Natural dialogue in natural language

Frédéric Landragin

INF642

8 January 2021

Content

1. Definitions and examples
2. Dialogue models – management, control, history
3. Speech acts – dialogue act processing
4. Referring acts – multimodal reference resolution

First part

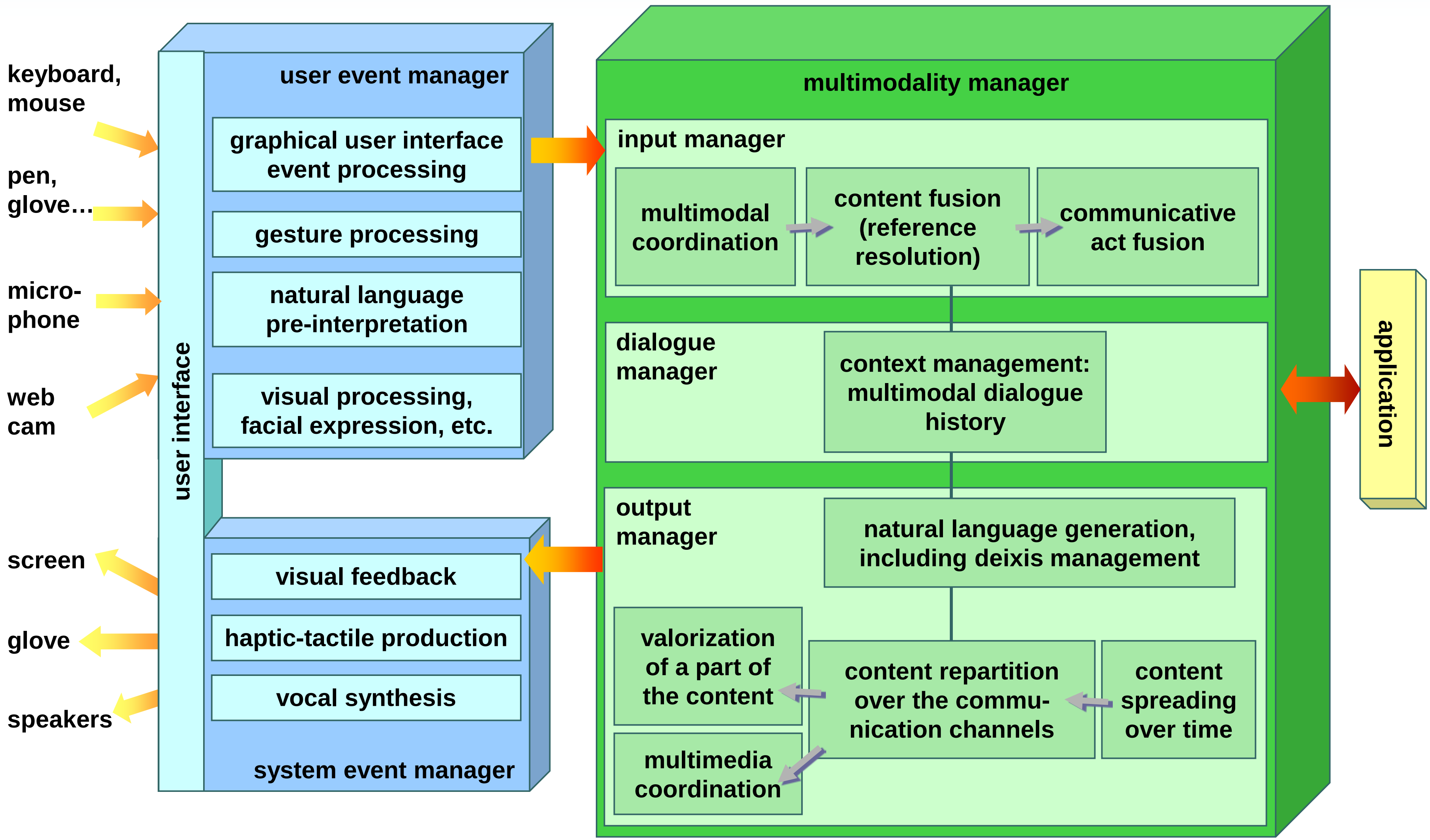
1. Definitions and examples
   Human-machine dialogue: between linguistics and computer science, between NLP, AI, QAS, ECA, GUI and cognitive sciences.
   Examples : ELIZA, SHRDLU, SRAM, OZONE, MIAMM…
   And what about machine learning?

2. Dialogue models – management, control, history

3. Speech acts – dialogue act processing

4. Referring acts – multimodal reference resolution

General architecture



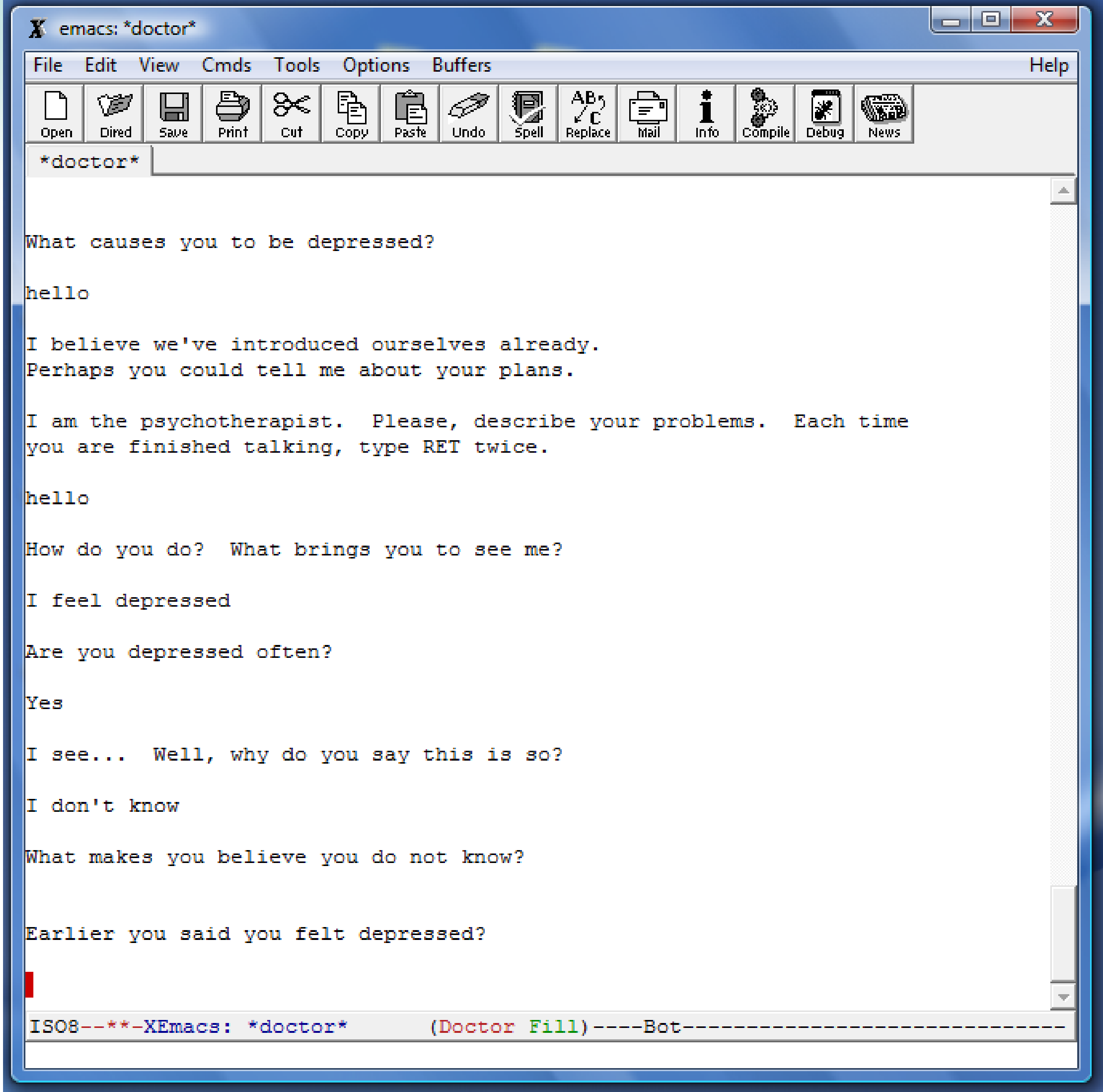
Definitions (human communication)

- Natural dialogue in natural language
  - Natural language $\Rightarrow$ lexicon, syntax, prosody, semantics, pragmatics
  - Natural dialogue $\Rightarrow$ human dialogue structure, coherence, cohesion
- Multimodality
  - Speech
  - Gesture
  - Visual context
- Pragmatics
  - 1st level pragmatics = indexicality, reference to objects
  - 2nd level pragmatics = literal meaning and communicated meaning
  - 3rd level pragmatics = speech acts theory

Definitions (Human-Machine Dialogue)

- Task-oriented dialogue / interaction
  - Train ticket reservation = task-oriented dialogue
  - IBM Watson (Jeopardy!) = task-oriented interaction
- Dialogue in open domain / in closed domain
  - Train ticket reservation = closed domain
  - IBM Watson and QAS = open domain
- Context
  - Domain context
  - Visual context: communication situation
  - Linguistic context: dialogue history

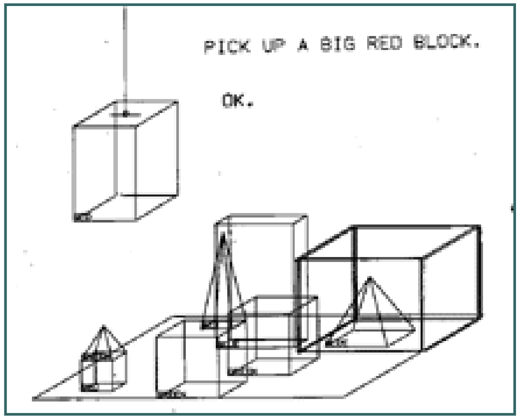
- ELIZA (1966)
- Simplest approach
  - deictic
  - sentence templates
  - typical answers
  - memory
  - lexicon (family terms)



Reference to objects in dialogue

- SHRDLU (1970)
  - Person: PICK UP A BIG RED BLOCK.
  - Computer: OK. (does it)
  - Person: GRASP THE PYRAMID.
  - Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.
  - Person: FIND A BLOCK WHICH IS TALLER THAN THE ONE YOU ARE HOLDING AND PUT IT INTO THE BOX.
  - Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.

    Computer: OK. (does it)
  - Person: WHAT DOES THE BOX CONTAIN?
  - Computer: THE BLUE PYRAMID AND THE BLUE BLOCK.
  - Person: WHAT IS THE PYRAMID SUPPORTED BY?
  - Computer: THE BOX.
  - Person: HOW MANY BLOCKS ARE NOT IN THE BOX?
  - Computer: FOUR OF THEM.



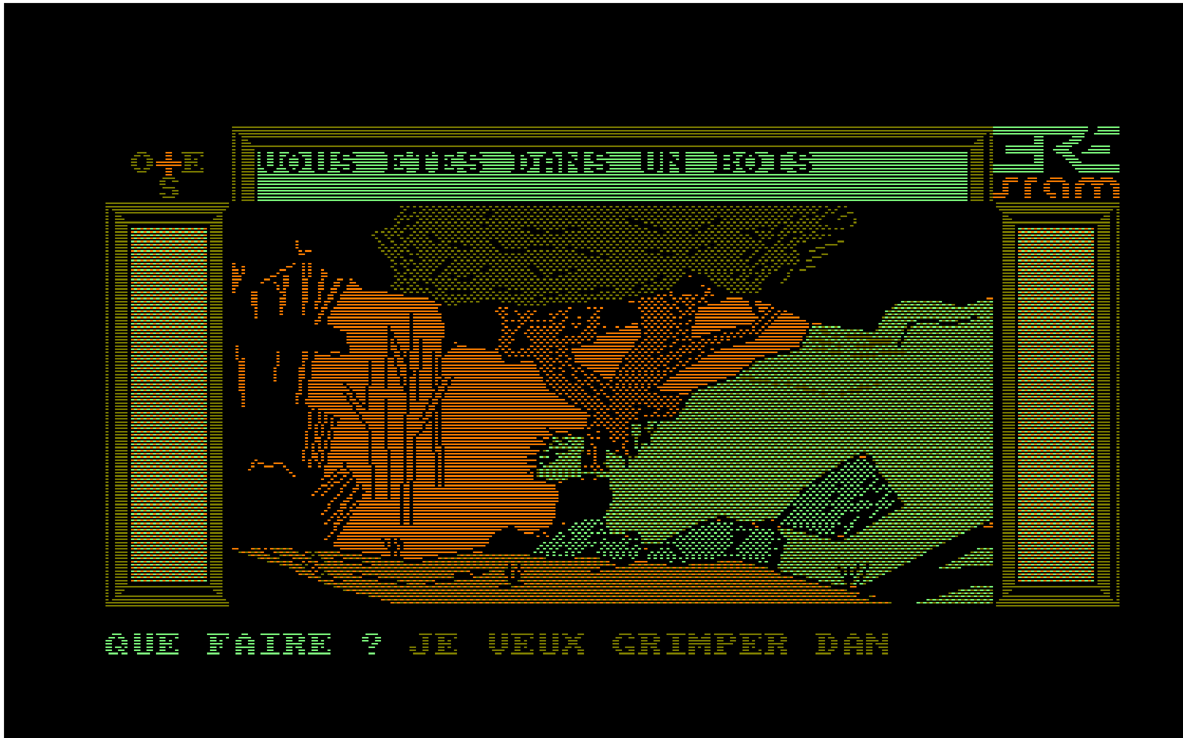
Models in SHRDLU system

- Models for the physical world
  - typology of geometrical forms, colours, sizes
  - spatial relations, inclusion, superposition
  - simple physical model (gravity)
- Models for linguistic reference resolution
  - model for determination (indefinites, definites)
  - singularising properties
  - model for the interpretation of pronouns
  - most complex case: “a block which is taller than the one you are holding”
- Models for human-machine interaction
  - model for speech acts processing (orders and questions)

Models in SRAM video game

(1985, first textual adventure game in France)

- Model of the world: very simple, only the visible objects
- Interaction model: the user has to answer to “what to do?”
- Language model, also very simple:
  - made to be robust
  - detection of a verb and a complement
  - lexicon = list of verbs, list of nouns, and list of possible associations verb + noun
  - typical sentences; time, repetitions, and insults management



Models in IBM Watson system
(2011, first participant to a TV game)

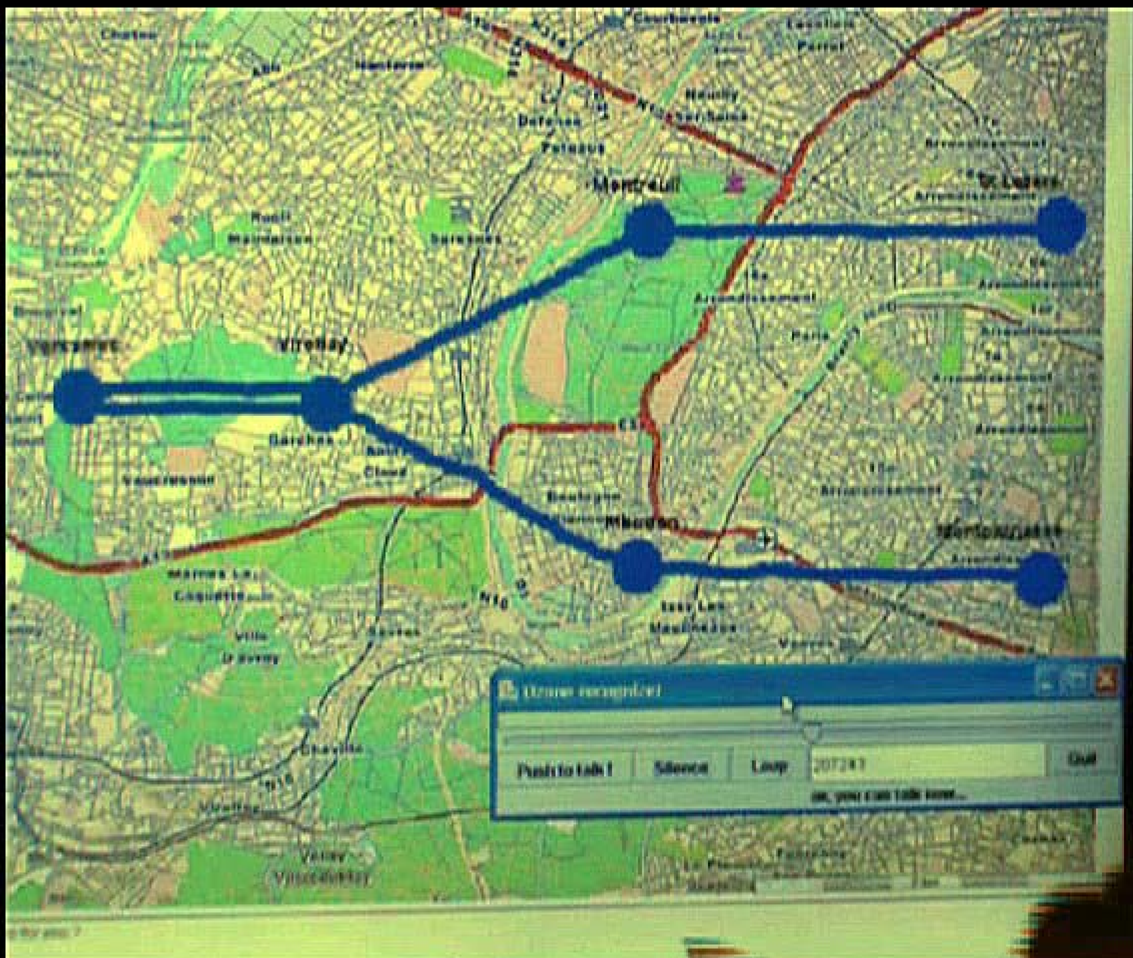
- **Context:** Jeopardy! TV game = the presenter utters an answer and the candidates have to find the related question
- **Interaction model:** the rules of the game + a set of behaviours, with models such as states diagrams, decision trees, typical sentences…
- **Language models:** numerous and complex…

IBM Watson (DeepQA project)

- Models of the world = in a huge local database:
  - NB: 2880 processors system, with 15 TO of RAM for the database
  - models for general knowledge, models for the links between knowledge, models for the management of several information sources
  - algorithmic models (NLP oriented) and strategic models (game oriented)
  - models for speech recognition and text-to-speech synthesis

Train ticket reservation example

from the 2004 OZONE project (not real time!)



- Touch screen (with a pen)
- Voice input (with a ‘push-to-talk’ button)
- Free software for speech synthesis
- Simple multimodal references processing
- Simple speech acts processing

General methodology for HMD

- Specifying the system's task, roles and abilities
- Specifying a first set of covered phenomena
- Carrying out experiments (Wizard of Oz) and corpus studies
- Specifying the conceptual architecture and related processes
- Resources writing and development (NB: the modules are not implemented in the same order than the processing chain)
- User tests
- Evaluation (more complex methods than for any other NLP system)

Advanced methodology for HMD

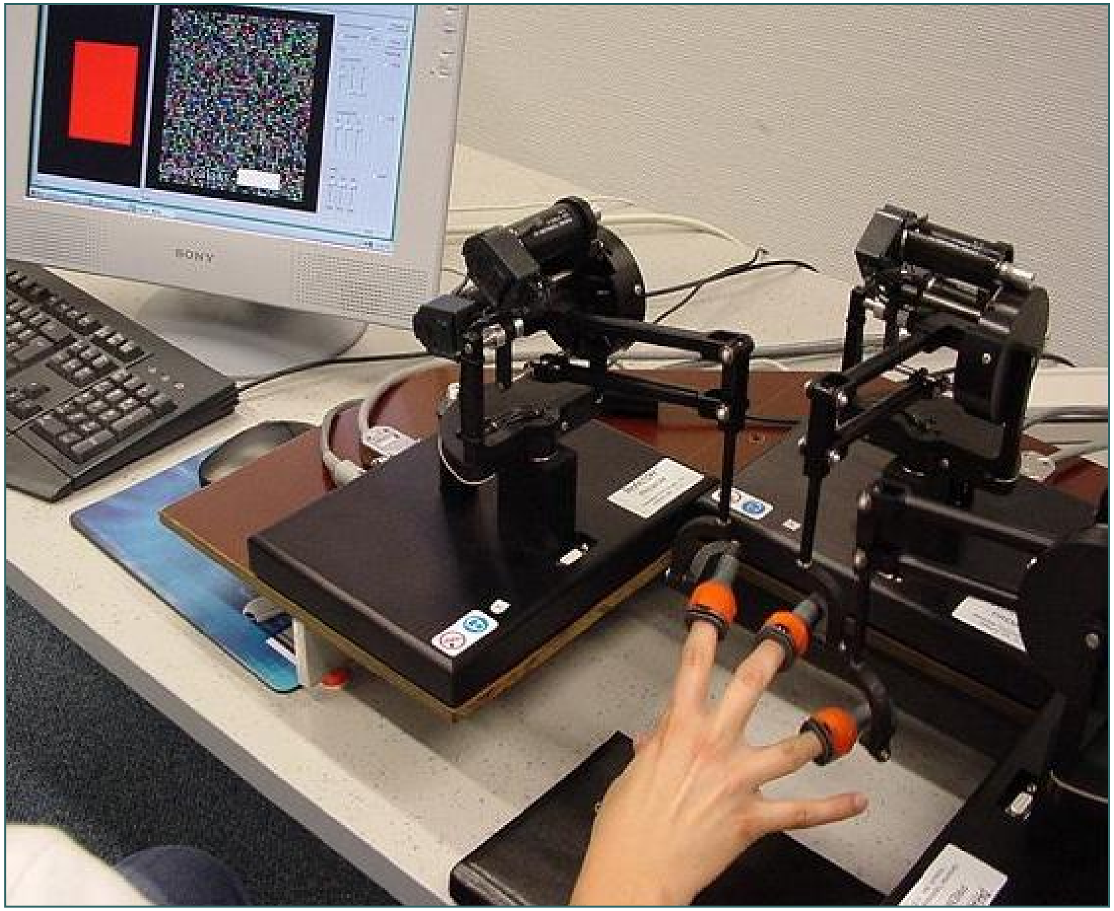
- Reference to objects
  - How a human user refer to objects in a multimodal context?
  - Research on possible phenomena linked to particular devices
- Speech acts
  - How to do things with words?
  - Research on composite speech acts and multimodal dialogue acts
- Dialogue systems architecture and design process
  - Exploring reusability and adaptability to a new task
  - Specifying modules and protocols (MultiModal Interaction Language)
  - Applying machine learning techniques for each module
  - Applying machine learning techniques for the whole system

MIAMM particular interaction

Virtual:
PDA with force feedback buttons

Multidimensional Information Access using Multiple Modalities

For real:
Phantom devices for haptic-tactile interaction



What about machine learning?

- Reference to objects
  - "Show me the cup": reference with continuous representations
- Speech acts
  - Speech act classification using a convolutional neural network
- Natural language processing and dialogue processing
  - **Word2Vec**: used to produce word embeddings in order to apply easily computable methods (from mathematics)
  - **Seq2Seq**: turns one sequence into another sequence - first designed for machine translation, then applied to other applications like HMD
  - **Mem2Seq**: effectively incorporating knowledge bases into end-to-end task oriented dialogue systems
  - **WMM2Seq**: working memory model for dialogue response generation

Modern HMD systems

- Speech recognition
  - Dedicated neural network
- Intention identification (among 80)
  - Weather, traffic…
  - Call someone, play music…
  - Dedicated neural networks (“transformers”)
- A dedicated process/server for each intention
  - “Alexa, call Mom!” $\rightarrow$ simple system using a set of rules
  - “Alexa turn the volume up in the kitchen!” $\rightarrow$ more complex system
  - Possibly, an additional neural network can be used for advanced order processing
  - “What is the height of the Eiffel tower?” $\rightarrow$ search engine

Second part

1. Definitions and examples
2. Dialogue models – management, control, history
   How can a dialogue system determine relevant answers?
   What information has to be saved, i.e., what is the content of the dialogue history?
3. Speech acts – dialogue act processing
4. Referring acts – multimodal reference resolution

Dialogue manager

- DM = heart of the HMD system
  - where all the machine understanding process results arrive
  - where all the different information is compared, including that stored as the dialogue progressed
  - where decisions are taken, which might involve a problem resolution phase or a database information research phase
  - where the messages for the user are generated as an abstract form which is then materialized by the modules in charge of automatic generation and synthesis
- The dialogue manager makes decisions
  - which go in the direction of a natural dialogue with the user
  - which manifest a perceptible ‘persona’ to the user that is visible in the semantic content uttered and the method of its transmission

Dialogue?

- To have a dialogue is to do everything to continue having a dialogue, as long as the goal has not been reached
- The dialogue manager materializes the cooperation
  - ‘cooperative’ dialogue system = answers to the user, but remains within its role of system in charge of satisfying the user’s queries
  - ‘collaborative’ dialogue system = shows an increased level of cooperation, for instance suggesting a change of direction to the user (and therefore leading to the joint construction of a common goal)
- When listening to the user, the dialogue system is able to
  - produce backchannels
  - show its attention (eye direction, body movements)

Some examples

- Choosing the right (and not too long) answer
  - U: “how long with this itinerary?” – S: “two hours”
  - U: “how long with this itinerary?” – S: “two hours due to a change at Versailles”
  - U: “how long with this itinerary?” – S: “two hours, due to a change at Versailles, but if you go through Meudon you'll get there in fifty minutes”
- Need for variability and naturalness, not only within an answer
  - U: “I would like to go to Paris”
    S: “what day?”
    U: “it will be tomorrow”
    S: “what time?”
    U: “let's say around nine o'clock”
    S: “what class?”
    U: “first, please”…

Backchannels

- Two channels
  - the main channel = speaker leading the dialogue at a given moment
  - the backchannel = occupied by listening hints given by the hearer
- Examples: “hmm”, “yes”, “oh really”…
- Backchannels and TRP (transition-relevance place)
  - because they are brief and do not really constitute a speaking turn, they do not prevent the speaker from going on with his intervention
  - This can thus lead to voices being superimposed
  - Many technical issues: speech recognition works at the same time as the system is making noise; a backchannel can happen at the moment when the user was about to stop talking; a backchannel is not always produced at the most relevant moment, due, for example, to the slight temporal delay with regard to the best places, the TRP

Theoretical models for DM
(often exploited *a posteriori*)

- Linguistic models
  - the Geneva model of discourse analysis (Roulet *et al*., 1985) = emphasizes general indication so that the dialogue progresses toward the satisfaction of both speaker and hearer
  - argumentative acts (Moeschler, 1985) = the linguistic form contains instructions (connectors: “but”, “however”, “thus”, “because”, “so”)
  - Relevance Theory (Sperber & Wilson, 1986)
  - coherence and cohesion
- Socio-cognitive models
  - conversational analysis (Sacks *et al*., 1974) = diversity, in human dialogue, of speaking turn organization phenomena, sequence and segment organization phenomena, reparation organization phenomena
  - (Allwood *et al*., 2000) = cooperation principles with a number of criteria: joint purpose, trust, cognitive, ethical…

Speech acts = bricks for the DM

- Interrogative
  - when the user asks the system a question, the choice of answer is =
  - either it knows the answer and gives it
  - or it does not know the answer and apologizes, tries to redirect the user on another path or, potentially, asks the user to rephrase his question
- Imperative
  - when the user gives the system an order, the choice of answer is =
  - at the same time as the system carries out the order, it can generate an utterance which on the one hand announces the action being carried out (especially if it is invisible)
  - and on the other hand allows the user to follow-up and continue the natural dialogue

Speech acts = bricks for the DM

- Assertive (including performative)
  - when the user generates an assertion, such as “I do not have a senior discount card”, this is the point where the system must prove it is able to manage a dialogue:
  - an assertion contributes a seemingly new piece of information to the system or it is useless, and this newness must trigger inferences
  - on the basis of these inferences, and on what has already been said and what both speaker and hearer know, the system should understand if the assertion is filling a lack which was blocking a situation
  - which the inferences unblock and will allows the system to know what to answer
  - or, on the contrary, if the assertion should incite the system to suggest something such as “let us see if you can get one. How old are you?”

Computer models for the DM

- Dialogue management groups three phases that more or less overlap one another
  - **dialogue control** which tries to manage the interactive process to determine a type of reaction after a structured chain of utterances
  - **dialogue history** which groups what has been said and the way in which the utterance content is grounded
  - **dialogue initiative** which adds a specific behavior to the previous considerations
- “Grounded”?
  - **common ground** = information shared by the speaker and hearer, either because it mutually manifests, for example because it has been verbalized, or because it can be deduced from what has been said
  - U: “How to go to Paris?” – S: “To Paris. When would you like to leave?” “To Paris” = user's utterance is **grounded** by the system

Dialogue control

- Techniques (domain plans and/or discourse plans)
  - **finite states automata** and **dialogue grammars**: the point is to determine all the possible situations, and ways of going from one situation to the next
  - **patterns** and **templates**-based methods: allow to recognize situations without forcing the initiative to remain on one side or the other
  - methods based on an **information state**: add a memory, which contains anything required, knowing that this memory will help determine the possible follow-ups of the dialogue. Depending on the authors, the information state contains the dialogue history, common ground, a model with mental states, a model of the user, etc.
  - **joint action** theories (with the notion of common ground)
  - dialogue control inspired by **game theory**, each utterance being a turn played during which the speaker tries to maximize his gain
  - **machine learning** techniques

Machine learning for DM

- Techniques
  - **MDP**: Markov Decision Process
  - **POMPD**: Partially Observable Markov Decision Process, a model which extends that of the information state by adding a probabilistic means to decide on the future action depending on the current state
  - **reinforcement learning**: to achieve an optimal decision set. The decision rules are refined by carrying out several thousand of exchanges between the system and the user
  - but also techniques that are used for natural language processing
- Data
  - sets of complete dialogues, with (manual) annotations
  - lack of data? $\rightarrow$ user simulation

Dialogue history

- Modeling the dialogue context
  - storing what have been said
  - formalizing and storing in a dedicated structure the contextual interpretation results
  - highlighting in this history a structure describing the dialogue's progression, for example the path followed to solve the task
- Modeling the common ground
  - **communal common ground**, where the shared knowledge goes beyond the interaction between the speaker and the hearer
  - **personal common ground** in which the knowledge is only good for the speaker and the hearer
  - modeling the **grounding process**: 1. automatically update the common ground at each utterance, 2. only when a grounding criterion is reached (discourse contributions model), 3. grounding acts model

Dialogue history

- Other structures, in addition to the dialogue history
  - visual history
  - gesture history
  - multimodal history
  - **history page** (Vilnat, 2005) = contains the speaker's identifier for each intervention, the semantic and pragmatic representations, the related topic, the related goal, the state of the dialogue's structure, the state of the interaction variables, the state of the plan being developed…
- To simplify
  - assertive: **"saying that"** → updates the **CG, Common Ground**
  - imperative: **"telling to"** → updates the **TDL, To Do List**
  - interrogative: **"asking"** → updates the **QUD, Questions Under Discussion** (or Questions Under Debates)

Dialogue initiative

- Questions
  - choosing to give an information or not
  - choosing to display an information or not
  - choosing to highlight a part of the information or not
  - choosing what to say and how to say it
  - choosing the head movements, eye direction…
  - for multimodal dialogue and multimodal user interfaces
    → **IMMPS**, Intelligent MultiMedia Presentation Systems
- Towards (very) intelligent systems
  - able to clarify the conditions that allow the system to abandon its goals, decide on a path aiming to satisfy a specific goal and in general explore the interaction between goals, beliefs, and intentions
    (Cohen & Levesque, 1990)

Dialogue initiative

- Examples with train ticket reservation
  - if the system decides to present the details of 30 train itineraries, or to show a geographical map annotated with the elements answering the user's query, it must be able to control the way in which it transmits this information
  - this can be done by **planning**, that is allocating the transmission over several speaking turns
- (Horchani, 2007) proposed three strategies
  - **enumeration**, in the case in which the number of solutions is reasonable (but then a threshold has to be decided)
  - **restriction**: the system suggests criteria to limit the scope of the research space
  - **relaxation**: the system suggests either alternative solutions or alternative research criteria

Conclusion of the second part

- Natural dialogue in natural language
  - the task to be accomplished is the driving force behind the dialogue: the dialogue progresses when the task progresses
  - however, a realistic human-machine dialogue should not be built around this priority alone: it also has to take into account fluidity and the linguistic spontaneity of exchanges
- Speech acts are the bricks for the dialogue management
  - but there are not always as simple as interrogative, imperative and assertive
    $\rightarrow$ third part

Third part

1. Definitions and examples
2. Dialogue models – management, control, history
3. Speech acts – dialogue act processing
   How to do things with words?
   How to derive meaning out of direct and indirect speech acts
4. Referring acts – multimodal reference resolution

Content of the third part

- Objectives
- Speech act processing for human-machine dialogue
  - Collaborativeness in human-machine dialogue
  - Existing dialogue systems
- Speech act segmentation
  - Indirect speech act processing
    - Process parameters
    - Example
  - Composite speech acts processing
    - Process parameters
    - Example
- Conclusion

Objectives and underlying questions

- Indirect and composite speech act processing:
  - Does the system need to make the distinction between surface act and profound act?
  - Does the system need to identify the various acts in a composite speech act?
  - To which act the system has to react?
  - What are the criteria for making a choice between the various possibilities?
- Data that are mandatory to the process:
  - When interpreting complex speech acts, does the system need to make hypotheses on the user's mental states? (BDI)
  - What kind of structures has the system to manage? (CG, TDL, QUD)
  - What are the simplest (with an implementational point of view) configuration and process?
- Links between speech act processing and collaborativeness:
  - By interpreting indirect and composite speech acts, does a system increase the collaborative aspects of the human-machine dialogue?
  - Which collaborative criteria can help the system to resolve indirect speech acts?

Complex speech acts: definitions and main principles

Indirect speech acts

(saying something with the appearance of something else)

- Example: “can I have a train ticket to go to Paris?”
  surface act = question
  profound act = request
- The surface act is useful to determine the linguistic form of the answer
- The profound act is useful to determine the content of the answer

Composite speech act

(how to do several things with one utterance)

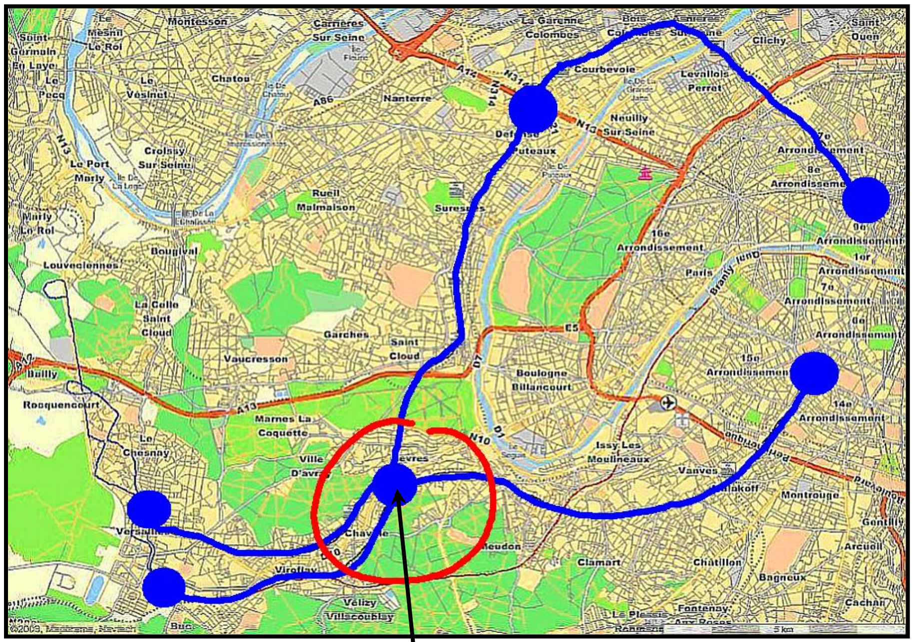
- Example: “how much time with this way that seems to be the shortest one?”
  primary act = question: “how much time with this way?”
  secondary act = comment: “this way seems to be the shortest one”
- The system must answer to the primary act
- It can also treat the secondary act, for instance with a confirmation of the comment

Speech act processing in human-machine dialogue

- When interpreting as well as when generating, to the message content is added an illocutionary force that expresses the act that is realized by the enunciation, and that depends on an underlying intention
  - **“saying that”** = speaker expresses an **assertion** in order to **make** the addressee **know** something
  - **“telling to”** = speaker expresses a **demand** in order to **make** the addressee **do** something
  - **“asking”** = speaker expresses a **question** in order to **know** something from the addressee (two cases: the close question, i.e., ask-if, whose answer is yes or no, and the open question, i.e., ask-WH, whose answer is an information)

- Example in natural language and multimodal generation
  - the way of generating depends on the illocutionary force:
    if we just want to **inform** (say that), we can use a certain way of generating that totally differs from the way of generating we use to **encourage to act** (tell to)
  - the dialogue system may need a confirmation of the message reception:
    then we can distinguish a “say that without feedback” from a “say that with a mandatory feedback” (in user interfaces it is the use of the “OK” or “OK/Cancel” dialogue boxes)
  - that can be modeled using composite speech acts:
    “say that with feedback” = “**say that** information” + “**tell to/ask if** feedback”

A transport application interface and scenario



Meudon

Paris

User: “I want to go to Paris”

System: “You can take the following train lines” + *display*

U: “How much time with this way that seems to be the shortest one?” + *gesture*

S: “40 minutes”

U: “OK, can I buy a train ticket?”

S: “Yes, a single or return ticket?”

U: “single”

S: “OK […]”

Searle’s indirect speech act processing

10 steps for the derivation of the profound act from the surface act (exchange: “let’s go to the cinema this evening”, “I have to work for my exam”)

1. conversational facts: the proposal and the answer to it
2. collaborative conversational principles : the answer must be relevant
3. possible speech acts for the answer: acceptation, reject, counter-proposal…
4. by inferring 1 and 2, there is a contradiction between 3 and the literal form of the answer
5. by inferring 2 and 4, there is an indirect speech act
6. background factual information: time that working and going to the cinema require
7. inference from 6: working and going to the cinema are not reconcilable in the same evening
8. preparatory condition for a proposal acceptation = capability to accomplish the act
9. inference from 1, 7 and 8: the answer implies the incapability to accomplish the act
10. profound act of the answer = proposal reject

Intentional states and related acts:

1. belief: affirmations, assertions, remarks, postulations, deductions, arguments...
2. intention: promises, oaths, menaces, engagements
3. desire: requests, orders, commands, demands, supplications...
4. pleasure: compliments, congratulations...

Collaborativeness

Grice’s maxims

- deal with collaborativeness but are impossible to implement in a human-machine dialogue system because they are too general
  (how can “be relevant” be implemented?)
- such principles seem to intervene at another level than semantic contents and pragmatic forces

Dialectic functions in argumentative dialogue

- to each utterance correspond:
  1. a semantic content
  2. an illocutionary force
  3. a dialectic function (evaluation of the message with an argumentative point of view)
- examples of dialectic functions:
  ‘arg-pro-mt’ = the speaker provides an argument in favour of his own thesis
  ‘req-tdp-mt’ = the speaker asks his opponent to take a position with respect to his thesis
- managing dialectic functions is a way to increase collaborativeness but is for now too hard to implement

Collaborative aspects in human-machine dialogue

Collaborativeness degrees

- communicant system (answers to questions, executes orders)
  "6 hours" is the simplest answer to the user's question "how much time to go to Massy?"
- cooperative system (each interlocutor has an aim and the roles are not reversed)
  "6 hours because of a train change at Bordeaux"
- collaborative system (co-construction of a common objective)
  "6 hours because of a train change at Bordeaux, but if you go to Paris instead of Massy, it will take you only 5 hours"

A collaborative system may verbalize

- control utterances
  "I understand your request"
- its own intentions
  "I am going to search for trains from Toulouse to Massy"
- its own beliefs
  "but I was believing you wanted to go to Paris!"

Existing dialogue systems

- Command dialogue and query dialogue:
  - Only one speech act is possible
  - Whatever the form of an utterance, its illocutionary force is brought back to the expected one
- Combined dialogue with static identification of speech acts:
  - Several speech acts are possible, and the identification relies on a hashtable that links syntactic forms with speech acts
  - The hashtable is predefined
- Combined dialogue with dynamic identification of speech acts:
  - Several speech acts are possible, and the identification relies on :
    - rules such as conversational postulates
    - statistic criteria with machine-learning techniques
    - reasoning with the user's mental states

Speech act segmentation

Implementational aspects

- The recognition module provides utterances that correspond to a noisy signal between two silences $\Rightarrow$ this is a first unit-related constraint
- The syntactic and semantic modules are often designed for the process of simple utterances (sometimes with only one verb)
- The last step of interpretation consists of resolving references to actions (making the link between the result of the semantic/pragmatic analysis and the application primitives)
  sometimes: to one speech act corresponds one executive function of the application

Theoretical aspects

- One composite speech act when several discourse segments can be identified
- One composite speech act when several distinct meanings can be identified
- Example: appositives
  "We called Dr. Hughes, a famous scientist" can be interpreted as a composite speech act
- Counterexample: restrictive clauses
  "I dislike driving in a town where there are no stop signals" is a simple speech act

Parameters for indirect speech act processing

Determining if a speech act is indirect or not will rely on:

1. linguistic and semantic characteristics of the utterance
   imperative, interrogative, indicative + semantic category of the the verb, etc.
2. the dialogue history
   the previous utterances and their interpretations (the user can use frequently the same form of utterance with the same illocutionary aim)
3. a lexicon of dialogue pairs
   categories of acts and list of possible reactions
   (we answer to a proposition by an acceptation, a reject, or a counterproposition)
4. classical conventional uses
   constructions such as “can you give me the salt?”
5. the list of system abilities and all task constraints
   if the system helps the user finding a restaurant, a question like “can you list me the Chinese restaurants near Palaiseau” is of course a request
6. hypotheses on the user’s mental states
   in particular, does he knows the answer to his question? if certainly yes, it is an additional argument for interpreting the question as an indirect speech act

Details for the first exchange of the example

U: “I want to go to Paris”

Semantic analysis: say_that (U, S, want (U, go_to (Paris)))

Roles and abilities of the system:

help (S, U)

sell (S, U, train_tickets_to (Paris, Meudon…))

know (S, means_to_go_to (Paris, Meudon…))…

Hypotheses on the user’s mental states:

desire (U, go_to (Paris))

desire (U, help (S, U, go_to (Paris))

desire (U, know (U, means_to_go_to (Paris)))

Identification of an indirect speech act + collaborativeness:

help (S, U, go_to (Paris))

know (S, means_to_go_to (Paris)))

Determination of the profound act: ask_for_value (U, S, λx.means_to_go_to (Paris) (x))

Reaction from the system: say_that (S, U, means_to_go_to (Paris) = {way_1, way_2…})

S: “You can take the following train lines” + *display*

Parameters for composite speech act processing

Determining if a speech act is composite and not, and determining which is the primary act, will rely on:

1. linguistic and semantic characteristics of the utterance
   epithets, evaluative adverbs, appositions, subordinate clauses, etc.
2. the dialogue history
   when the same composite act is produced again by the user, the same classification can be made by the system
3. a lexicon of dialogue pairs
   categories of acts and list of possible reactions
   (to a composite act like a "question + comment", we react with the answer to the question and/or with a confirming or infirming reaction to the comment)
4. classical conventional uses
   constructions such as "who **the hell** did that?"
5. task constraints (= ordering the user's requests using their importance for the task)
   the primary act is the most relevant to the ongoing task
6. hypotheses on the user's mental states
   the primary act is the one that has the most important contextual effects to the hypotheses on the user's mental states

Details for the second exchange of the example

U: "How much time with this way that seems to be the shortest one?" + *gesture*

Analysis of the main proposition and multimodal reference resolution:

ask_for_value (U, S, λx.travel_duration (way$_1$, x))

Analysis of the comment: say_that (U, S, way$_1$= shortest_travel_to_go_to (Paris))

Hypotheses on the user's mental states:

desire (U, know (U, travel_duration (way$_1$)))

belief (U, way$_1$= shortest_travel_to_go_to (Paris))

desire (U, know (U, way$_1$= shortest_travel_to_go_to (Paris)))

Abilities of the system:

it knows the answer to the question: know (S, travel_duration (way$_1$))

it knows that the comment is true: know (S, shortest_travel_to_go_to (Paris) = way$_1$)

Choice of the primary act + collaborativeness:

answering is more informative than confirming a belief $\Rightarrow$ primary act = ask_for_value

Reaction from the system (= reaction to the only primary act):

say_that (S, U, travel_duration (way$_1$) = 40 minutes)

S: "40 minutes"

Open question: logical forms or feature structures?

On the management of logical forms in dialogue systems:

- Each semantic content has to be represented using a precise logic that is often complex
- Each inference has to be translated into an information that is exploitable by the dialogue manager
- Inferences have to be filtered (because a lot are useless)
- Integrating an inference module into a dialogue system is not so easy
- That requires huge resources

On the management of feature structures:

- The representation is based on events with their characteristics: MMIL-like structure with events, participants, and relations
- The possible relations between events are specified
- A particular unification process can be designed
- The implantation is easier
- The resources that are needed are reasonable (the execution is fast because all the processes consist of the comparison of values)

Conclusion of the third part

Between speech act processing and collaboration:

- It is when processing indirect and composite speech acts that collaboration intervenes
- Thus: a system that aims at being collaborative must include a module for complex speech acts processing
- Thus: a system that aims at being collaborative must be able to produce and to manage hypotheses on the user’s mental states

Experience on dialogue system design:

- Separating the specificities of the application from the specificities of natural language dialogue and pragmatics can make the design very complex
- Logical forms are difficult to handle together with operational or industrial constraints (for now, a feature structure system is sufficient and much more simpler to design)
- The way to manage speech acts for the messages between the user and the system can be applied to the messages between the various modules of the system, for instance between the dialogue manager and the application

Fourth part

1. Definitions and examples
2. Dialogue models – management, control, history
3. Speech acts – dialogue act processing
4. Referring acts – multimodal reference resolution
   Making the correct links between a referring expression and an object of the application

Content of the fourth part

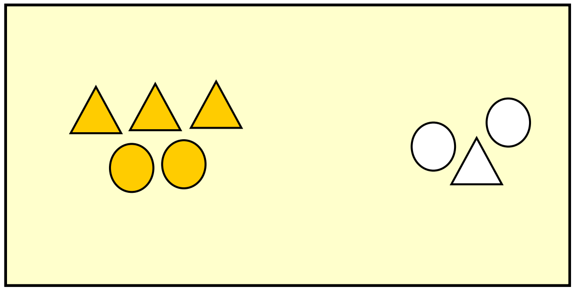
The problem: resolving references in multimodal dialogue systems by modeling the contents of the modalities and the context of their use (visual context, dialogue history, involved cognitive factors)

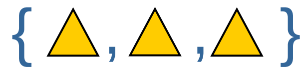
The concepts:
- reference domains
- salience

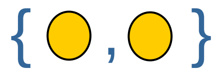
The model:
- interpretation of gesture / gesture trajectories
- interpretation of referring expressions
- architecture for referents and domains management

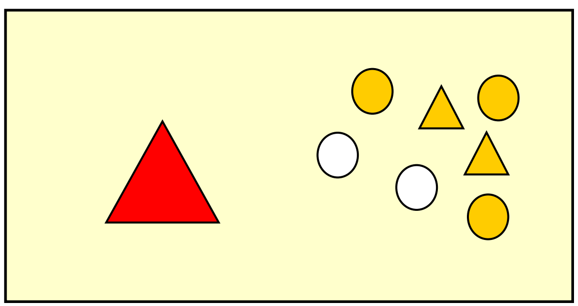
Extension: taking the referring mode into account

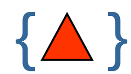
The problem: several modalities…



“these three triangles” → 

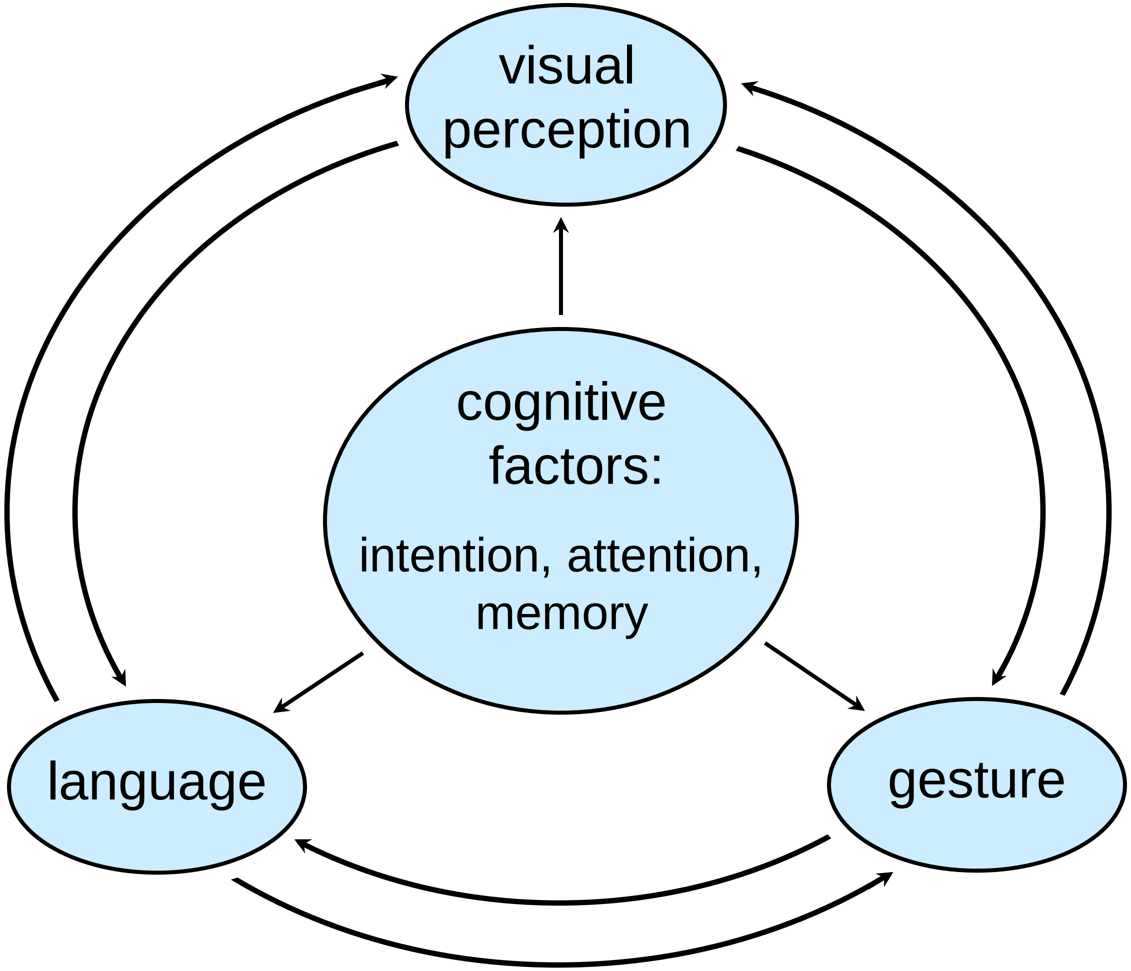
“the two circles” → 



“the triangle” → 

⇒ Analyzing and modeling the modalities contents

…that interact with each others



⇒ Analyzing and modeling cognitive factors

Approaches

From computer science

- very few exploitations of finer points of language and implicitness
- multimodal fusion at a low level and not during the interpretation process
- resolving references is not the main preoccupation

→ some models of architecture

From linguistics

- no deep interaction between language, gesture and visual context
- the gesture is often seen as pointing out directly the referents

→ some models for determination

From psychology

- a lot of theories, a lot of incompatibilities, few formalizations

→ some important basic principles

Approaches (cont.)

From pluridisciplinary approaches:

modeling cognitive factors within a computational framework

- (Clark 1983, 1986) = reference to objects is seen as a collaborative process; role of salience when producing referential expressions
- (Grosz & Sidner, 1986) = model involving a discursive structure, an intentional structure, and an attention structure
- (Hajičová, 1993) = salience
- (Beun & Cremers, 1998) = model for the management of salience, spatial attention, and functional attention

But no satisfying integration of the three modalities…

Objectives

- Modeling the contextual interpretation process for multimodal referring expressions:
  - taking into account the interactions between the modalities
  - taking into account the cognitive factors
  - based on the linguistic model of “reference domains”
  - with pluridisciplinary concerns

- Designing multimodal human-machine dialogue systems:
  - formalizing linguistic and cognitive theories
  - with a particular concern on integration

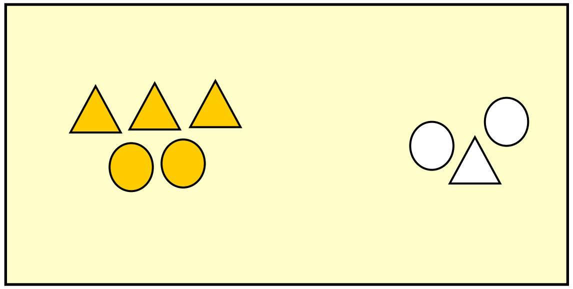
Reference domains: principles

- The identification of the referents pass through the identification of the contextual subset (reference domain) to which they belong

- Several contextual sources have to be integrated, each one implying a structuring into sets and subsets
  - reference domains linked to the linguistic context
  - reference domains linked to the shared visual context
  - reference domains linked to the task context

- Same structural properties $\Rightarrow$ confrontation, integration

Reference domains: example



RD$_1$

T: form
GF: coordination

DC: form

circle	triangle

T = type

GF = grouping factor

DC = differentiation criterion

RD$_2$

T: circle
GF: proximity

partitions

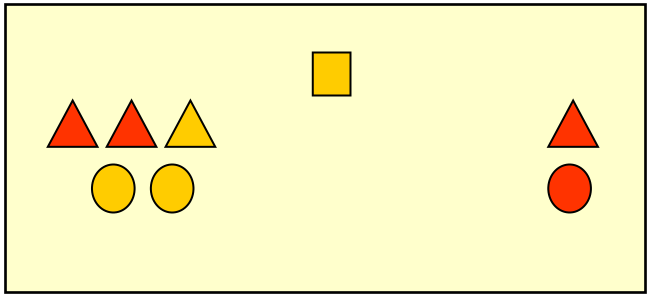
DC: x-axis

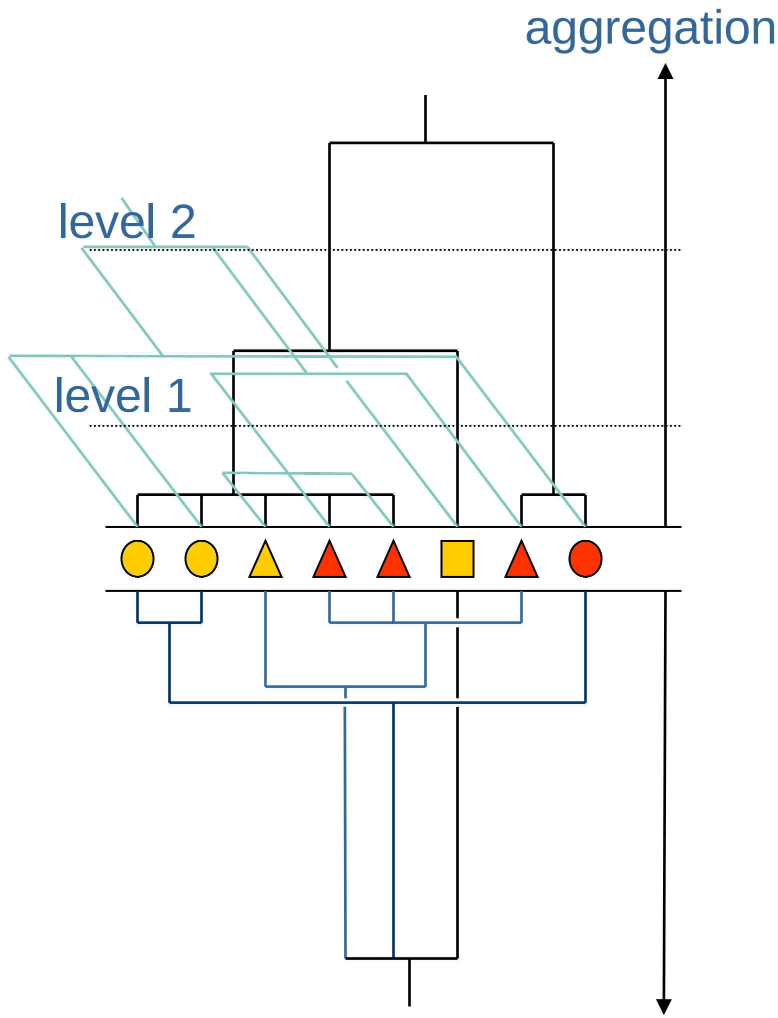
object$_1$	object$_2$

DC: color

object$_2$	object$_1$

Visual context modeling





- Grouping by proximity

  level 1: 5 + 1 + 2
  level 2: 6 + 2

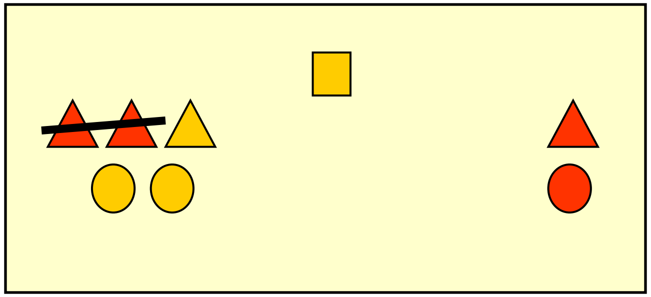
- Grouping by similarity

  level 1: 3 + 2 + 1 + 1 + 1
  level 2: 4 + 3 + 1

- Grouping by continuity

  level 1: 3 + 1 + 1 + 1 + 1 + 1
  level 2: 4 + 3 + 1

Visual context and gesture



- Grouping by proximity

  level 1: 5 + 1 + 2
  level 2: 6 + 2

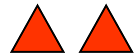
- Grouping by similarity

  level 1: 3 ▲ + 2 ● + 1 + 1 + 1
  level 2: 4 △ + 3 ○ + 1

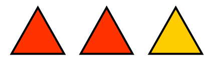
- Grouping by continuity

  level 1: 3△ + 1 + 1 + 1 + 1 + 1
  level 2: 4△ + 3 ○ + 1

sim level 1 + prox level 1

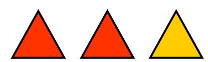


sim level 2 + prox level 1

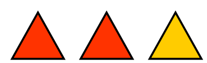


sim level 2 + prox level 1 + cont level 1

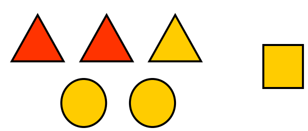
sim level 2 + prox level 2 + cont level 1



sim level 2 + prox level 2 + cont level 2

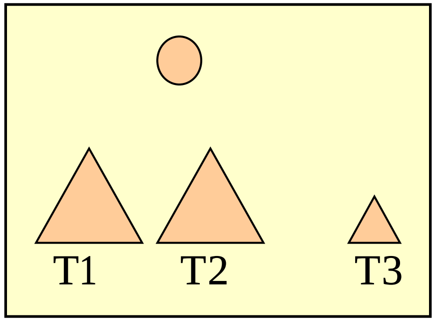


sim level 3 + prox level 2 + cont level 3



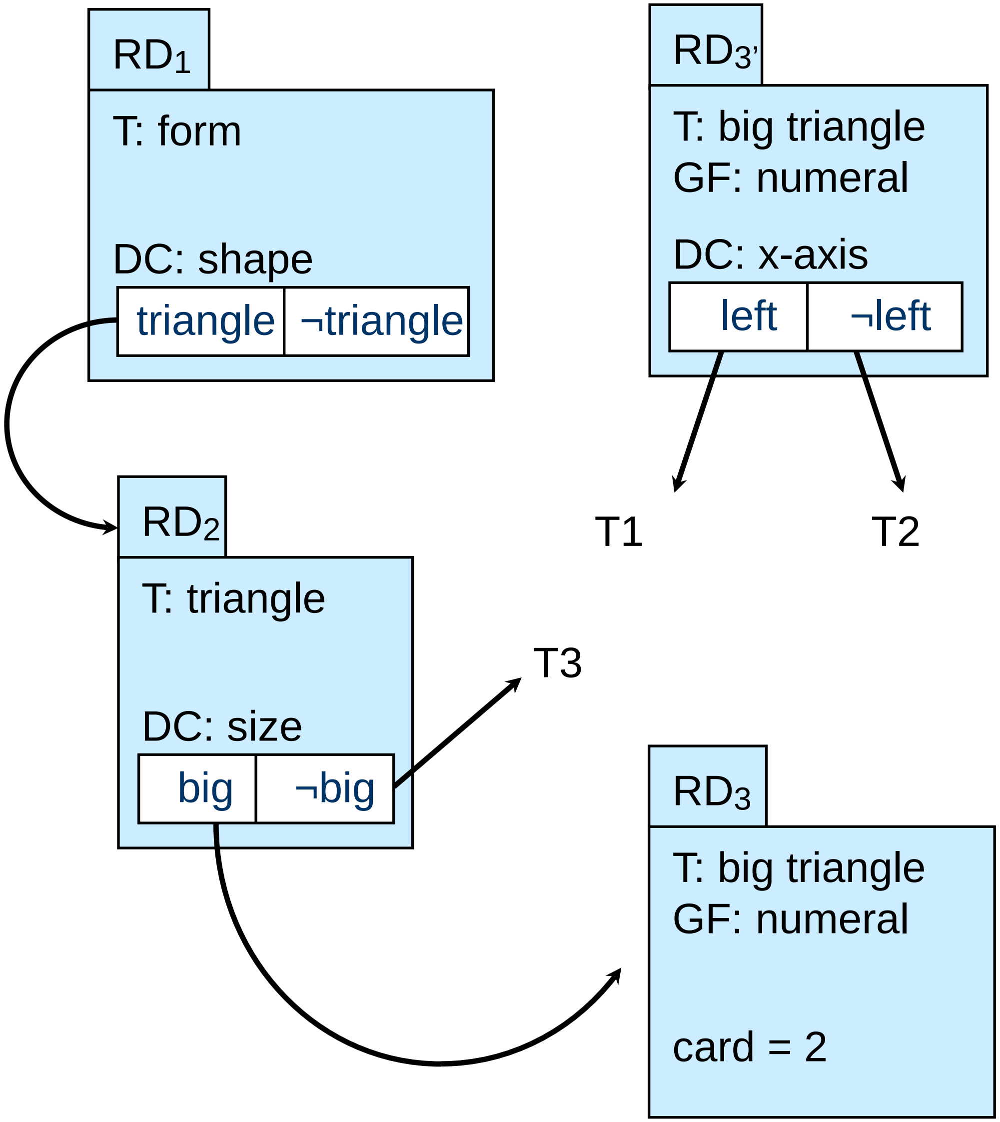
Linguistic context modeling

Visual scene



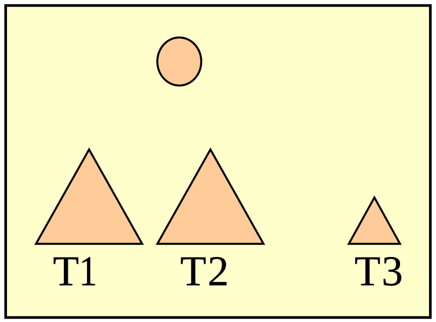
Referring expressions

- “the two big triangles”
- “the left triangle”
- “the right triangle”
- “the small triangle”



Task context modeling

Same example



Task

- applies to all triangles
- treat / name / delete / tidy…
- …following an implicit order

RD_1

T: triangle
GF: task-linked

DC: treatment order

T1	T2	T3

DC: treatment order

T3	T2	T1

RD_2

T: form

DC: task-linked

treated	¬treated

Quick characterization of visual salience

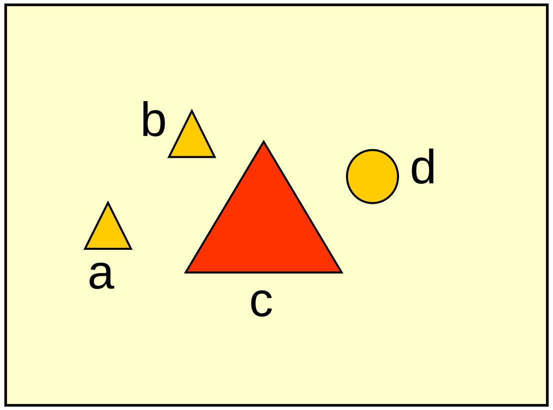
Factors linked to the structure of the visual scene:

- salience due to zones that catch the gaze
- salience due to perspective
- salience due to symmetry and/or balance

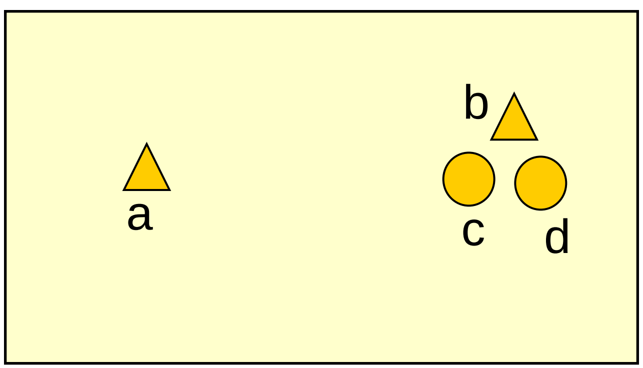
Factors linked to the visual rendering of the objects:

- salience due to an explicit obviousness
- salience due to the object category and physical properties
- salience due to the object localization in the scene
- salience due to the object incongruity
- salience due to the object dynamics

Numeric scores for visual salience



	category	size	color	disposition	
a:	0	0	0	0	→ 0
b:	0	0	0	0	→ 0
c:	0	1	1	0	→ **0.5**
d:	1	0	0	0	→ 0.25



a:	0	0	0	1	→ **0.25**
b:	0	0	0	0	→ 0
c:	0	0	0	0	→ 0
d:	0	0	0	0	→ 0

Scores exploitation when interpreting

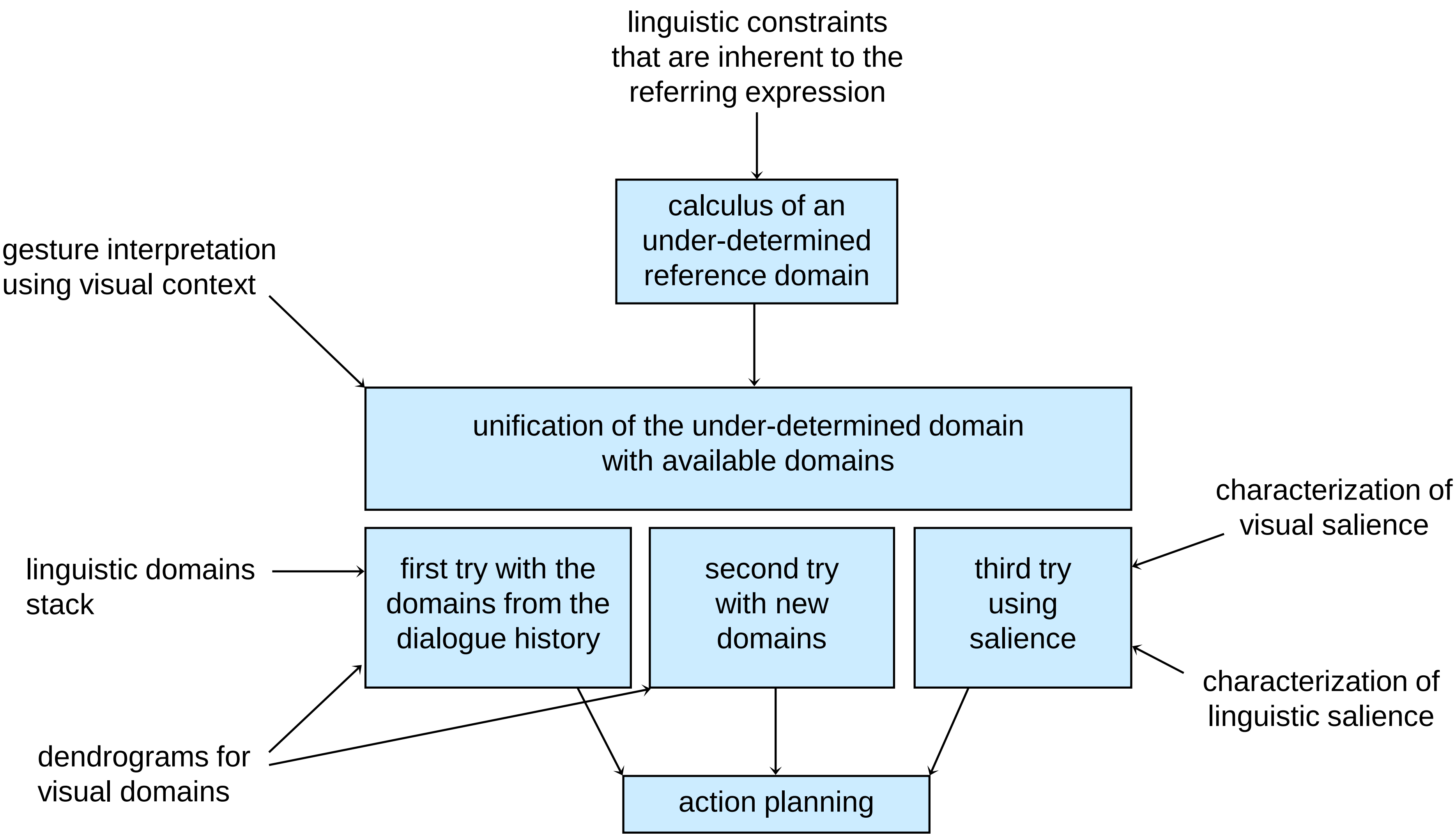
Interpreting a linguistic referring expression that is ambiguous considering the context:

- “the N” in a visual scene including several N, with only one of them being salient
- “the N” in an interactive context where one N has just been manipulated or mentioned (anaphora)
- “the N” interpreted as “the next N following the tasks succession” in a particular task context (script, plan)

Interpreting a multimodal referring expression that is ambiguous considering the context

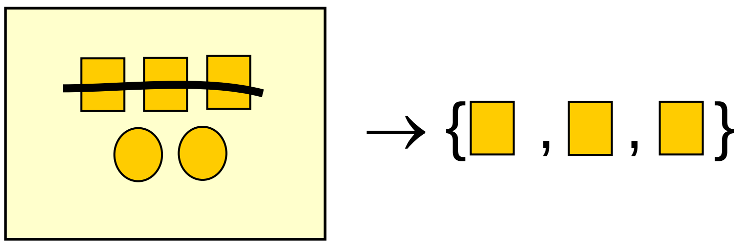
Predicting on which entity the user may focus and build on his next utterance

General algorithm for reference resolution

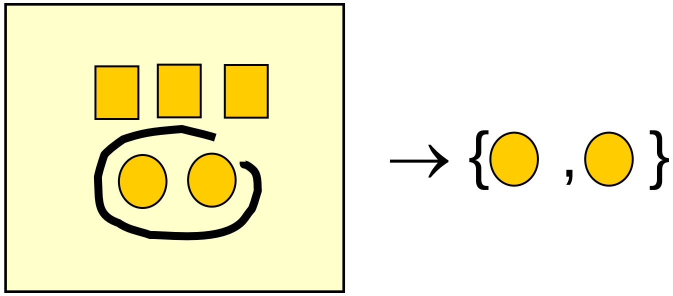


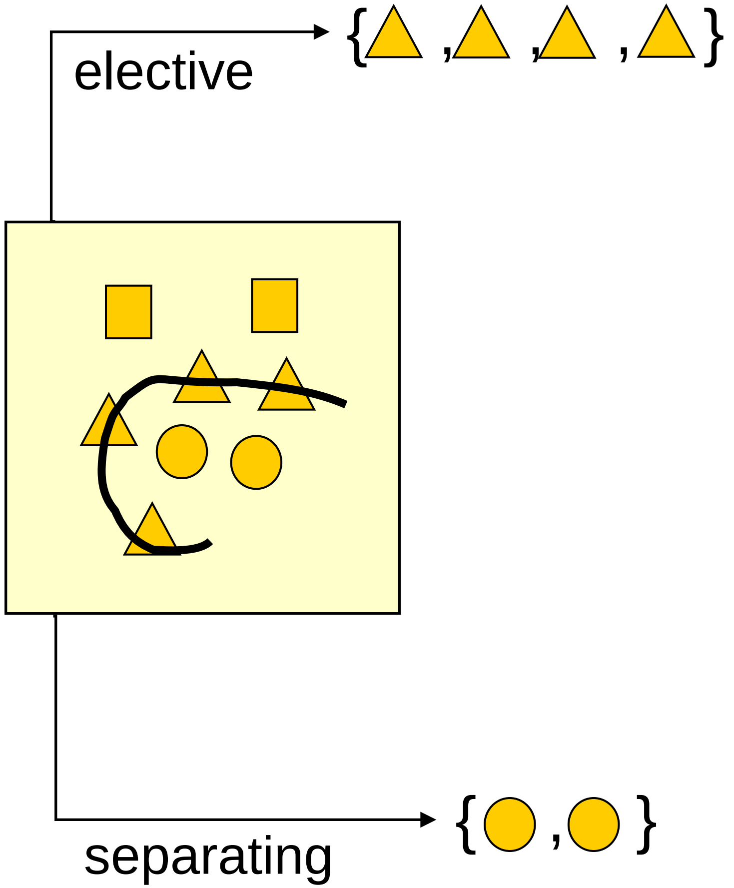
Categories of gesture trajectories

- Elective gesture

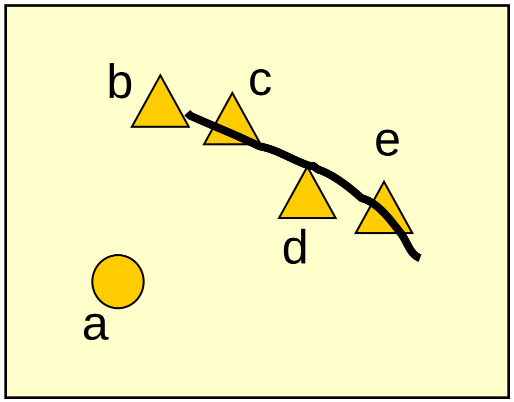


- Separating gesture





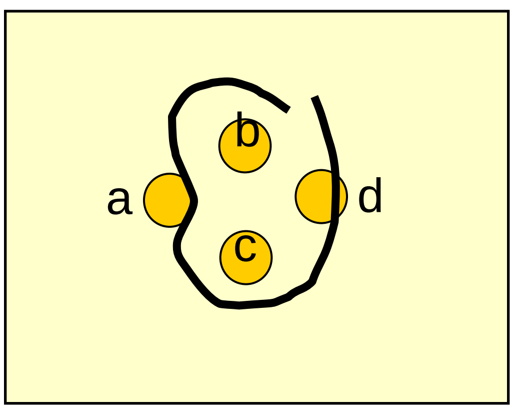
Scores for interpreting a trajectory



in immediate continuity

close at hand

	a	b	c	d	e
elective:	0	.5	1	.5	1
separating:	1	0	0	0	0

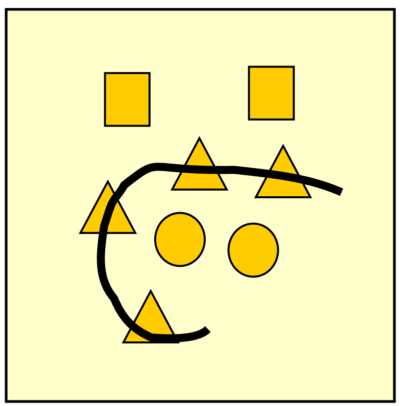


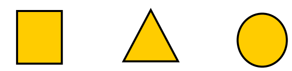
	a	b	c	d
elective:	.7	0	0	1
separating:	0	1	1	.8

covering ratio

semantic exclusion

Scores for interpreting a trajectory (cont.)





elective:	0	1	0
separating:	0	0	1

Using under-determined domains

Indefinite noun phrase

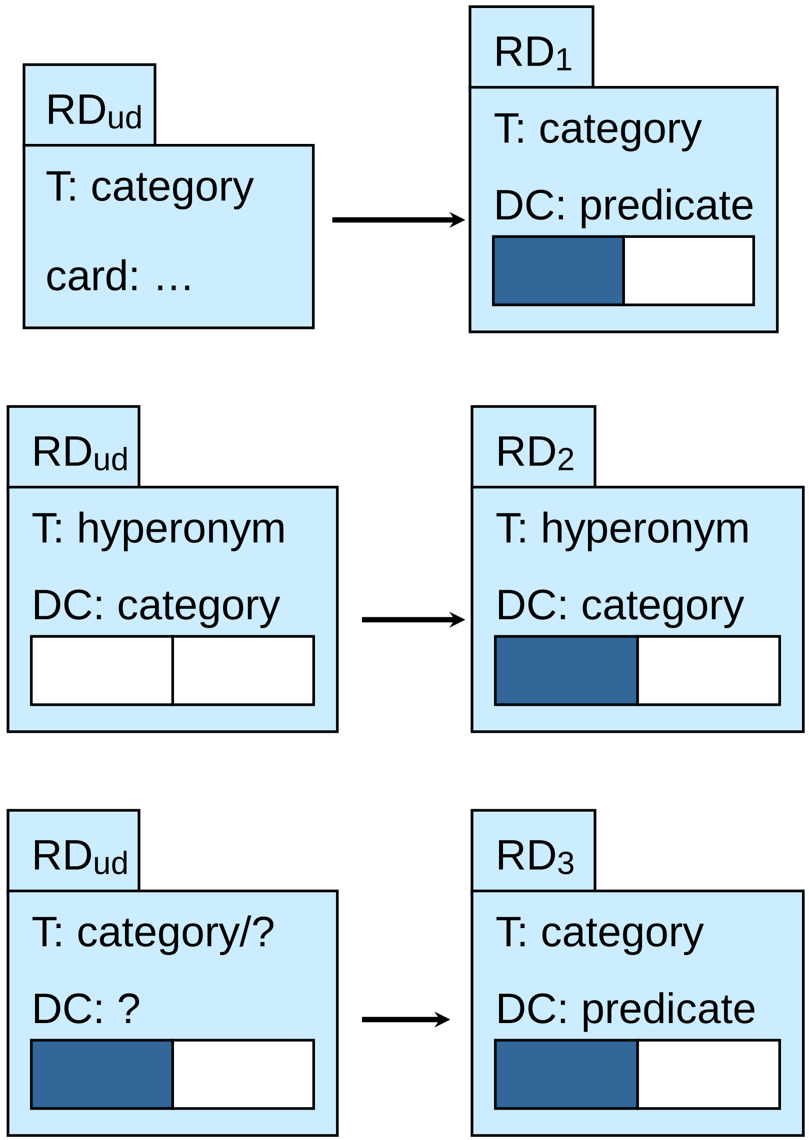
- Constraint on the type of the elements
- Creation of a new partition

Definite noun phrase

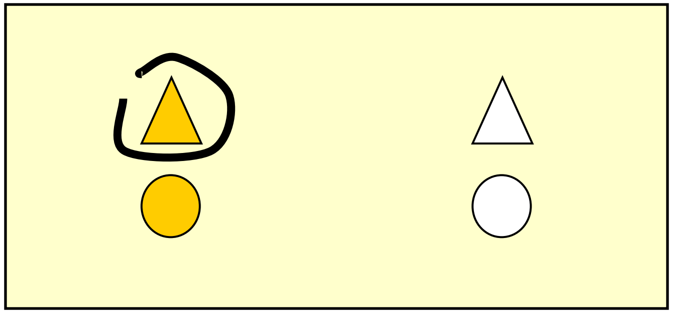
- Constraint on the existence of a partition
- Extraction and focusing

Demonstrative noun phrase

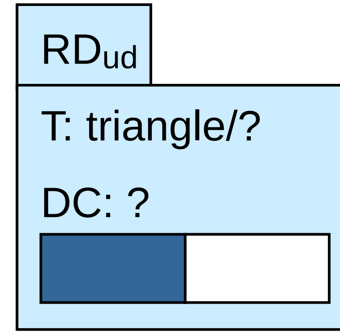
- Constraint on a focused partition
- Insertion in a new domain



Example with a demonstrative and a gesture

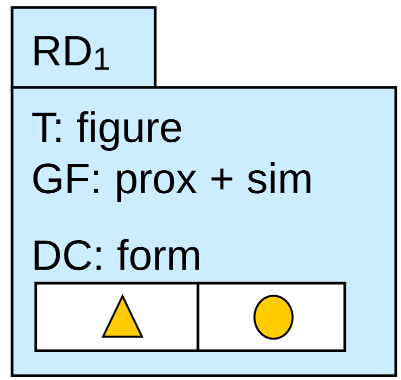


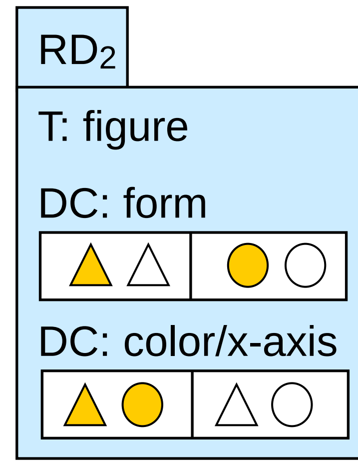
"this triangle"

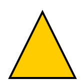


Available reference domains:

RD$_2$ is better than RD$_1$ because it justifies best the demonstrative





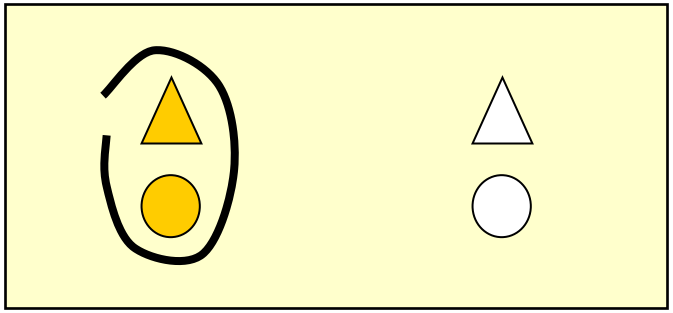
Pairing: the gesture focus 

Next utterance:

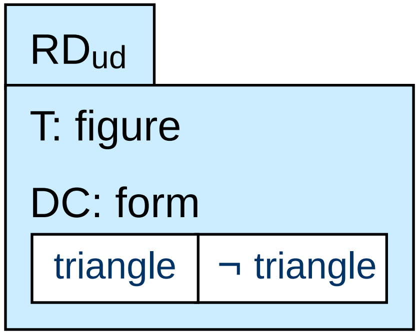
"the circle" activates RD$_1$

"the other one/triangle" activates RD$_2$

Example with a definite and a gesture

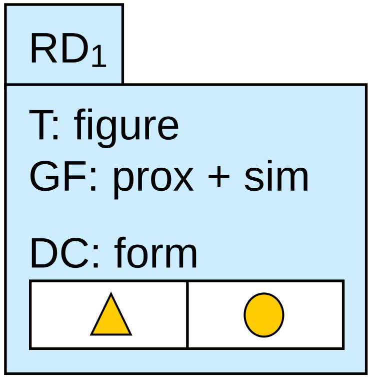


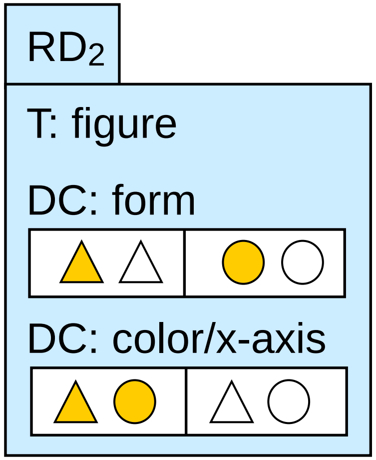
"the triangle"



Available reference domains:

RD$_1$ is much better than RD$_2$ because it corresponds to the gesture and it suits well the singular





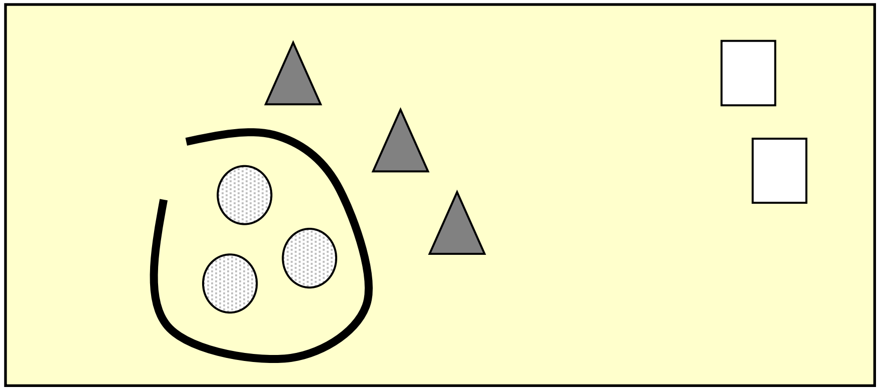
Next utterance:

"the circle" can be well interpreted in RD$_1$

"the other triangle" activates RD$_2$ but is nearly unacceptable

"the other one" is ambiguous (rather RD$_1$ for "the other figure")

A borderline example

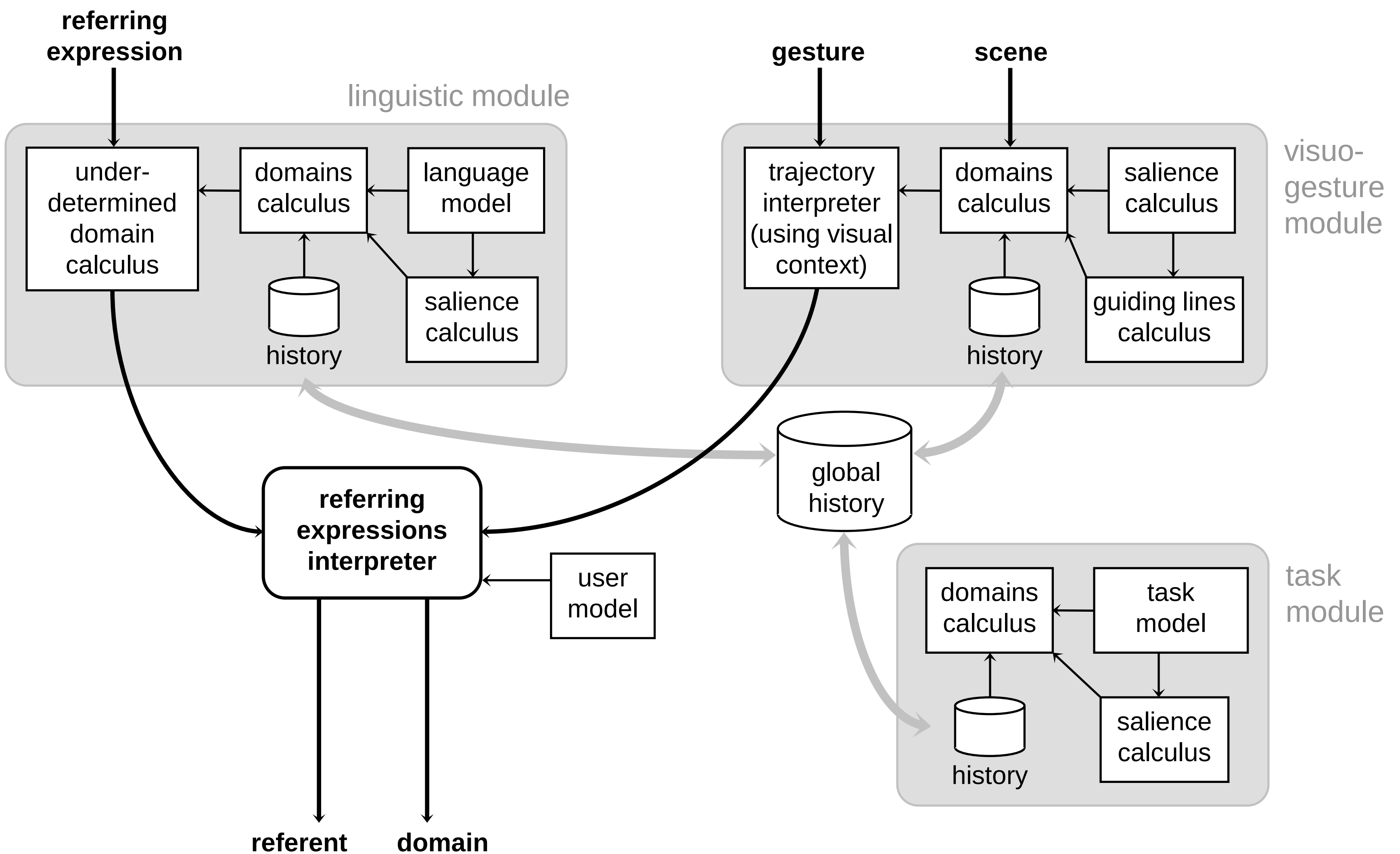


“the clearest figures”

+ gesture pointing out some figures that are not the clearest…

Solution: interpreting with the reference domain that corresponds to the perceptive group in the closeness of the gesture trajectory

Architecture for reference domains management



Conclusion of the fourth part

- **Referential terms are not linked to referring modes**
  The choice of a determiner, of the singular or plural form, of a coreferent pointing gesture, lead to clues that specify some aspects of the interpretation process

- **A referring action implies a reference domain**
  Reference domains allow to integrate several modalities and to model several aspects of implicitness:
  - context restriction
  - attention
  - **continuity** in a reference sequence

- **Continuity is the most important aspect**
  because it means that the reference model is a part of the dialogue model

General conclusions

1. Natural dialogue in natural language is still an on-going challenge, implying lot of links with several research disciplines
   linguistics, pragmatics, cognitive sciences

2. Human communication is multimodal,
   then human-machine communication has to be multimodal

3. Human-machine dialogue using socio-emotional embodied conversational agents is perhaps the most complex application of Natural Language Processing and Affective Computing

4. Automatic comprehension of language has to take into account various cognitive aspects in order to be efficient and 'intelligent'
   salience, vagueness, underspecification, ambiguity, implicitness, etc.

5. Machine learning techniques have increased the performance of some tasks, but classical symbolic techniques are still relevant, and the future may be hybrid…