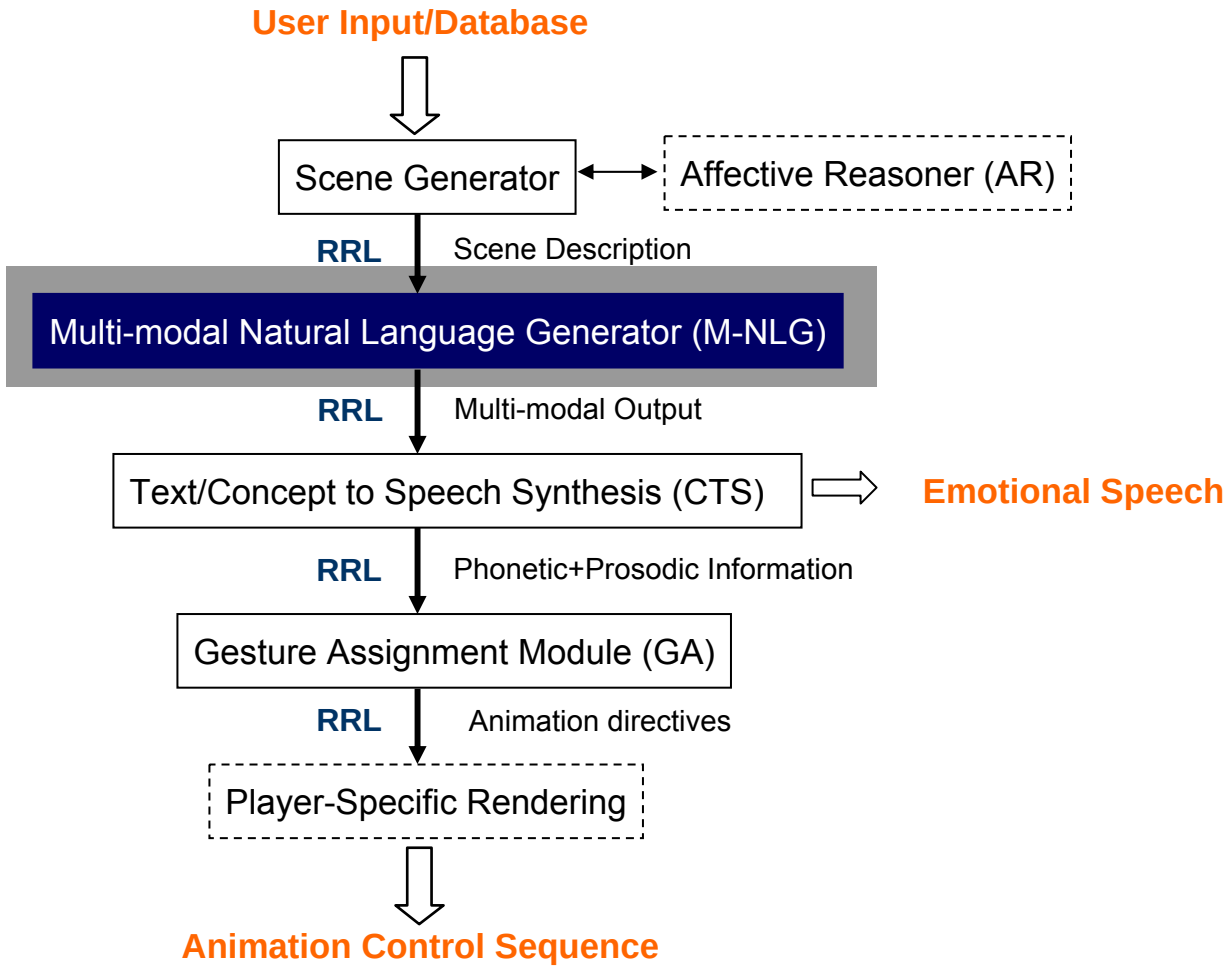

User Input/Database
Scene Generator
Affective Reasoner (AR)
RRL
Scene Description
Multi-modal Natural Language Generator (M-NLG)
RRL
Multi-modal Output
Text/Concept to Speech Synthesis (CTS)
Emotional Speech
RRL
Phonetic+Prosodic Information
Gesture Assignment Module (GA)
RRL
Animation directives
Player-Specific Rendering
Animation Control Sequence


User Input/Database
Scene Generator
Affective Reasoner (AR)
RRL
Scene Description
Multi-modal Natural Language Generator (M-NLG)
RRL
Multi-modal Output
Text/Concept to Speech Synthesis (CTS)
Emotional Speech
RRL
Phonetic+Prosodic Information
Gesture Assignment Module (GA)
RRL
Animation directives
Player-Specific Rendering
Animation Control Sequence

MSAgents Script (HTML Java Script)

```
ritchieRequest = ritchie.Play("GS_Cup");
ritchieRequest = ritchie.Speak("It has airbags.");
ritchieRequest
ritchie.Play("AS_LookLeft");
tina.Wait(ritchieRequest);
tinaRequest = tina.Play("GS_Attention");
tinaRequest = tina.Speak("Does it have power
  windows?");
```


Document Planner
Micro Planner
Surface Realizer


User Input/Database
Scene Generator
Affective Reasoner (AR)
RRL
Scene Description
Multi-modal Natural Language Generator (M-NLG)
RRL
Multi-modal Output
Text/Concept to Speech Synthesis (CTS)
Emotional Speech
RRL
Phonetic+Prosodic Information
Gesture Assignment Module (GA)
RRL
Animation directives
Player-Specific Rendering
Animation Control Sequence

Two Principal Requirements

1. The linguistic resources should support *seamless integration* of canned text, templates and full grammar rules.
2. The generator should allow for the *combination* of different types of constraints on its output (syntactic, semantic, pragmatic).

The MNLG Pipeline (Piwek, 2003)

1. Parsing of RRL XML Scene (using Pillow; Cabeza & Hermenegildo) to Prolog terms (Erbach's Profit: Prolog with Feature Structures).
2. Generating Deep Structures;
3. Generating Ref. Exp. Deep Structures (extension of Dale & Reiter with salience, following Krahmer & Theune);
4. Lexical Realization (inflection, agreement, punctuation);
5. Gesture Generation;
6. Mapping to RRL XML output.

The MNLG Pipeline

1. Parsing of RRL XML Scene (using Pillow; Cabeza & Hermenegildo) to Prolog terms (Erbach's Profit: Prolog with Feature Structures).
2. Generating Deep Structures;
3. Generating Ref. Exp. Deep Structures (extension of Dale & Reiter with salience, following Krahmer & Theune);
4. Lexical Realization (inflection, agreement, punctuation);
5. Gesture Generation;
6. Mapping to RRL XML output.

The MNLG Pipeline

1. Parsing of RRL XML Scene (using Pillow; Cabeza & Hermenegildo) to Prolog terms (Erbach's Profit: Prolog with Feature Structures).
2. Generating Deep Structures;
3. Generating Ref. Exp. Deep Structures (extension of Dale & Reiter with salience, following Krahmer & Theune);
4. Lexical Realization (inflection, agreement, punctuation);
5. Gesture Generation;
6. Mapping to RRL XML output.

The MNLG Pipeline

1. Parsing of RRL XML Scene (using Pillow; Cabeza & Hermenegildo) to Prolog terms (Erbach's Profit: Prolog with Feature Structures).
2. Generating Deep Structures;
3. Generating Ref. Exp. Deep Structures (extension of Dale & Reiter with salience, following Krahmer & Theune);
4. Lexical Realization (inflection, agreement, punctuation);
5. Gesture Generation;
6. Mapping to RRL XML output.

The MNLG Pipeline

1. Parsing of RRL XML Scene (using Pillow; Cabeza & Hermenegildo) to Prolog terms (Erbach's Profit: Prolog with Feature Structures).
2. Generating Deep Structures;
3. Generating Ref. Exp. Deep Structures (extension of Dale & Reiter with salience, following Krahmer & Theune);
4. Lexical Realization (inflection, agreement, punctuation);
5. Gesture Generation;
6. Mapping to RRL XML output.

The MNLG Pipeline

1. Parsing of RRL XML Scene (using Pillow; Cabeza & Hermenegildo) to Prolog terms (Erbach's Profit: Prolog with Feature Structures).
2. Generating Deep Structures;
3. Generating Ref. Exp. Deep Structures (extension of Dale & Reiter with salience, following Krahmer & Theune);
4. Lexical Realization (inflection, agreement, punctuation);
5. Gesture Generation;
6. Mapping to RRL XML output.

The MNLG Pipeline

1. Parsing of RRL XML Scene (using Pillow; Cabeza & Hermenegildo) to Prolog terms (Erbach's Profit: Prolog with Feature Structures).
2. Generating Deep Structures;
3. Generating Ref. Exp. Deep Structures (extension of Dale & Reiter with salience, following Krahmer & Theune);
4. Lexical Realization (inflection, agreement, punctuation);
5. Gesture Generation;
6. Mapping to RRL XML output.

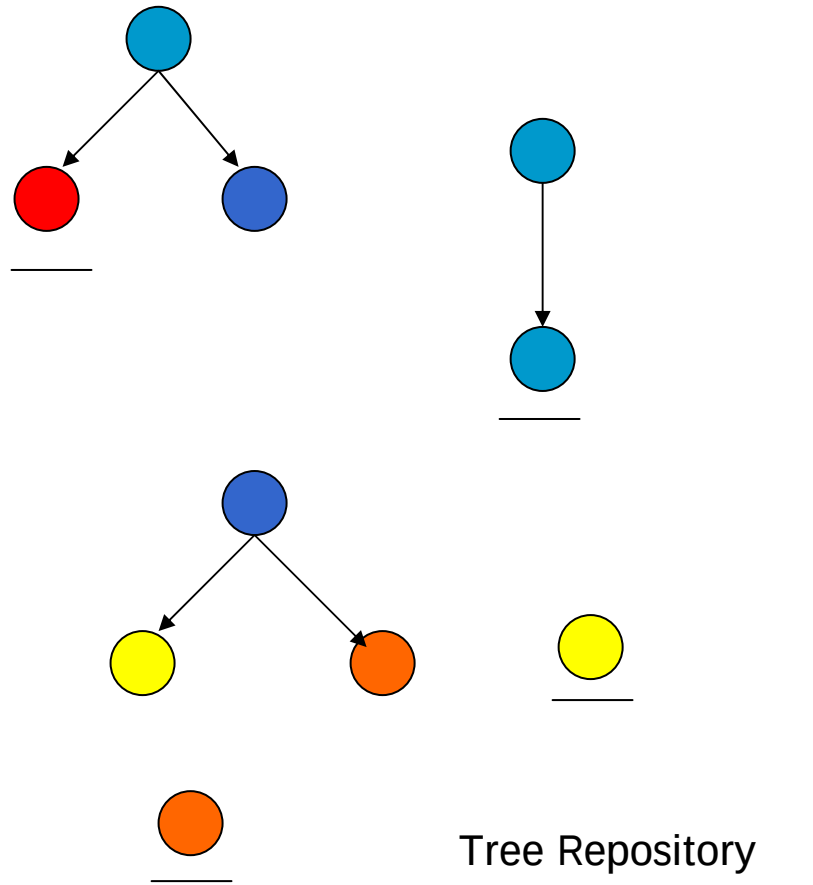
The MNLG Pipeline

1. Parsing of RRL XML Scene (using Pillow; Cabeza & Hermenegildo) to Prolog terms (Erbach's Profit: Prolog with Feature Structures).
2. Generating Deep Structures;
3. Generating Ref. Exp. Deep Structures (extension of Dale & Reiter with salience, following Krahmer & Theune);
4. Lexical Realization (inflection, agreement, punctuation);
5. Gesture Generation;
6. Mapping to RRL XML output.

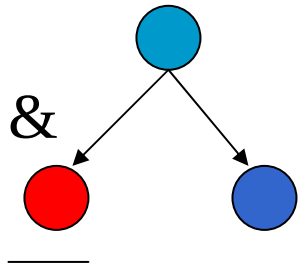
Approach to Generation

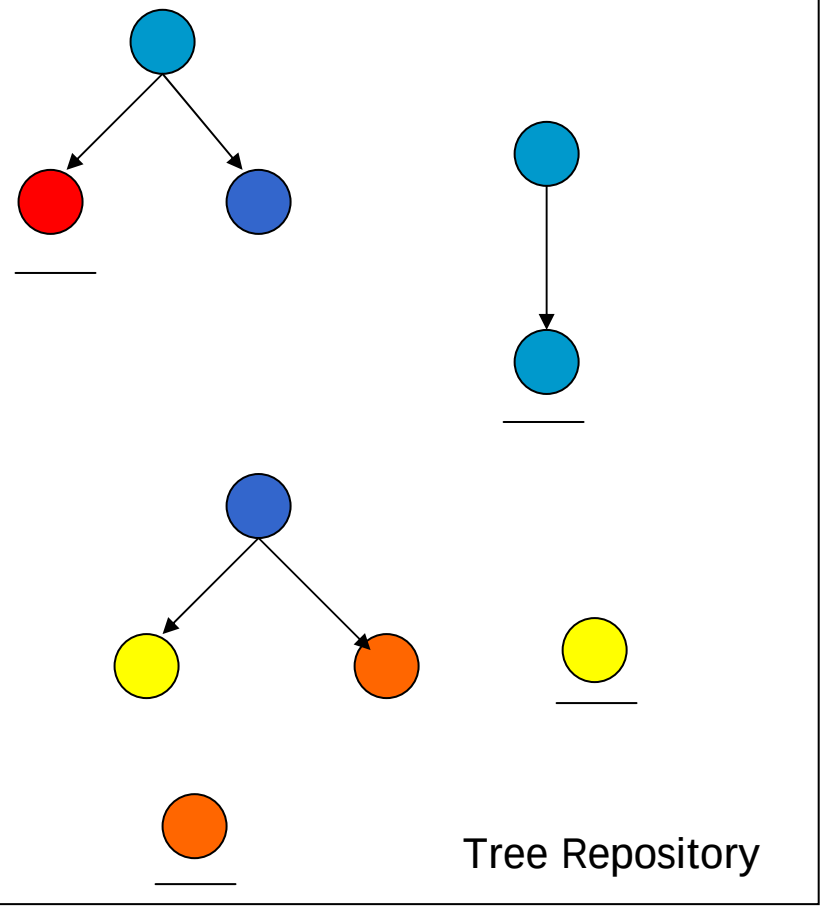
- Input: Root Node.
- Nodes are feature structures, e.g., of type s (sentence), with the following main features:
  - currentAct (speaker, addressees, type, …)
  - Sem (DRS, Kamp & Reyle)
- From a tree repository, a tree is selected whose root node matches with the input node. Subsequently the ‘open’ daughter nodes are taken as the new input and expanded.

(<s &
sem!.. &
...)

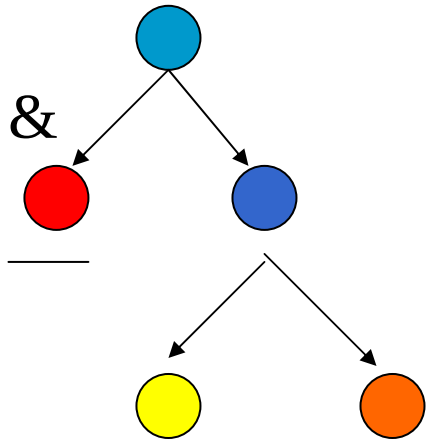


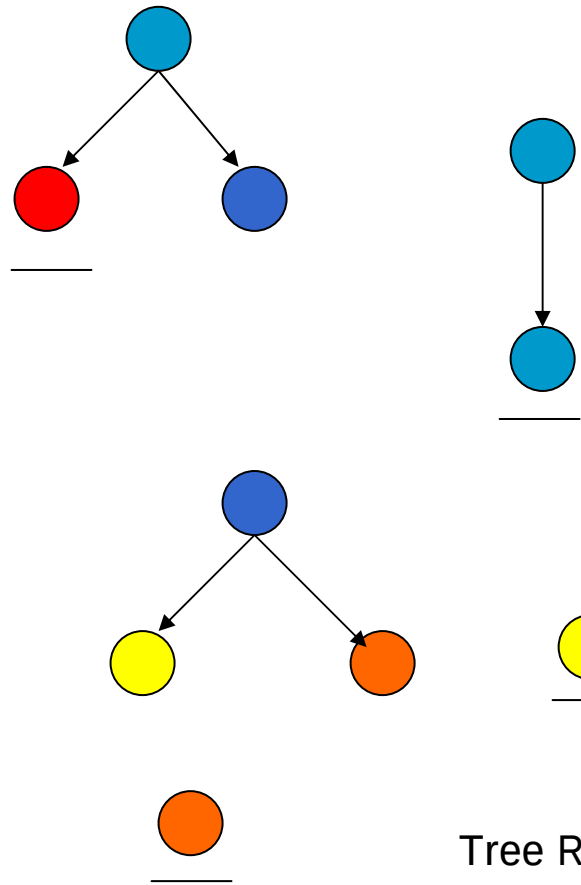
(<s &
sem!.. &
...)





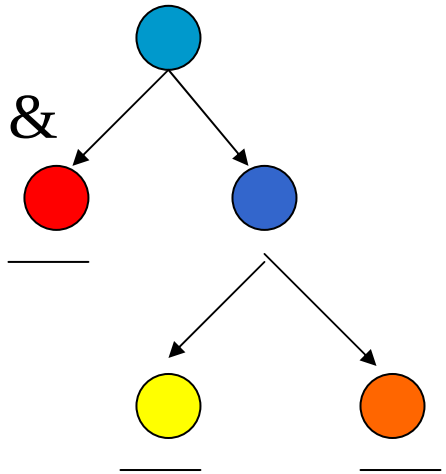
(<s &
sem!.. &
...)

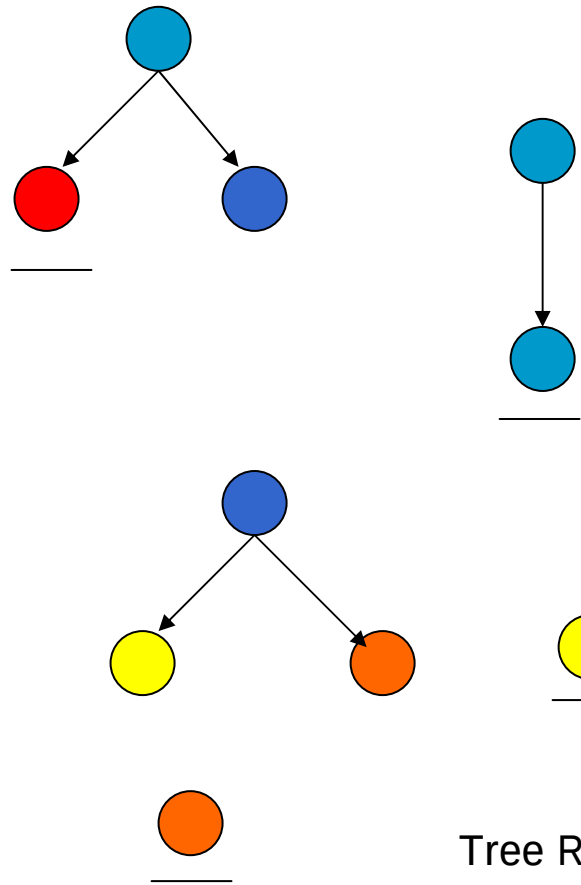




Tree Repository

(<s &
sem!.. &
...)





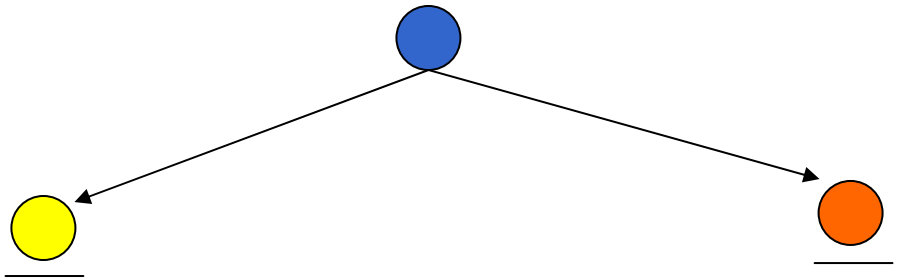
Tree Repository

Canned Text

```
  <s &
  currentAct!dialActTypeAttr!"agree" &
  sem!"none" &
form!"you are right"
```

Templates

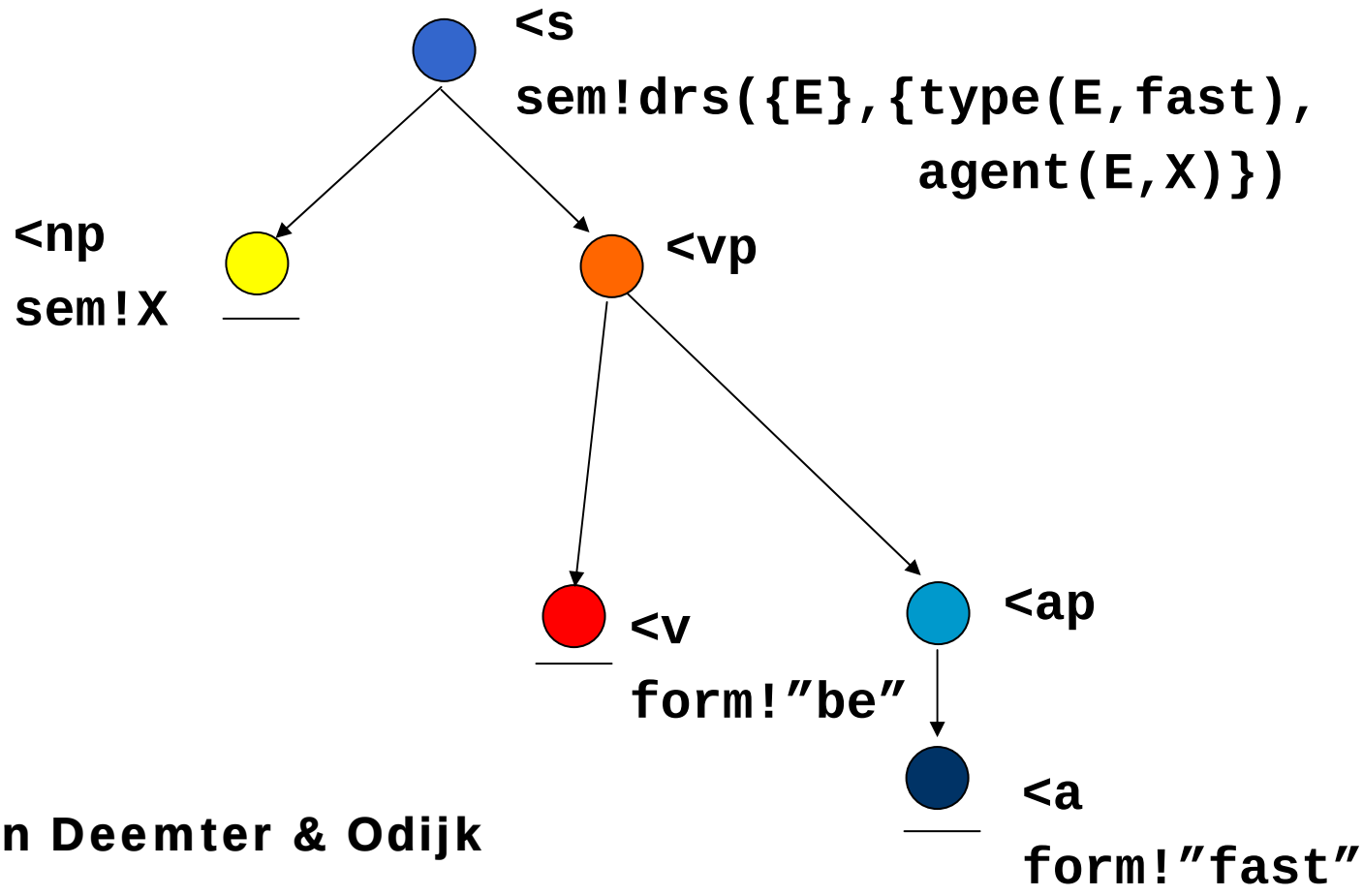
```
<s &
 currentAct!dialActTypeAttr!"greeting" &
  currentAct!speaker!polite!false &
 sem!"none" &
 currentAct!speaker!Speaker
```



```
<fragment &
form!“hey there, I am"
```

```
<np &
sem!referent(Speaker)
```

Sophisticated Templates

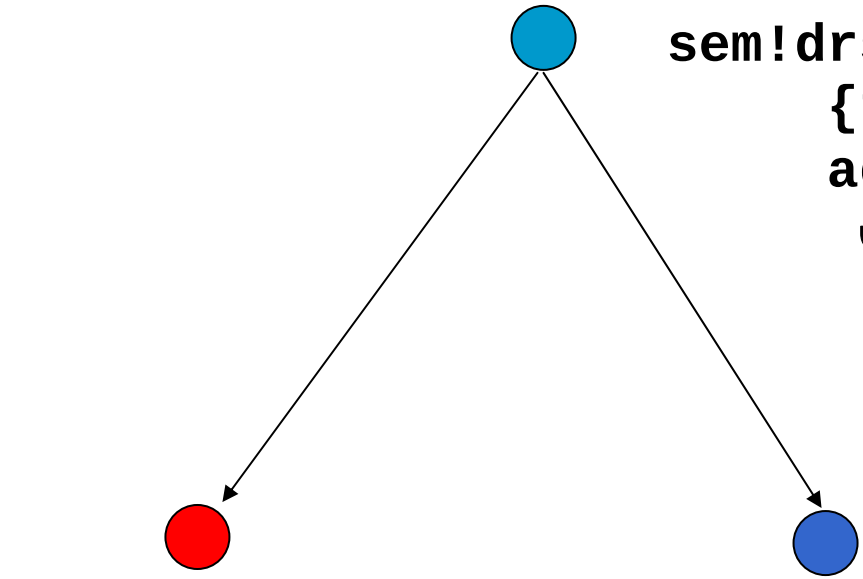


Cf. DYD; Van Deemter & Odijk

CFG Rules

S → NP VP

…



<s
currentAct!CA
sem!drs({E}∪R1,
{type(E,T),
agent(E,X)}
∪R2)

<np
currentAct!CA
sem!X

<vp
currentAct!CA
sem!drs({E} ∪R1,
{type(E,T)}
∪R2)

The MNLG Pipeline

1. Parsing of RRL XML Scene (using Pillow; Cabeza & Hermenegildo) to Prolog terms (Erbach's Profit: Prolog with Feature Structures).
2. Generating Deep Structures;
3. Generating Ref. Exp. Deep Structures (extension of Dale & Reiter with salience, following Krahmer & Theune);
4. Lexical Realization (inflection, agreement, punctuation);
5. Gesture Generation;
6. Mapping to RRL XML output.

Gestures in current Demonstrator

Speaker Gestures (vs Adaptors and Feedback Gestures)

- Topical Gestures
  - Iconic
  - Metaphoric
  - Deictic
- Interactive Gestures
  - Turn taking
  - Expression of the discourse/dialogue function
  - Beats, batons
- Emblematic Gestures

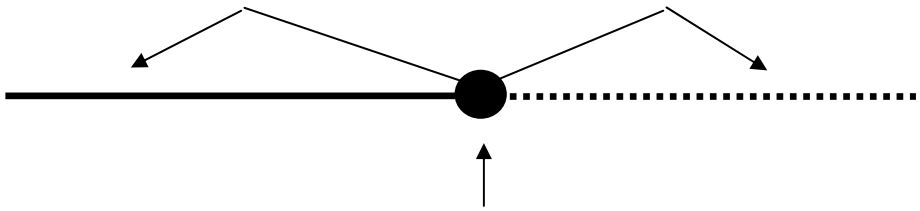
Tilburg University

The Open University

Automated Generation of Scripted Dialogue

Controlling Global Properties
(Piwek & Van Deemter, 2003)

- **Problem**: How can global properties of the generated dialogues be controlled?
- **Example**: length or style of the dialogue.
- Enforcing global properties requires non-local generation decisions



Possible Solutions

- Monitoring (e.g., Hovy, 1988)
- Estimation (e.g., Reiter, 2000)
- Revision (e.g., Callaway & Lester, 1997; Reiter, 2000; Robin & McKeown, 1996)

- Revision for Scripted Dialogue:
  - arguments in literature on text generation regarding maintainability and effectiveness of revision
  - no left to right generation constraint

1. Interacting global constraints?

- Number of turns for expressing the selected content

  TURN = Max / Min

- Degree of dialogue level emphasis

  EMPH = Max / Min

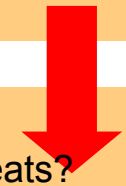
  E.g., maximize number of subdialogues which are about information which needs to be emphasized (Piwek & Van Deemter, 2002)

2. Which operations at what stage?

Adjacency Pair Insertion Example

C: Does it have leather seats?
S: Yes.
...

Paraphrase of Initial Dialogue Plan



C: Does it have leather seats?
S: Yes.
C: Really, leather seats?
S: Yes indeed it does.
...

After applying the INSERT revision operation

2. Which operations at what stage?

Adjacency Pair Aggregation Example

C: Does it have airbags?
S: Yes.
...
C: Does it have ABS?
S: Yes.

Paraphrase of Initial Dialogue Plan

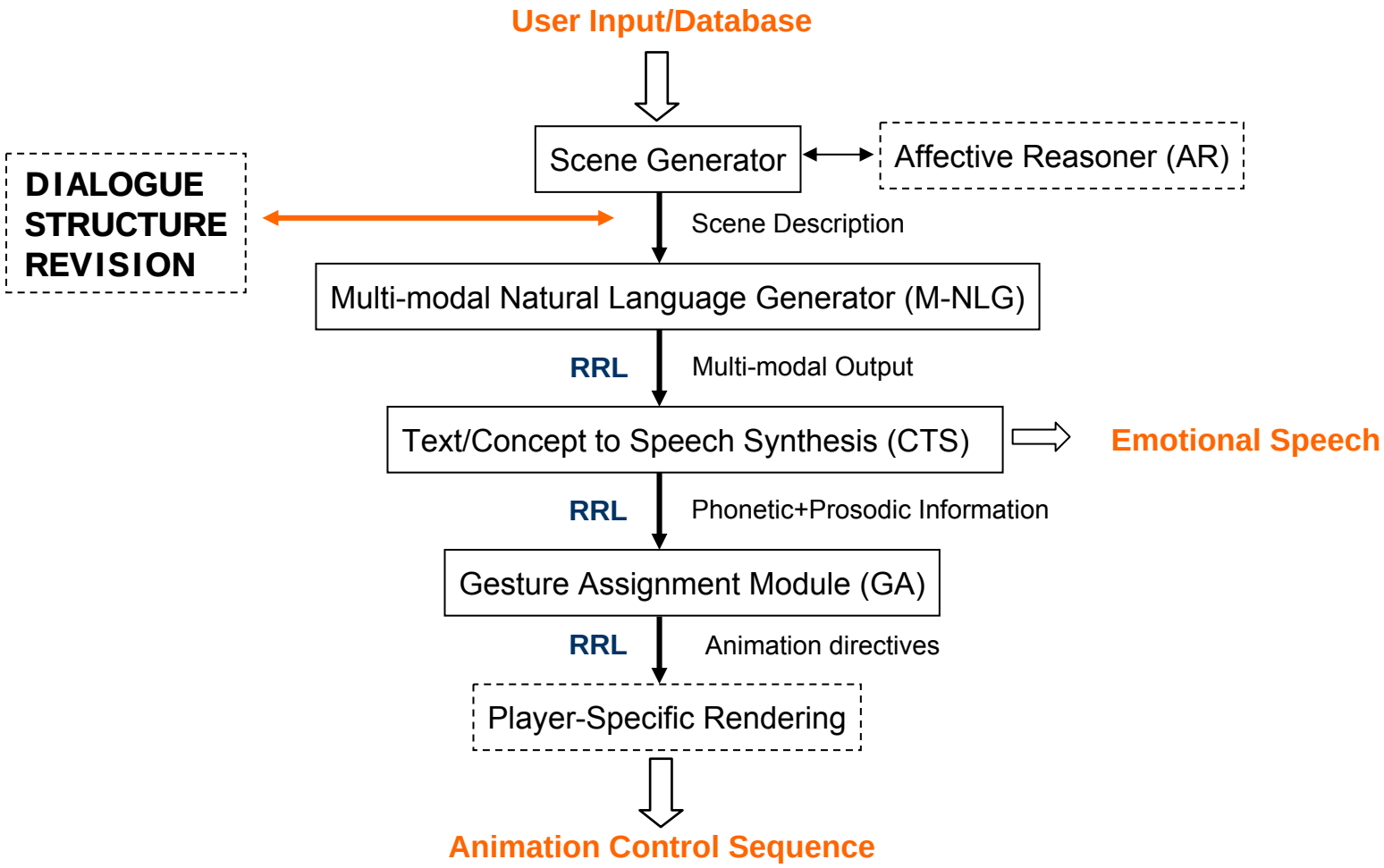
C: Does it have airbags and ABS?
S: Yes.
...

After applying the AGGR revision operation

3. How to apply revision operations?

Naive Sequential Strategy

- Take output of conventional top down planner.


User Input/Database
Scene Generator
Affective Reasoner (AR)
DIALOGUE STRUCTURE REVISION
Scene Description
Multi-modal Natural Language Generator (M-NLG)
RRL
Multi-modal Output
Text/Concept to Speech Synthesis (CTS)
Emotional Speech
RRL
Phonetic+Prosodic Information
Gesture Assignment Module (GA)
RRL
Animation directives
Player-Specific Rendering
Animation Control Sequence

3. How to apply revision operations?

Naive Sequential Strategy

- Take output of conventional top down planner.
- Apply operations as folllows:
  - If EMPH = Max then apply INSERT as often as possible;
  - If EMPH = Min then do not apply INSERT;
  - If TURN = Max then do not apply AGGR;
  - If TURN = Min then apply AGGR as often as possible.
- Problems:
  - Ordering: AGGR, INSERT
  - TURN = Min, EMPH = Max

3. How to apply revision operations?

Generate and Test Approach

1. Take dialogue plan from conventional dialogue planner.
2. Generate all possible plans that can be generated by applying the revision operations zero or more times in any order to the initial plan.
3. Assign a score to each of the revised plans.
4. Select the best plan, according to some arbitration scheme.

3. How to apply revision operations?

Assigning scores: Each plan p is assigned a tuple $\langle score_T(p), score_E(p) \rangle$ each in [0,100] such that:

- If TURN=Max then:
  - $score_T$(Shortest Plan) = 0
  - $score_T$(Longest Plan) = 100
- If EMPH=Max then:
  - $score_E$(Most Emphasis Plan) = 100
  - $score_E$(Least Emphasis Plan) = 0
- Etc.

3. How to apply revision operations?

- plan with maximal $score_E(p) + score_T(p)$

3. How to apply revision operations?

- plan with maximal $score_E(p) + score_T(p)$
- **Problem**: (45,45) loses from (90,10).

3. How to apply revision operations?

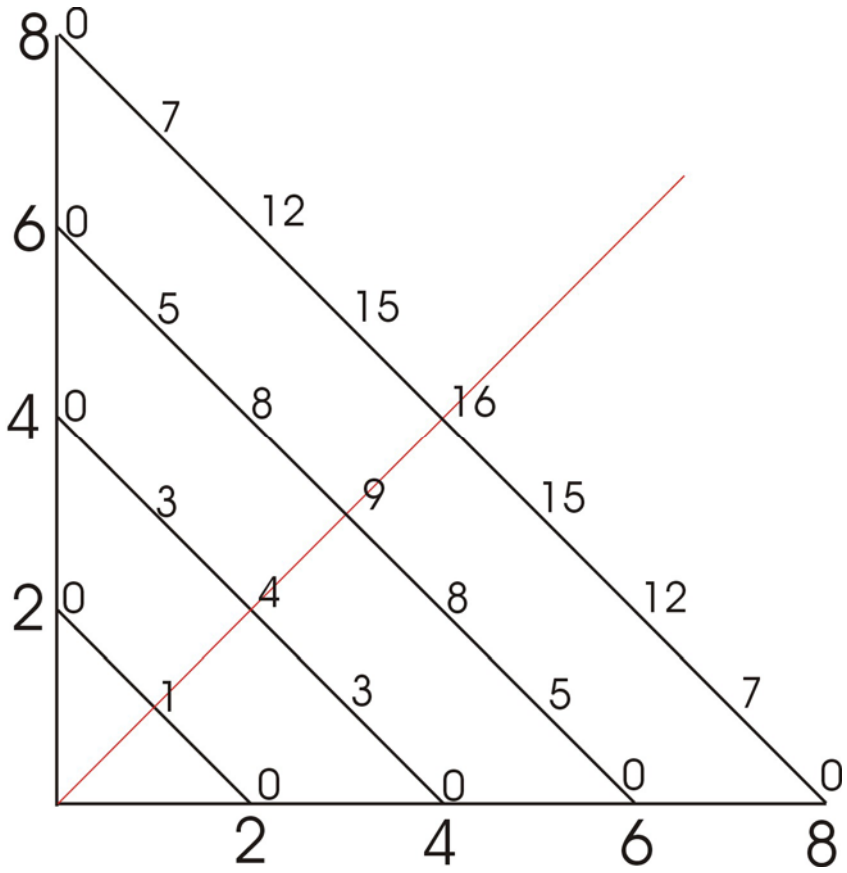
Nash Bargaining Solution:

plan with maximal $score_E(p) \times score_T(p)$ wins

3. How to apply revision operations?

Nash Bargaining Solution:

plan with maximal $score_E(p) \times score_T(p)$ wins



3. How to apply revision operations?

- **Summing Scores**: Individual Constraints do not matter
- **Multiplying Scores (NBS)**: Fair/balanced treatment of individual constraints (analogous to a fair solution for negotiation/bargaining situations)

Summary NECA

- **Starting point**: generating dialogue as discourse
- **For**: performance by team of agents
- **Architecture**: pipeline which incrementally specifies a script
- **Extension**: revision module to deal with global constraints

TILBURG UNIVERSITY

The Open University

For more information see ...

http://www.itri.bton.ac.uk/projects/neca

Plan

1. Introduction: NLG for Embodied Agents
2. What is Natural Language Generation?
3. Case: Generation of Referring Expressions (GRE)
4. NLG for Embodied Agents: Requirements
5. Case: GRE for Embodied Agents
6. Realization: Audiovisual Speech and Gestures
7. Case: NLG for Teams of Embodied Agents
8. Evaluating NLG for Embodied Agents

Plan

1. Introduction: NLG for Embodied Agents
2. What is Natural Language Generation?
3. Case: Generation of Referring Expressions (GRE)
4. NLG for Embodied Agents: Requirements
5. Case: GRE for Embodied Agents
6. Realization: Audiovisual Speech and Gestures
7. Case: NLG for Teams of Embodied Agents
8. Evaluating NLG for Embodied Agents

Evaluation?

- Why?
- How?

Why?

- Because I want to find out how good my ECA is (compared to alternatives).
- Because I want to *learn* how to build better ECAs. We want to move from a craft/art to a science of building ECAs.
- When? Formative versus Summative.
- Are there other options? Evaluation is one possibility: it is assumed that we can also learn from human (conversational) behaviour, because humans treat “computers as social actors” (CASA; Reeves & Nass, 1996).

What can we learn from an evaluation?

- Design guidelines which relate

  1. Parameters of the design of the agent (look, behaviours) and
  2. Context of use of the agent (type of user, task) to
  3. Evaluation parameters (does the agent achieve a particular effect, goal?)

  E.g.: Extrovert agents are *liked better* by extrovert users (than by introvert users) (Nass & Lee, 2000)

  In a bank setting, a clerk with formal dress is perceived as more *trustworthy* than one in jeans.

- More abstract guidelines: e.g., keep design parameters consistent with each other.

How? Design parameters (Ruttkay et al., 2004)

- Embodiment
  - Look
  - Communication modalities
    - language, text or speech, facial display (e.g., lip-sync), body, modality coordination and motion generation
- Mental Aspect
  - Personality
  - Social Role
  - Emotions
  - Adaptation to the user
  - Discourse capabilities
  - Control
  - Input modalities of the user
- Technical aspects
  - Animation technology; speech synthesis method, etc.

How? Context of use parameters (Ruttkay et al., 2004)

- Education, information giving, sales, …
- User characteristics:
  - Demographic
    - Gender, age, ethnicity, language, computer skills, familiarity with ECA technology
  - Psychological data
    - Personality, Affect intensity, Cognitive style, Perception and body capabilities
  - Culture

How? Evaluation parameters/dimensions (Ruttkay et al., 2004)

- Usability
  - Learnability, Memorizability and Ease of Use
  - Efficiency
  - Errors

- Evaluation of User Perception of ECAs
  - Satisfaction
  - Engagement
  - Helpfulness
  - Naturalness and believability
  - Trust
  - Perceived Task Difficulty
  - Likeability
  - Entertainment

How? Methods

Evaluation parameters from two perspectives (Dehn & Van Mulken, 2000):

1. User's behaviour during interaction
2. User's subjective perception of the interaction

- Observation (1)
- Experiment (1)
- Questionnaire (2)
- Interview (2)
- Usage data (1)
- Biomedical data (1)

Long term or short term effects?

How? Some Problems

- Separating the ECA from the application.
- Limited comparability of results owing to
  - vague terms: fun, likeability, friendliness
  - ambiguous terms: believable (alive versus believable actions)
- Many factors (e.g., non-essential technologies) that can influence the evaluation variables.
- See Dehn & Van Mulken (2000) for critical review of existing evaluations.

Illustration: Evaluation using physiological responses

- Prendinger, Mori & Ishizuka (2004)
- Learn about: The impact of emphatic ECA behaviour on the user
- How: 1. measure physiological response of user to emphatic behaviour, and 2. questionnaire

- Measure used: Galvanic skin response. Indicator of skin conductivity. Various with level of arousal and increases with anxiety and stress (Picard, 1997; Healey, 2000). Wilson & Strasse (2000): increase as result of low frame rates (in videoconferencing) not noticed by users.

Setup

- **AGENT**: Virtual quiz master Agent
- Technology: MSAgents
- Language and culture: Japanese

- **TASK**: Five numbers are displayed after each other. Competition for prize: The subject has to sum them and subtract the *i*th number (i<5) and get it right. The system occassionally delayed the presentation of the 5th number to frustrate the user.

- **CONDITIONS**:
  - Affective version: correct/incorrect: happy for and sorry for (speech and face); delay: express empathy through apology.
  - Non-affective version: Only say “right” or “incorrect”.

- Measurements: DELAY, (last number & user answer) RESPONSE_TO_ANSWER (correct/incorrect), RESPONSE_TO_DELAY (express empathy or segment following RESPONSE_TO ANSWER).

Results

GSR measurements

- GSR difference between RESPONSE_TO_DELAY than DELAY higher for Affective version (suggesting reduction of frustration).
- RESPONSE_TO_ANSWER: no significant difference.

Objective performance no significant differences

Questionnaire

	NA	A
Difficulty	7.5	5.4
Frustration	5.2	4.2
Enjoyment	6.6	7.2

Done

1. Introduction: NLG for Embodied Agents
2. What is Natural Language Generation?
3. Case: Generation of Referring Expressions (GRE)
4. NLG for Embodied Agents: Requirements
5. Case: GRE for Embodied Agents
6. Realization: Audiovisual Speech and Gestures
7. Case: NLG for Teams of Embodied Agents
8. Evaluating NLG for Embodied Agents

Done

1. Introduction: NLG for Embodied Agents
2. What is Natural Language Generation?
3. Case: Generation of Referring Expressions (GRE)
4. NLG for Embodied Agents: Requirements
5. Case: GRE for Embodied Agents
6. Realization: Audiovisual Speech and Gestures
7. Case: NLG for Teams of Embodied Agents
8. Evaluating NLG for Embodied Agents

Done

1. Introduction: NLG for Embodied Agents
2. What is Natural Language Generation?
3. Case: Generation of Referring Expressions (GRE)
4. NLG for Embodied Agents: Requirements
5. Case: GRE for Embodied Agents
6. Realization: Audiovisual Speech and Gestures
7. Case: NLG for Teams of Embodied Agents
8. Evaluating NLG for Embodied Agents

Finally…

- **Acknowledgements**: Thanks are due to Marc Swerts, Ielka van der Sluis, Lennard van der Laar, Richard Power and Kees van Deemter for their help.

- For *more information* on

  - Functions of Audiovisual Prosody:
    foap.uvt.nl

  - Generation of Referring Expressions: **www.csd.abdn.ac.uk/~agatt/tuna/**

  - NECA Generation for Teams of Agents
    www.itri.bton.ac.uk/projects/neca/ and **mcs.open.ac.uk/pp2464**

- *Slides* will be available at: **mcs.open.ac.uk/pp2464/easss05/**